Journal of the American Medical Informatics Association, 29(8), 2022, 1350–1365
https://doi.org/10.1093/jamia/ocac045
Advance Access Publication Date: 13 May 2022
Research and Applications

Research and Applications

Demonstrating an approach for evaluating synthetic geospatial and temporal epidemiologic data utility: results from analyzing >1.8 million SARS-CoV-2 tests in the United States National COVID Cohort Collaborative (N3C)

Jason A. Thomas[1], Randi E. Foraker[2,3], Noa Zamstein[4], Jon D. Morrow[4,5], Philip R.O. Payne[2,3], and Adam B. Wilcox[2,3]; the N3C Consortium[†]

[1]Department of Biomedical Informatics & Medical Education, University of Washington, Seattle, Washington, USA, [2]Division of General Medical Sciences, School of Medicine, Washington University in St. Louis, St. Louis, Missouri, USA, [3]School of Medicine, Institute for Informatics, Washington University in St. Louis, St. Louis, Missouri, USA, [4]MDClone Ltd., Be'er Sheva, Israel, and [5]Department of Obstetrics and Gynecology, New York University Grossman School of Medicine, New York, New York, USA

[†]A list of consortial authors and N3C core team contributions are given in 'Author Contributions' and 'Acknowledgments' sections, respectively.

Corresponding Author: Jason A. Thomas, PhD, Philips North America, LLC, 22100 Bothell Everett Hwy, Bothell, WA 98021, USA; thomasjt@uw.edu

Received 14 July 2021; Revised 11 March 2022; Editorial Decision 14 March 2022; Accepted 28 March 2022

ABSTRACT

Objective: This study sought to evaluate whether synthetic data derived from a national coronavirus disease 2019 (COVID-19) dataset could be used for geospatial and temporal epidemic analyses.

Materials and Methods: Using an original dataset ($n = 1\,854\,968$ severe acute respiratory syndrome coronavirus 2 tests) and its synthetic derivative, we compared key indicators of COVID-19 community spread through analysis of aggregate and zip code-level epidemic curves, patient characteristics and outcomes, distribution of tests by zip code, and indicator counts stratified by month and zip code. Similarity between the data was statistically and qualitatively evaluated.

Results: In general, synthetic data closely matched original data for epidemic curves, patient characteristics, and outcomes. Synthetic data suppressed labels of zip codes with few total tests (mean = 2.9 ± 2.4; max = 16 tests; 66% reduction of unique zip codes). Epidemic curves and monthly indicator counts were similar between synthetic and original data in a random sample of the most tested (top 1%; $n = 171$) and for all unsuppressed zip codes ($n = 5819$), respectively. In small sample sizes, synthetic data utility was notably decreased.

Discussion: Analyses on the population-level and of densely tested zip codes (which contained most of the data) were similar between original and synthetically derived datasets. Analyses of sparsely tested populations were less similar and had more data suppression.

Conclusion: In general, synthetic data were successfully used to analyze geospatial and temporal trends. Analyses using small sample sizes or populations were limited, in part due to purposeful data label suppression—an attribute disclosure countermeasure. Users should consider data fitness for use in these cases.

Key words: data utility, data sharing, synthetic data, COVID-19, electronic health records

INTRODUCTION

Background and significance

Coronavirus disease 2019 (COVID-19) has illustrated the need to disseminate accurate, timely, and useful epidemiologic public health data—especially data related to ongoing pandemics or pandemic preparedness. It has also highlighted the need to protect the privacy of individuals.[1,2] The National COVID Cohort Collaborative (N3C) was created to share and harmonize individual-level electronic health record (EHR) data into a single dataset.[3] The N3C has received, ingested, harmonized, and characterized[4] data from across the United States. To balance data access and privacy, N3C created 2 levels of datasets: (1) the limited dataset (LDS) which has 16 HIPAA Privacy Rule[5,6] direct identifiers stripped out except dates and zip codes, and (2) synthetic data which are computationally derived from the LDS to mimic the LDS data statistical distributions, covariance, and higher order interactions. Synthetic data generation can potentially protect privacy because synthetic data rows are not directly tied to the original source data.[7–11] Pending a pilot study and privacy validation, synthetic datasets are the only data under consideration to be shared outside of the N3C enclave.[3]

Applying privacy-preserving methods to data comes at varying cost to utility, producing a privacy-utility tradeoff.[9,12–15] De-identification removes granular geographic information such as street-level address. Obscuring dates reduces the utility of temporal data for some analyses, such as epidemic curves. However, these geographic and temporal data are critical components needed to measure key indicators of COVID-19 community spread[16] used to inform pandemic management decisions such as determining when to reopen schools[17] and businesses.[18] Thus, synthetic data may be the only privacy-preserving (pending privacy evaluations) N3C data that can be used to analyze some of the most critically important data related to pandemic management and preparedness while also providing citizens more transparency into the underlying data. However, previous research has reported deficits in how well-synthetic data mimic original data including limitations in their: ability to capture longitudinal relationships, model multiple data types, and perform well on small sample sizes[10,19,20] Due to the combination of potential widespread synthetic data dissemination, heightened research interest in COVID-19,[21] and the rise of “citizen science,”[22–24] the user base and applications of pandemic-related synthetic data will likely be heterogeneous and broad. Therefore, it is important to evaluate N3C synthetic data in a manner that can inform users with a wide range of intended use cases and definitions for synthetic data fitness for use.[25]

The utility of synthetic health data has been evaluated in other work[15,19,20,26–30] outside of N3C which applied a variety of the ways one can validate synthetic data.[31] However, N3C synthetic data utility has only been evaluated once before. Recently, the N3C synthetic data validation task team evaluated the utility of N3C synthetic data (MDClone, Be’er Sheva, Israel) across 3 use cases, one of which had a geospatial and temporal focus.[32] Foraker et al found the synthetic data had high utility for construction of a single aggregate epidemic curve of COVID-19 cases. However, it showed that rural zip codes with smaller population counts were more likely than urban zip codes to have zip code labels censored (suppressed) in the synthetic data, which is where a categorical variable’s value is replaced with the word “censored.” Zip code censoring is a method that aims to protect privacy of patients with particularly uncommon, and thus identifiable, features. To date, no analyses have been conducted on the N3C synthetic data to assess utility for analyses by individual zip codes and/or aggregate indicators beyond case counts (eg, percent positive) over time.

Objective

In this article, we describe the N3C synthetic data validation task team methods and results focused on evaluating whether synthetic N3C data can be used for geospatial and temporal epidemic analyses. Our replication studies focused on what we deemed were important and common analyses to be performed, such as epidemic curves for key indicators and creation of public-facing dashboards.[33–35] Our validation included replication of studies and general utility metrics[31] for: analyses at the zip code-level over time, construction of epidemic curves, and aggregate population characteristics. We believe these approaches balance the need to provide broad utility results for a wide range of analyses while also providing specific validation results relevant to analyses of common interest.

MATERIALS AND METHODS

Synthetic data

The MDClone ADAMS Synthetic Engine (MDClone Ltd., Be’er Sheva, Israel) derives a novel, synthetic dataset from input data, specifically computed to preserve the statistical properties, correlations, and higher-order relationships between variables while containing none of the individuals from the original data. The synthetic process fits new data points to a derived, multi-dimensional model so that information cannot be learned with certainty about any one individual in the population that cannot be learned about a group of other similar individuals.

An authenticated researcher specifies the patient cohort of interest from the underlying local data lake using the graphical query tool in ADAMS. The user selects the variables to be included in the output and can specify temporal relationships of interest. The derivative synthetic dataset is then computed from the original data for the selected cohort and variables, without exposing the user to the underlying original data.

For continuous variables, such computationally derived synthetic data are inherently privacy-preserving because, unlike de-identified data, the synthetic data process begins with a statistical model of the original data and samples entirely novel points to fit that model, maintaining the distribution, density, and co-variance between and among features within that model. There is no one-to-one correspondence between points in the original dataset and sampled points in the computed synthetic derivative; this prevents information from being learned from the latter about individuals from the former, other than their overall inter-variable relationships and their population-level descriptive properties. Indeed, if the process is run repeatedly on the same source data, each set of output will be unique, all sharing the same statistical properties as the source but none identical to it or to one another.

For categorical variables (eg, ZIP codes, genders), the finite number of categories presents the inherent possibility of an inference attack or other privacy-threatening methodology.[36–38] The MDClone engine therefore enforces a number of proprietary techniques, beyond the control of the user, to mitigate this risk.[39] The rows (ie, patients) in the source database are sorted by the discrete values of categorical variables, thus grouping identical rows. If in any group there are fewer than a system-defined number of rows, termed κ, some of these discrete values are replaced by the word “censored.”

This process, which reduces the variance in the categorical variables, is repeated until all rows are members of groups of size ≥κ.

Each resulting unique group of rows includes a matrix representing the group's associated numeric variables. Because knowledge of some of the numeric values might permit an attacker to discover something about other variables, MDClone replaces each matrix with an alternative matrix of similar statistical properties. Specifically, the rows are clustered into sets of <κ rows, minimizing the scaled Euclidean distance between data points and preserving, within each cluster, statistical characteristics for every pair of variables. As there is an unlimited number of possible alternative matrices for each cluster that satisfies this requirement, the algorithm selects each solution randomly, resulting in an irreversible process and preventing the re-creation of the original data from the result.

Finally, to protect against a difference attack based on knowing the exact size of the original population,[40] the number of rows that are created to fit the overall data model is altered slightly. This small, arbitrary change in the number of rows prevents an attacker from deducing the exact size of the original population but does not affect the overall statistical properties of the resulting dataset.

Data model

The N3C data analyzed include individual-level EHR data enriched with social determinants of health (SDOH) at the 5-digit zip code level. The data have been harmonized into the Observational Medical Outcomes Partnership (OMOP) common data model (CDM) v5.3.1[3,41] and are the same datasets described in a previous N3C synthetic data validation use case.[32] The N3C LDS as of November 30, 2020—which included 34 data source partners—was used as the data source. MDClone received a copy of the LDS then transformed these data from the N3C harmonized data model into MDClone's data model. Afterwards, the required data needed for the study team's analyses were extracted by MDClone from the transformed LDS for use as the "original" dataset. A synthetic derivative of this transformed original dataset was then created by MDClone. MDClone provided both the original and synthetic datasets to the research team for evaluation within the N3C secure enclave environment (Supplementary Figure S4 in the flowchart). Information on the MDClone data model and pre-processing steps specific to this study are described further in the Supplementary Material and in general in past analyses of MDClone data.[27]

Both the original and synthetic data were formatted as a single table adhering to the same schema, with each row representing a single COVID-19 test. The table had the following columns: test result (positive/negative; only each patient's first negative and/or first positive test included), age at confirmed test result; admission start date days from reference if admission occurred within ±7 days of COVID-19 positive test result; death (null/yes) during admission; admission length of stay (LOS); patient's state of residence; source partner with which the patient was affiliated; and patient's 5-digit zip code. The data also included the following SDOH columns determined by the patient's zip code: total population in zip code; percent of residents under the poverty line; percent without health insurance; and median household income.

As in Foraker et al, we used consistent definitions for censored and uncensored zip codes. Censored zip codes were those present within the original data not found ($n = 11\,222$) within the synthetic dataset either because the zip code was suppressed by labeling the zip code "censored" or removed within the synthetic dataset to protect privacy. Conversely, uncensored zip codes were defined as discrete zip codes found in the original and the synthetic data ($n = 5819$).

Analysis (excluding the *post hoc* privacy evaluation)

All analyses were conducted solely by 1 author (JAT). All code was written in Python (v3.6.10) and—as required by N3C—ran within the secure N3C enclave using the Palantir Foundry Analytic Platform (Palantir Technologies, Denver, CO, USA). The entirety of code used in this analysis is contained within a single Foundry Code Workbook using a saved Spark environment to preserve required software versions and dependencies. The code workbook and source data have been stored within the N3C enclave so that they may inform and be reused in future validation work.

Summary of data

Descriptive statistics were calculated and reported in Table 1 for age, number of unique zip codes present, LOS, and admission date after positive test stratified by patients who were tested, positive, admitted, and who died during admission. Number of unique zip codes present excluded null or censored zip codes. The difference between original and synthetic values was reported as the raw synthetic difference (synthetic—original). The difference as a percentage of the original value was reported as synthetic difference percentage (raw synthetic difference/original).

Aggregate epidemic curves

We constructed aggregate epidemic curves using each dataset spanning January 1 through November 30, 2020 (Figure 1). The following key indicators were calculated and visualized: tests, cases (reproduced from Foraker et al to view others in context), percent positive, admissions, and deaths during admission. Each indicator had the following daily metrics calculated: count (discrete indicators) or value (continuous indicators), 7-day midpoint moving average, and 7-day slope (count or daily value—its value 6 days prior). To assess the statistical difference between original and synthetic epidemic curves, we conducted the paired 2-sided *t*-test (scipy v1.5.3, stats.ttest_rel) and 2-sided Wilcoxon signed-rank test (scipy v1.5.3, stats.wilcoxon) for all metrics across all indicators (Table 2), treating each dataset's daily results as a pair.

Distribution of tests; censoring of zip codes

To assess the distribution of tests by zip code and threshold of zip code censoring, we calculated the total number of tests per zip code in the original and synthetic data. In the synthetic data, we excluded rows with a censored ($n = 44\,337$; 2.4%) or null ($n = 444\,092$; 23.9%) zip code. In the original data, we excluded rows with a null ($n = 444\,380$; 24.0%) zip code. We computed the 99th, 97.5th, and 90th percentiles of tests per zip code in the original data. The distributions of tests by zip code were plotted as a histogram (Figure 2) with the synthetic and original data overlaid. Additionally, we calculated the distribution of tests by zip code in the original data that were censored in the synthetic data, then plotted the result as a histogram (Supplementary Figure S3). We then calculated the difference in patients' SDOH values within the original data, comparing patients whose zip codes were censored within the synthetic data to those whose zip codes were not censored (Table 3).

Top 1% paired zip codes' epidemic curves

Next, we assessed synthetic epidemic curves' performance at the zip code level, focusing on zip codes with relatively abundant data. We

Table 1. Testing and outcomes characteristics: comparison of original versus synthetic data

	Original	Synthetic	Synthetic difference (raw)	Synthetic difference (%)
Tests (*n*)	1 854 968	1 854 950	−18.00	0.00
Age (mean)	44	44	0.00	0.00
Age (stdev)	22.16	22.16	0.00	0.00
Age (median)	43.52	43.51	−0.01	−0.02
Age (IQR)	35.08	35.04	−0.04	−0.11
Unique zip codes (*n*)	17 041	5819	−11 222.00	−65.85
Positive (count)	195 200	195 198	−2.00	0.00
Positive (%)	10.52	10.52	0.00	0.00
Age (mean)	41.54	41.53	−0.01	−0.02
Age (stdev)	20.4	20.42	0.02	0.10
Age (median)	39.65	39.56	−0.09	−0.23
Age (IQR)	31.84	31.81	−0.03	−0.09
Unique zip codes (*n*)	6660	1798	−4862.00	−73.00
Negative (*n*)	1 659 768	1 659 752	−16.00	0.00
Negative (%)	89.48	89.48	0.00	0.00
Age (mean)	44.29	44.29	0.00	0.00
Age (stdev)	22.34	22.34	0.00	0.00
Age (median)	44.08	44.08	0.00	0.00
Age (IQR)	35.36	35.34	−0.02	−0.06
Unique zip codes (*n*)	16 668	5805	−10 863.00	−65.17
Admitted (*n*)	23 044	23 044	0.00	0.00
Admitted (%)	1.24	1.24	0.00	0.00
Age (mean)	57.87	57.85	−0.02	−0.03
Age (stdev)	19.77	19.74	−0.03	−0.15
Age (median)	59.98	60	0.02	0.03
Age (IQR)	28.2	28.22	0.02	0.07
Days after positive test (mean)	−0.07	−0.1	−0.03	42.86
Days after positive test (stdev)	1.77	1.74	−0.03	−1.69
Days after positive test (median)	−0.05	−0.04	0.01	−20.00
Days after positive test (IQR)	0.88	0.88	0.00	0.00
LOS (mean)	6.48	8.32	1.84	28.40
LOS (stdev)	290.81	10.66	−280.15	−96.33
LOS (median)	5	5	0.00	0.00
LOS (IQR)	8	8	0.00	0.00
Unique zip codes (*n*)	3132	1515	−1617.00	−51.63
Died (*n*)	2032	2032	0.00	0.00
Died (%)	0.11	0.11	0.00	0.00
Age (mean)	71.81	71.81	0.00	0.00
Age (stdev)	14.57	14.65	0.08	0.55
Age (median)	73.26	73.21	−0.05	−0.07
Age (IQR)	19.68	19.58	−0.10	−0.51
Days after positive test (mean)	−0.32	−0.32	0.00	0.00
Days after positive test (stdev)	1.39	1.36	−0.03	−2.16
Days after positive test (median)	−0.14	−0.11	0.03	−21.43
Days after positive test (IQR)	0.91	0.93	0.02	2.20
LOS (mean)	13.69	13.71	0.02	0.15
LOS (stdev)	12.93	13.05	0.12	0.93
LOS (median)	10	10	0.00	0.00
LOS (IQR)	13	13	0.00	0.00
Unique zip codes (*n*)	831	16	−815.00	−98.07

created a list of zip codes from the original data in the 99th percentile ($n = 171$) by total number of tests, then removed any zip codes without an uncensored matched zip code pair in the synthetic data ($n = 0$). We randomly sampled 10 zip codes from the list and constructed epidemic curves for these zip codes' original and synthetic data Figures 3 and 4). Each epidemic curve was constructed using the same date range, methods, and metrics as the aggregate epidemic curves described above with the following change: we only assessed tests and admissions indicators due to the infrequency of death during admission at the zip code level and manuscript space limitations.

Monthly zip code pairwise synthetic error

We compared the difference in monthly counts of tests, cases, and admissions between the original data and paired uncensored synthetic zip codes. To do so, we calculated each dataset's num-

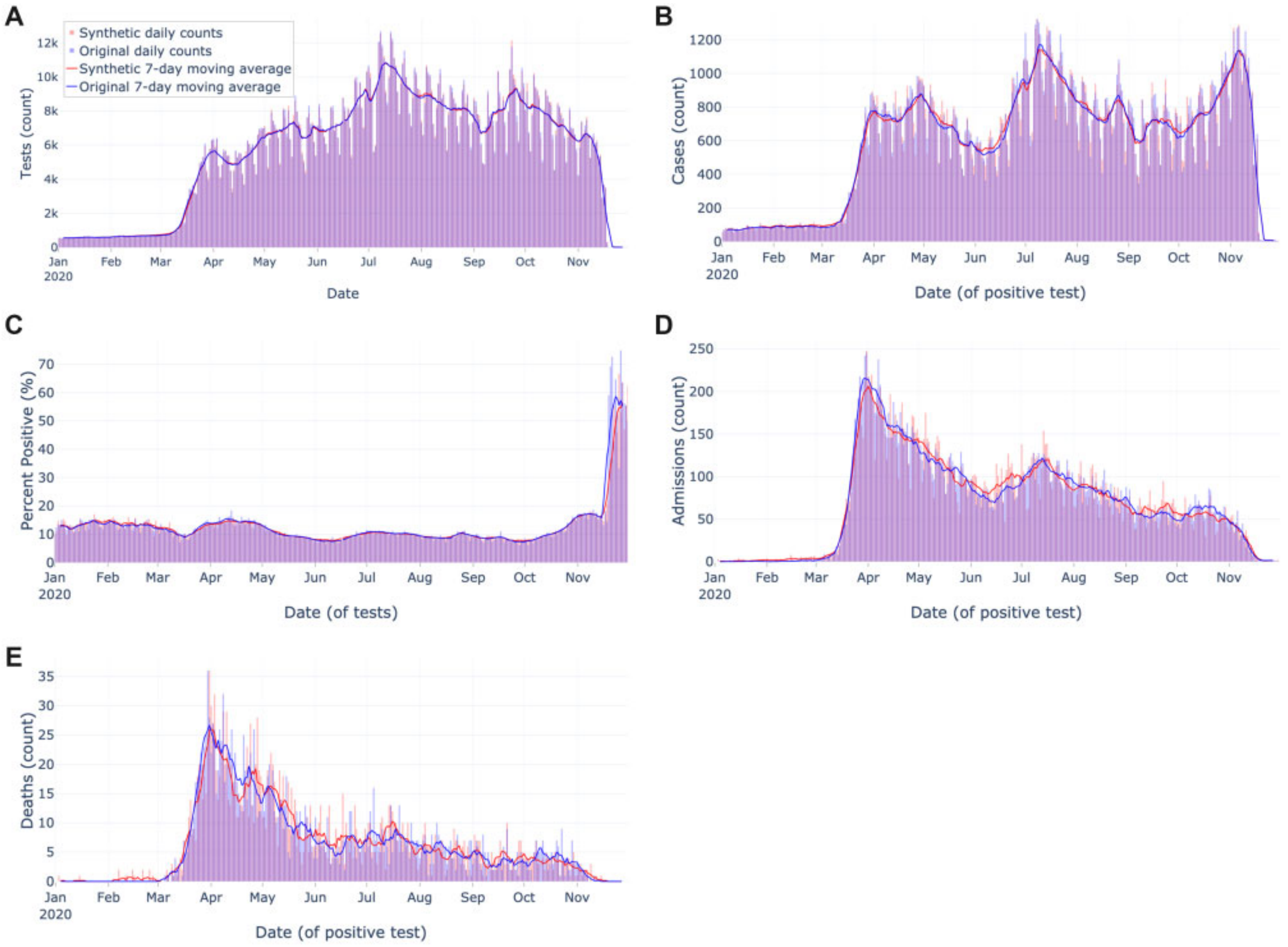

Figure 1. Aggregate epidemic curves with counts (vertical bars) and 7-day moving averages (smoothed line) for (A) tests, (B) cases, (C) percent positive, (D) admissions, and (E) deaths during admission. Color encodings include original data (light blue) and synthetic data (light red), with their overlap (purple). As counts get smaller from tests to deaths, the epidemic curves visually appear less similar.

Table 2. Tests for significant differences between aggregate original and synthetic epidemic curves

Key indicator	Metric	Wilcoxon result	*P*-value	*T*-test stat	*P*-value
Tests	Counts	25 354.5	0.300	−0.007	0.994
	7-day average	25 458.5	0.428	−0.025	0.980
	7-day slope	26 075	0.735	−0.002	0.998
Cases	Counts	26 288	0.496	−0.002	0.998
	7-day average	26 005	0.775	−0.006	0.996
	7-day slope	25 788.5	0.898	−0.002	0.998
Percent positive	Counts	26 407	0.426	−0.932	0.352
	7-day average	24 038	0.072	−2.258	**0.025**
	7-day slope	27 083	0.972	0.129	0.896
Admissions	Counts	21 405	0.247	−0.007	0.995
	7-day average	24 299	0.197	−0.030	0.976
	7-day slope	22 825.5	0.894	−0.011	0.991
Deaths	Counts	13 881	0.748	0	1
	7-day average	19 171.5	0.247	−0.023	0.982
	7-day slope	16 632	0.866	−0.011	0.992

Boldface values significant at < 0.05.

ber of tests, cases, and admissions for every zip code stratified by month for each month the zip code had ≥1 test. Then, the datasets were outer merged on month and zip code (Figure 5). Synthetic error, defined as the difference between the synthetic monthly count and the original data monthly count value, was computed for every zip code month pair. The distribution of synthetic error was visualized (Figure 6) for tests, cases, and admissions.

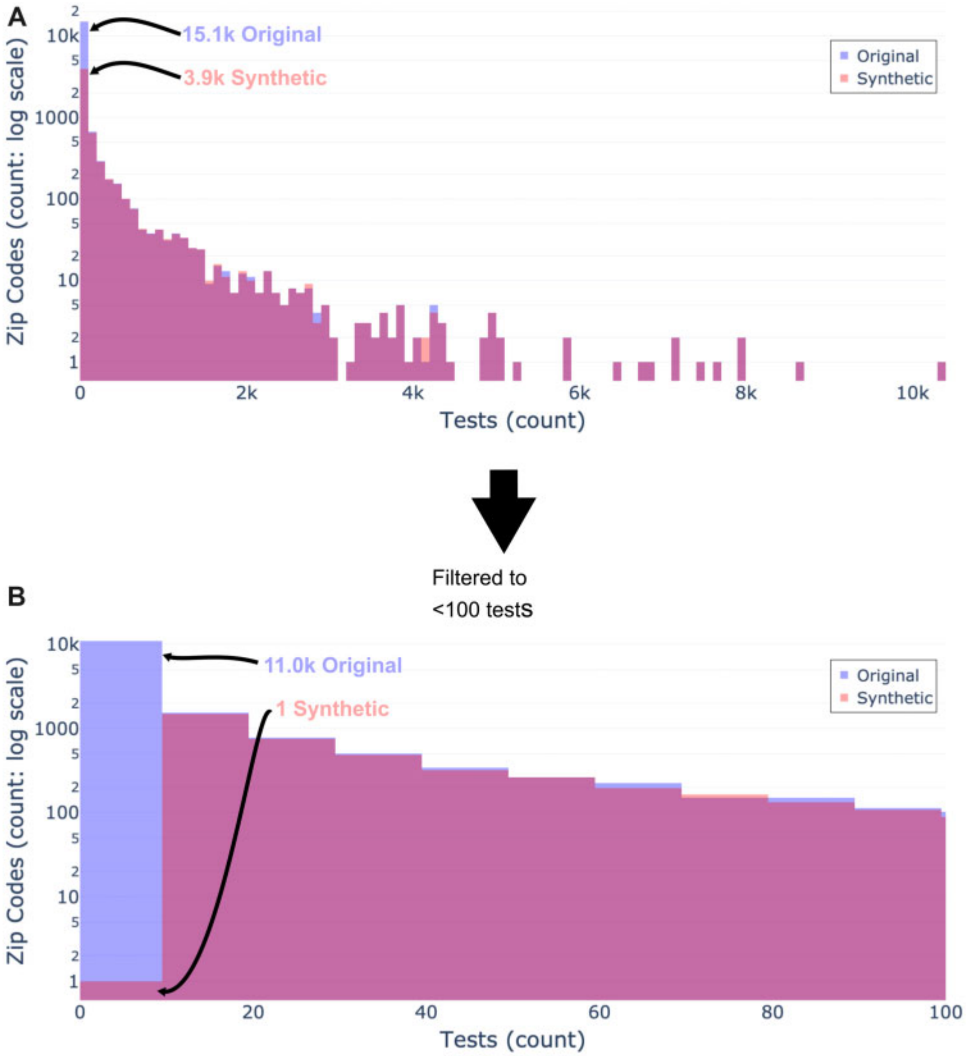

Figure 2. Distributions of total tests by zip code shown by original data (light blue) and synthetic data (light red), and their overlap (purple). (A) All data binned by 100. (B) Filtered data with a bin size of 10 to only show the distribution of tests by zip code in zip codes with <100 tests. Both *y*-axes use a log scale. As seen in panel A, the vast majority of tests are conducted in a minority of zip codes. As seen in panels A and B, the distribution of the synthetic data closely matches the original data at >10 tests per zip code.

Table 3. SDOH and age of patients in the original data whose zip codes were censored versus uncensored

SDOH	Censored status	mean	Standard deviation	Median	IQR
Age (years)	Uncensored	44.0	22.2	43.5	35.0
	Censored	46.4	22.0	48.7	40.1
	Uncensored Difference (raw)	−2.4	0.2	−5.2	−5.1
Median household income ($)	Uncensored	64 092.6	23 973.9	59 324.0	29 241.0
	Censored	63 101.5	28 964.1	55 625.0	28 857.0
	Uncensored Difference (raw)	991.1	−4990.2	3699.0	384.0
Percent under the poverty line	Uncensored	13.7	9.0	11.3	11.2
	Censored	13.3	9.6	11.2	10.9
	Uncensored Difference (raw)	0.4	−0.6	0.1	0.3
Percent without health insurance	Uncensored	8.7	5.1	7.6	7.0
	Censored (raw)	9.2	6.7	7.8	7.7
	Uncensored Difference (raw)	−0.5	−1.6	−0.2	−0.7
Total population of zip code	Uncensored	29 758.7	17 992.4	28 479.0	25 220.0
	Censored	15 493.9	17 967.1	7935.0	23 119.3
	Uncensored Difference (raw)	14 264.8	25.3	20 544.0	2100.7

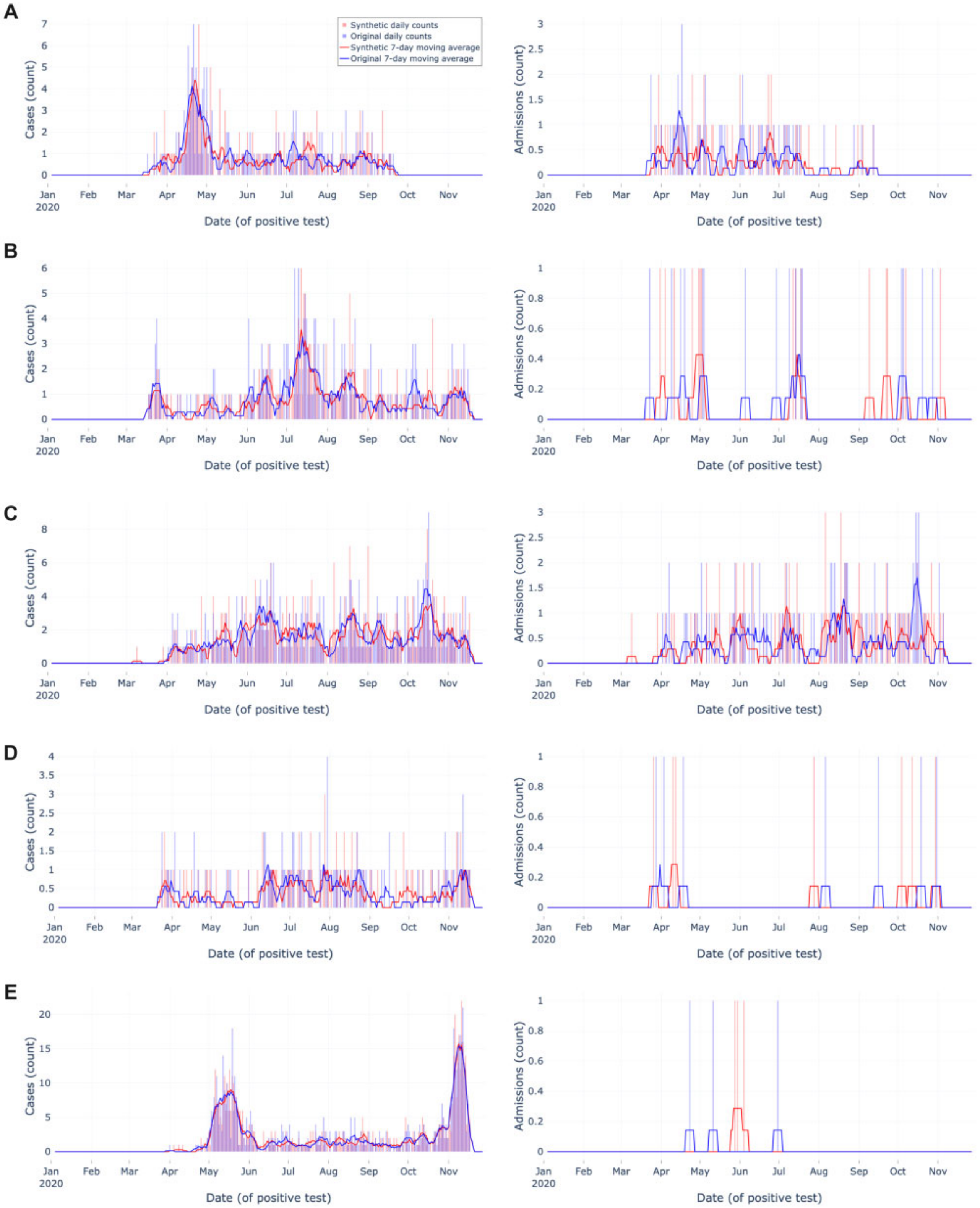

Figure 3. Zip code-level epidemic curves with counts (vertical bars) and 7-day moving averages (smoothed line). Color encodings include original data (light blue) and synthetic data (light red), with their overlap (purple). Each row (A–E) corresponds to a different randomly sampled zip code visualizing cases (left column) and admissions (right column). Synthetic data are more similar to original data when indicator density is higher. Overall, synthetic data closely match overall trends and closely match start and end dates.

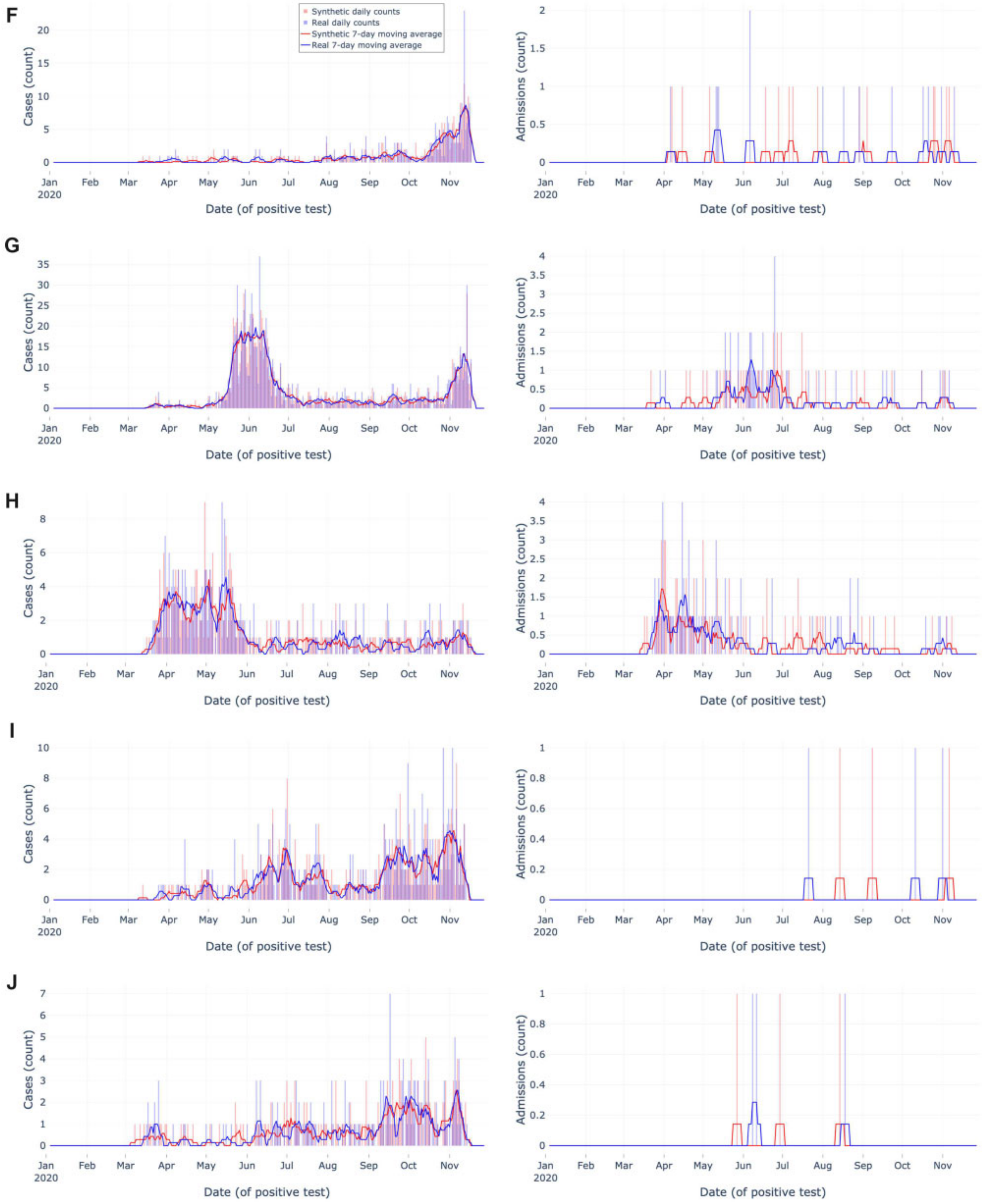

Figure 4. Zip code-level epidemic curves with counts (vertical bars) and 7-day moving averages (smoothed line). Color encodings include original data (light blue) and synthetic data (light red), with their overlap (purple). Each row (A–E) corresponds to a different randomly sampled zip code visualizing cases (left column) and admissions (right column). Synthetic data are more similar to original data when indicator density is higher. Overall, synthetic data closely match overall trends and closely match start and end dates.

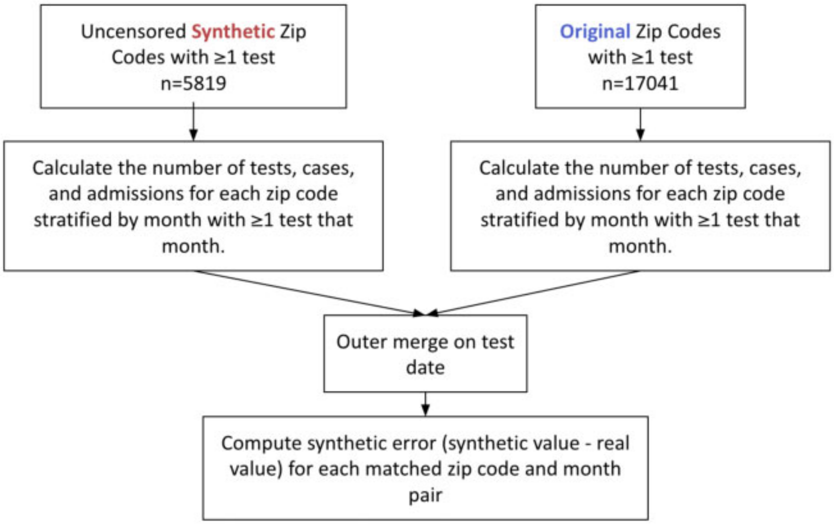
Uncensored Synthetic Zip Codes with ≥1 test n=5819
Original Zip Codes with ≥1 test n=17041
Calculate the number of tests, cases, and admissions for each zip code stratified by month with ≥1 test that month.
Calculate the number of tests, cases, and admissions for each zip code stratified by month with ≥1 test that month.
Outer merge on test date
Compute synthetic error (synthetic value - real value) for each matched zip code and month pair

Figure 5. Workflow of synthetic error experiment showing synthetic data on the left, original data on the right which are then merged to allow the calculation of synthetic error to be made.

Visualizations

All visualizations (Plotly v4.14.1, Plotly Technologies Inc.) were interactive, allowing N3C enclave users to zoom in/out, pan, and hover to see values and/or labels. In this manuscript, static figures are presented. Log scales were avoided when possible and, when used, annotated to draw attention to the scale.

Visualizations that overlaid both datasets adhered to consistent style conventions. We encoded synthetic and original data sources as red and blue, respectively. Vertical overlaid bars were set to an opacity of 0.35 to (1) provide contrast between 2 datasets and (2) allow additional tracings, such as 100% opacity 7-day moving averages used in epidemic curves, to be seen on top of the bars.

All visualizations were created using colorblind-safe color mappings. Categorical mappings encoding values besides data source (synthetic or original) used hexadecimal color codes found in the seaborn colorblind palette.[42,43] Each visualization was qualitatively tested for colorblind deuteranopia, protanopia, and tritanopia interpretability by 1 member of the research team (JAT) using Color Oracle.[44]

Post hoc privacy evaluation

A *post hoc* assessment was done to determine the privacy preservation of the synthetic data produced for this study, by addressing the possibility that the presence of an individual in the original dataset could be inferred from the synthetic data. Specifically, we queried whether there were any rows in the synthetic dataset that share identical attributes across continuous and categorical values with rows in the original dataset. Since exact matches across continuous variables are expected to be rare,[45] we also examined whether subjects with a unique value in a categorical variable or bearing a rare combination of categorical values were reproduced in the derivative synthetic dataset.

RESULTS

There were nearly 2 million tested patients (original $n = 1\,854\,968$; synthetic $n = 1\,854\,950$) in each dataset. As seen in Table 1, the overall central tendencies of variables of interest overall were similar between the synthetic data and original data, especially for age and percent positive/admitted/died. The raw synthetic difference was 0, rounded to 2 decimal points, roughly one-third (18/50 rows in Table 1) of the time. The variable with the greatest synthetic difference was unique zip codes, with between a 65% and 98% reduction in unique zip codes. Median LOS and interquartile range (IQR) for admitted patients were exactly the same, yet the mean LOS was 6.48 (±290.81) and 8.32 (±10.66) days for original and synthetic values, respectively. The extreme LOS standard deviation observed in the original data was due to an erroneous outlier. A single row in the original data had an extreme negative LOS (∼−44 000 days; ∼−120 years) and 11 rows with a LOS = −1. The synthetic data also had negative LOS values ($n < 10$), but the values were greatly attenuated, ranging from −1 to roughly −175. As a result of noticing this extreme LOS, all columns in the original and synthetic data were assessed for implausible outliers likely to be the result of data quality issues. None were found.

In our statistical analysis, no differences were found between the aggregate epidemic curves besides the 7-day average of percent positive ([*t*-test *P*-value = .025; Wilcoxon *P*-value = .072], Table 2).

Differences were observed between patients' SDOH values whose zip codes were uncensored in the synthetic data compared to patients whose zip codes were censored in the synthetic data (Table 3). The largest differences were found in the total population of zip code and age. Patients with uncensored zip codes lived in more populous zip codes (median total population: uncensored = 28 479, censored = 7935) and were younger (median age: uncensored = 43.5, censored = 48.7).

The randomly sampled top 1% paired zip codes' epidemic curves are presented in Figures 3 and 4.

Distribution of tests by zip code and of censored zip codes

The 90th, 97.5th, and 99th percentiles for total tests by zip code in the original data were 125, 784, and 1636 tests, respectively (see Figure 2A). Thus, a small minority of zip codes account for the vast majority of total tests. There were 15 108 (88.7%) unique zip codes in the original data with <100 total tests and 11 039 (64.7%) with

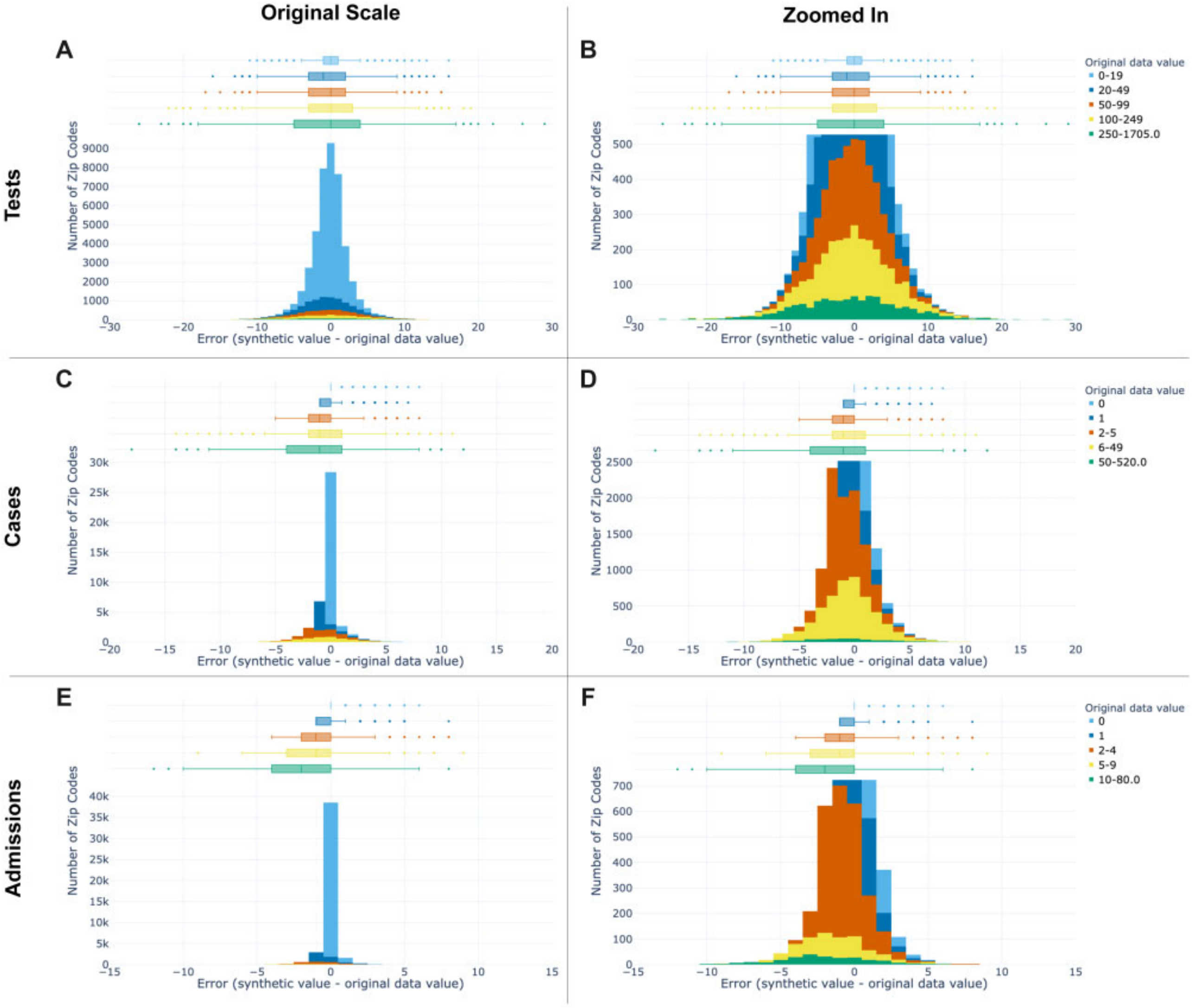

Figure 6. Synthetic error distributions per zip code stratified by month for tests (top row), cases (middle row), and admissions (bottom row) shown both at original scale (left column) and zoomed in to the peak of each row's middle bin (legend showing bin ranges and color encodings seen on the far right of each row). Original data value denotes the monthly count in the original data for the key indicator of interest. Box plots of synthetic error are shown in the top 30% of each sub-plot (A–F), with a histogram of synthetic error shown in the bottom 70%. Within each sub-plot, the box plot and histogram have a shared *x*-axis corresponding to synthetic error and shared bins corresponding to the original data value. The *y*-axis shows the number of zip codes stratified by month (eg, zip code month pairs). Boxes in the box plots span from Q1 to Q3, with median marked inside the box. Fences span ±1.5 times the IQR. Error increased as the size (count) of the original data increased, which allows users to estimate the level of error in their data of interest. The synthetic data systematically underestimate the monthly count of key indicators in zip codes with the most tests, cases, and deaths, and overestimate them in zip codes with the least.

<10 tests. Above this threshold ($n \geq 10$ tests), the synthetic data mimic the original data distribution closely (see Figure 2B). There were 17 041 unique zip codes and 5819 unique uncensored zip codes in the original and synthetic data, respectively. The vast majority of censored zip codes are those that had <10 total tests in the original data (mean = 2.9 ± 2.4; median = 2, IQR = 3; max = 16) as seen in Supplementary Figure S3.

Monthly zip code pairwise synthetic error

The absolute value of pairwise synthetic error stratified by month and zip code increased as the original data value of counts increased (see Figure 6; Supplementary Table S1). Thus, as sample size of data increased, so did the absolute synthetic error and vice versa. The synthetic error for tests ranged from an IQR = 2 when the original value of tests was between 0–19 and IQR = 9 when the original value of tests was between 250 and 1705. All synthetic error for zip codes with an original bin value of zero count was positive. All other bins' synthetic error across key indicators was skewed negative, indicating that the synthetic data had lower counts than the original data.

Post hoc privacy evaluation

In the *post hoc* privacy assessment, 6839 of 1 854 975 rows (0.37%) in the synthetic dataset contained all the same values in all 13 columns as corresponding rows in the original dataset. However, this included numerous values that were null or missing; all but 6 of the 6839 rows included at least 8 missing values among the 13 variables, which greatly mitigates the likelihood of a meaningful identify disclosure, particularly given the vastly larger number of rows compared to columns. In a second run of the synthetic algorithm, none of the 6 rows with fewer than 8 missing values appeared again in the new synthetic derivative, indicating that the initial replication was due to chance rather than individual characteristics of the rows. When the rows in both datasets were grouped into unique combinations of their categorical values (Supplementary Figure S2), groups (or equivalent classes) of individuals with fewer than 10 members existed in the original dataset but did not appear in the synthetic dataset; this is consistent with the censoring algorithm's minimum equivalence class of 10 rows, chosen in conformance with a generally accepted cutoff.[46,47]

DISCUSSION

Overall, analyses on the population-level and of densely tested zip codes (which contained most of the data) were similar between original and synthetically derived datasets. Analyses of sparsely tested populations with smaller sample sizes were notably less similar and had more data suppression, which is in agreement with prior work.[19,32] Synthetic data most closely matched the original data on aggregate data tasks such as aggregate epidemic curves (Figure 1) and broad summary statistics (Table 1). At the aggregate level, only one metric (percent positive, 7-day average) across all indicators showed a significant difference between synthetic and original data aggregate epidemic curves (Table 2). Scarcity of data—as data collection used in this article tapered off in November—is likely a contributing factor to the difference.

The summary statistics shown of both datasets' populations in Table 1 were similar. Major exceptions were the number of unique zip codes due to censoring in the synthetic data and attenuation in the synthetic data of a single extreme outlier (∼−44 000 day LOS) caused by a data quality issue in the original data. Other erroneous negative LOS values persisted within the synthetic data, yet the bulk of the erroneous values remaining were a LOS = −1 which has been reported as a data quality issue attributed to daylight savings.[48,49] Thus, we show that synthetic data can reduce the impact of data quality issues by removing or attenuating erroneous outliers with the aim of protecting the privacy of rare, and thus identifiable, data.

At the zip code and month level, the synthetic data error performed well on an absolute level; the error increased as the size of the original data increased (Figure 5 and Supplementary Table S1). Therefore, the amount of synthetic error is predictable which gives users the ability to estimate the level of error in their data of interest. Additionally, the synthetic error relative to the original data value is likely small enough for most uses of synthetic data. For example, a zip code in the synthetic data with a monthly positive count of 6–49 is off from the original data by an average of -0.59 ± 2.63. The overrepresentation of negative tests in the original data by 8.5-fold (Table 1) appears to bias synthetic error. Since it is impossible to have less than zero count, the synthetic data cannot add privacy-producing noise in the negative direction for zip code monthly counts equal to 0. Consequently, the synthetic data systematically underestimate the monthly count of key indicators in zip codes with the most tests, cases, and deaths, and overestimate them in zip codes with the least. Our results relate to Petti and Flaxman,[12] which observed a similar effect resulting from a non-negativity constraint in the US Census' TopDown differential privacy algorithm. The magnitude of the synthetic error skewing negative in a smaller concentration of zip codes increased as a key indicator became less frequent, which is fundamentally a signal problem in low-density datasets and is not specific to synthetic data generation.

The top 1% most tested zip codes' epidemic curves provide users with 10 qualitative examples of densely tested zip codes. Overall, the synthetic data closely matched the start and end dates of the original data and followed the overall trend of the original data over time (eg, Figure 3A matched spike in late April). The 10 examples show users the 99th percentile best-case scenario of key indicator original data availability and synthetic data performance at the zip code level, yet the size and testing density of N3C data will likely continue to increase.

Our findings show the importance of understanding the characteristics and limitations of the original data since we found these biases affected synthetic data utility. Data biases resulting in poorer performance of software tools, clinical guidelines, and other applications for groups underrepresented in source data have been previously reported for separate tasks.[12,50–53] Foraker et al found that censored zip codes had greater missingness of SDOH values in the original data than uncensored zip codes. In our study, we found the bulk of patients in the N3C data live in a small minority of zip codes (Figure 2), likely those most adjacent to institutions contributing data. These zip codes are therefore more likely to be urban and less likely to have their zip code censored (Table 3). As a consequence, rural zip codes, which are already underrepresented in the original data, become even less available to directly analyze. Additionally, patients with censored zip codes were older, potentially due to older patients traveling from sparsely tested regions to receive care offered at distant academic medical centers which participate in N3C. Traditional de-identification methods would likely censor or suppress zip codes with few tests as well or group them together into higher-level geographic regions. Thus, it is important to view our findings in relation to common alternatives.

While our results demonstrate the utility of using synthetic data for a broad range of geospatial analyses, a caveat to synthetic data use is its utility to analyze rural N3C populations since nearly all zip codes with <10 tests were censored and much more likely to be rural within the original data. Suppression of non-zero counts <10 is a common convention within state and federal guidelines to avoid inadvertent disclosure of protected health information for publicly released data.[46,47,54] Analyses such as choropleth maps at the zip code level including sparsely tested regions would benefit from using the LDS to obtain access to all zip codes without suppression, or by generating and using a different MDClone synthetic dataset that reports geospatial data at a lower level of granularity (eg, 3-digit zip codes). Our results may inform future N3C discussions about dataset balancing ranging from (1) creation of artificially balanced hybrid datasets to improve statistical models' performance on underrepresented data,[50,55] (2) source partners sending a random sample of negative tests alongside all positive tests, or (3) expansion of data ingestion from rural regions.

Whether these synthetic data are "good enough" hinges on a fitness for use determination to be made by each user. The authors believe the data will be useful enough for a wide variety of use cases. Educational software engineering projects or pandemic preparedness tool development could be especially well-served by these data. A major limitation of the data, however, is that they are output in a different data model than the OMOP CDM.[41] Thus, tools built on the synthetic data would not be transferable to run on the LDS without modification. Other users may find the synthetic data well suited to rapid, iterative hypothesis generation/testing without the delays of acquiring the relatively more restricted LDS.[3]

We performed a basic privacy assessment consistent with analyses in other published studies of synthetic data. This *post hoc* assessment demonstrated a lack of matches between the original and synthetic data, indicating that the data in the synthetic dataset do not represent specific individuals from the original dataset. Matches of values across datasets were rare (<0.4% of rows), non-informative (vast majority occurred for rows with sparse data), and random (matches were not duplicated in additional derived datasets). In addition, the absence of unique rows for categorical values demonstrated that value matches with categorical variables, which would be more common, are not unique. These results are not unexpected, given the applied algorithm's approach to generating synthetic data and censoring. While we consider this privacy assessment sufficient for this study, which focuses on demonstrating the utility of a synthetic dataset for analysis, more work can be done in evaluating privacy with synthetic data approaches. We are currently completing an independent and more rigorous evaluation of synthetic data privacy using an adversarial network approach. Additional research is still needed to evaluate synthetic data privacy validation approaches and the actual risk of information gain if variable sets are matched. These issues are beyond the scope of this article but represent the challenges in advancing the use of synthetic data. Others who have studied practical and legal implications of synthetic datasets have recommended their use over de-identified data[56]; this article demonstrates the utility of synthetic data for geographic and temporal analyses, which is a specific functional advantage over Safe Harbor de-identified data.

Limitations and future work

To date, no privacy analysis has been published on these synthetic data to provide context for its utility in relation to its privacy. N3C is currently assessing the privacy of the data used in this study. In a forthcoming manuscript, N3C will be able to quantify the privacy-utility tradeoff of these data through pairing the privacy analysis with these results. Such a full privacy analysis is well beyond the scope of the present paper. However, the methodology described above reflects that the synthetic data process, when computing a derivative dataset for a user-defined patient cohort and selection of properties, is inherently privacy-preserving. While the algorithm maintains the statistical properties—and therefore the utility—of the data, the underlying original data are not visible to the user during the synthesis process. Categorical values are censored, when necessary, to mitigate their inherent exposure to inference attack. The synthetic algorithm is intentionally non-reversible, with multiple layers of protection against privacy attack. The mathematical calculation of alternate matrices is based on Euclidean distance, which is not simply "straight-line" distance, but rather the shortest path through the matrix and is by nature non-reversible. The size of the output population is modified slightly, without altering the statistical model, to further thwart potential attack.

The data used in this article do not reflect the current size nor state of the N3C LDS. Other statistical techniques such as equivalence testing, bhattacharyya distance,[57,58] or adversarial challenges[28] could be used in the future to compare similarity between epidemic curves. The Wilcoxon signed-rank and paired *t*-tests assume the null hypothesis that the original and synthetic datasets are equivalent. Equivalence testing, which flips the null hypothesis, may be better suited. Equivalence testing was not used in this manuscript due to the challenge of selecting an equivalence bound without knowing what threshold(s) data end-users would find most applicable. Additionally, adjustments for multiple testing were not made for differences between synthetic and original epidemic curves. Had they been, no P-values would be <.05. Future work conducting equivalence testing specific to well-defined, high-impact use cases may be merited. However, the work required to do so in an *ad hoc* manner may suggest the LDS is a better alternative in those cases. In future work, the effect of data quality on synthetic data may be worth studying through generation of synthetic data at each cycle of iterative data quality improvement.

CONCLUSION

Overall, the synthetic data are promising for a wide range of use cases including: population-level summary statistics, epidemic curves for the data in aggregate and for the most densely tested zip codes, and analyses necessitating monthly counts of key indicators for the top third of zip codes by number of tests. However, analyses requiring unsuppressed zip code analyses on populations with <10 tests may be better served by the LDS. Biases found in the original data—namely an underrepresentation of positive tests and tests in rural zip codes—were reflected in the synthetic data. Therefore, it is important to understand the limitations and biases of the original data in addition to the synthetic data impacted downstream from it. We expect the user base of N3C synthetic data to be heterogeneous and the use cases of the data to be broad, resulting in a wide range of fitness for use definitions. To date, there is no published evaluation that quantifies the privacy afforded by this synthetic dataset specifically—nor of the MDClone system itself broadly—to contextualize this synthetic dataset's utility in relation to a privacy-utility tradeoff; such evaluations are beyond the scope of this work. Future privacy evaluations of MDClone will not necessarily reflect the privacy of the synthetic data analyzed in this study unless the same dataset and/or the same MDClone system version and parameters are evaluated. Our evaluation of the N3C synthetic data utility provides users the ability to assess whether the synthetic data are fit for use through its combination of general-purpose data utility assessments and visualized replications of analyses of common interest.

FUNDING

The analyses described in this publication were conducted with data or tools accessed through the NCATS N3C Data Enclave covid.cd2h.org/enclave and supported by NCATS U24 TR002306. This research was possible because of the patients whose information is included within the data from participating organizations (covid.cd2h.org/dtas) and the organizations and scientists (covid.cd2h.org/duas) who have contributed to the on-going development of this community resource.

The N3C data transfer to NCATS is performed under a Johns Hopkins University Reliance Protocol # IRB00249128 or individual site agreements with NIH. The N3C Data Enclave is managed under the authority of the NIH; information can be found at https://ncats.nih.gov/n3c/resources.

Data were provided from the following institutions: Stony Brook University—U24TR002306 • University of Oklahoma Health Sciences Center—U54GM104938: Oklahoma Clinical and Translational Science Institute (OCTSI) • West Virginia University—U54GM104942: West Virginia Clinical and Translational Science Institute (WVCTSI) • University of Mississippi Medical Center—U54GM115428: Mississippi Center for Clinical and Translational Research (CCTR) • University of Nebraska Medical Center—U54GM115458: Great Plains IDeA-Clinical & Translational Research • Maine Medical Center—U54GM115516: Northern New England Clinical & Translational Research (NNE-CTR) Network • Wake Forest University Health Sciences—UL1TR001420: Wake Forest Clinical and Translational

Science Institute • Northwestern University at Chicago—UL1TR001422: Northwestern University Clinical and Translational Science Institute (NUCATS) • University of Cincinnati—UL1TR001425: Center for Clinical and Translational Science and Training • The University of Texas Medical Branch at Galveston—UL1TR001439: The Institute for Translational Sciences • Medical University of South Carolina—UL1TR001450: South Carolina Clinical & Translational Research Institute (SCTR) • University of Massachusetts Medical School Worcester—UL1TR001453: The UMass Center for Clinical and Translational Science (UMCCTS) • University of Southern California—UL1TR001855: The Southern California Clinical and Translational Science Institute (SC CTSI) • Columbia University Irving Medical Center—UL1TR001873: Irving Institute for Clinical and Translational Research • George Washington Children's Research Institute—UL1TR001876: Clinical and Translational Science Institute at Children's National (CTSA-CN) • University of Kentucky—UL1TR001998: UK Center for Clinical and Translational Science • University of Rochester—UL1TR002001: UR Clinical & Translational Science Institute • University of Illinois at Chicago—UL1TR002003: UIC Center for Clinical and Translational Science • Penn State Health Milton S. Hershey Medical Center—UL1TR002014: Penn State Clinical and Translational Science Institute • The University of Michigan at Ann Arbor—UL1TR002240: Michigan Institute for Clinical and Health Research • Vanderbilt University Medical Center—UL1TR002243: Vanderbilt Institute for Clinical and Translational Research • University of Washington—UL1TR002319: Institute of Translational Health Sciences • Washington University in St. Louis—UL1TR002345: Institute of Clinical and Translational Sciences • Oregon Health & Science University—UL1TR002369: Oregon Clinical and Translational Research Institute • University of Wisconsin-Madison—UL1TR002373: UW Institute for Clinical and Translational Research • Rush University Medical Center—UL1TR002389: The Institute for Translational Medicine (ITM) • The University of Chicago—UL1TR002389: The Institute for Translational Medicine (ITM) • University of North Carolina at Chapel Hill—UL1TR002489: North Carolina Translational and Clinical Science Institute • University of Minnesota—UL1TR002494: Clinical and Translational Science Institute • Children's Hospital Colorado—UL1TR002535: Colorado Clinical and Translational Sciences Institute • The University of Iowa—UL1TR002537: Institute for Clinical and Translational Science • The University of Utah—UL1TR002538: Uhealth Center for Clinical and Translational Science • Tufts Medical Center—UL1TR002544: Tufts Clinical and Translational Science Institute • Duke University—UL1TR002553: Duke Clinical and Translational Science Institute • Virginia Commonwealth University—UL1TR002649: C. Kenneth and Dianne Wright Center for Clinical and Translational Research • The Ohio State University—UL1TR002733: Center for Clinical and Translational Science • The University of Miami Leonard M. Miller School of Medicine—UL1TR002736: University of Miami Clinical and Translational Science Institute • University of Virginia—UL1TR003015: iTHRIVL Integrated Translational health Research Institute of Virginia • Carilion Clinic—UL1TR003015: iTHRIVL Integrated Translational health Research Institute of Virginia • University of Alabama at Birmingham—UL1TR003096: Center for Clinical and Translational Science • Johns Hopkins University—UL1TR003098: Johns Hopkins Institute for Clinical and Translational Research • University of Arkansas for Medical Sciences—UL1TR003107: UAMS Translational Research Institute • Nemours—U54GM104941: Delaware CTR ACCEL Program • University Medical Center New Orleans—U54GM104940: Louisiana Clinical and Translational Science (LA CaTS) Center • University of Colorado Denver, Anschutz Medical Campus—UL1TR002535: Colorado Clinical and Translational Sciences Institute • Mayo Clinic Rochester—UL1TR002377: Mayo Clinic Center for Clinical and Translational Science (CCaTS) • Tulane University—UL1TR003096: Center for Clinical and Translational Science • Loyola University Medical Center—UL1TR002389: The Institute for Translational Medicine (ITM) • Advocate Health Care Network—UL1TR002389: The Institute for Translational Medicine (ITM) • OCHIN—INV-018455: Bill and Melinda Gates Foundation grant to Sage Bionetworks.

Additional data partners who have signed DTA and data release pending: The Rockefeller University—UL1TR001866: Center for Clinical and Translational Science • The Scripps Research Institute—UL1TR002550: Scripps Research Translational Institute • University of Texas Health Science Center at San Antonio—UL1TR002645: Institute for Integration of Medicine and Science • The University of Texas Health Science Center at Houston—UL1TR003167: Center for Clinical and Translational Sciences (CCTS) • NorthShore University HealthSystem—UL1TR002389: The Institute for Translational Medicine (ITM) • Yale New Haven Hospital—UL1TR001863: Yale Center for Clinical Investigation • Emory University—UL1TR002378: Georgia Clinical and Translational Science Alliance • Weill Medical College of Cornell University—UL1TR002384: Weill Cornell Medicine Clinical and Translational Science Center • Montefiore Medical Center—UL1TR002556: Institute for Clinical and Translational Research at Einstein and Montefiore • Medical College of Wisconsin—UL1TR001436: Clinical and Translational Science Institute of Southeast Wisconsin • University of New Mexico Health Sciences Center—UL1TR001449: University of New Mexico Clinical and Translational Science Center • George Washington University—UL1TR001876: Clinical and Translational Science Institute at Children's National (CTSA-CN) • Stanford University—UL1TR003142: Spectrum: The Stanford Center for Clinical and Translational Research and Education • Regenstrief Institute—UL1TR002529: Indiana Clinical and Translational Science Institute • Cincinnati Children's Hospital Medical Center—UL1TR001425: Center for Clinical and Translational Science and Training • Boston University Medical Campus—UL1TR001430: Boston University Clinical and Translational Science Institute • The State University of New York at Buffalo—UL1TR001412: Clinical and Translational Science Institute • Aurora Health Care—UL1TR002373: Wisconsin Network For Health Research • Brown University—U54GM115677: Advance Clinical Translational Research (Advance-CTR) • Rutgers, The State University of New Jersey—UL1TR003017: New Jersey Alliance for Clinical and Translational Science • Loyola University Chicago—UL1TR002389: The Institute for Translational Medicine (ITM) • New York University—UL1TR001445: Langone Health's Clinical and Translational Science Institute • Children's Hospital of Philadelphia—UL1TR001878: Institute for Translational Medicine and Therapeutics • University of Kansas Medical Center—UL1TR002366: Frontiers: University of Kansas Clinical and Translational Science Institute • Massachusetts General Brigham—UL1TR002541: Harvard Catalyst • Icahn School of Medicine at Mount Sinai—UL1TR001433: ConduITS Institute for Translational Sciences • Ochsner Medical Center—U54GM104940: Louisiana Clinical and Translational Science (LA CaTS) Center • HonorHealth—None (Voluntary) • University of California, Irvine—UL1TR001414: The UC Irvine Institute for Clinical and Translational Science (ICTS) • University of California, San Diego—UL1TR001442: Altman Clinical and Translational Research Institute • University of California, Davis—UL1TR001860: UCDavis Health Clinical and Translational Science Center • University of California, San Francisco—UL1TR001872: UCSF Clinical and Translational Science Institute • University of California, Los Angeles—UL1TR001881: UCLA Clinical Translational Science Institute • University of Vermont—U54GM115516: Northern New England Clinical & Translational Research (NNE-CTR) Network • Arkansas Children's Hospital—UL1TR003107: UAMS Translational Research Institute.

Authorship was determined using ICMJE recommendations. The project described was supported by the National Institute of General Medical Sciences, 5U54GM104942-04. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. Any opinions expressed in this document are those of the authors and do not necessarily reflect the views of NCATS, individual N3C team members, or affiliated organizations and institutions.

AUTHOR CONTRIBUTIONS

Masthead authors: ABW, JAT, NZ, and REF contributed to study conception and design. NZ contributed to the generation of the data. JAT conducted the experiment and data analysis. JDM contributed in part to the interpretation of the results. JAT wrote the manuscript with input from all authors, and JDM wrote portions of the Methods (section "Synthetic data") and Discussion (a portion of the privacy limitations discussed in "Limitations and future directions"). ABW and REF led the N3C Synthetic Data Validation Task

Team with support from PROP who led the broader N3C Synthetic Data Workstream.

Post hoc privacy evaluation: Methods design and results interpretation were performed by the Washington University team (ABW, REF, and PROP). NZ wrote and ran the code/queries to generate the privacy results. ABW, JDM, and NZ wrote the privacy evaluation methods and results section with input from all authors. ABW wrote the privacy evaluation discussion section with input from all authors.

*Other consortial authors**: Christopher G. Chute[1–10] and Melissa A. Haendel[6,10–12]

[1]Clinical data model expertise, [2]data curation, [3]data integration, [4]data quality assurance, [5]funding acquisition, [6]governance, [7]critical revision of the manuscript for important intellectual content, [8]N3C Phenotype definition, [9]project evaluation, [10]project management, [11]regulatory oversight/admin, [12]funding acquisition.

*Consortial authorship and corresponding contributions were self-reported as part of the N3C authorship committee review process.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

ACKNOWLEDGMENTS

We gratefully acknowledge contributions from the following N3C core teams (leads noted with *).

- Principal Investigators: Melissa A. Haendel*, Christopher G. Chute*, Kenneth R. Gersing, Anita Walden.
- Workstream, subgroup and administrative leaders: Melissa A. Haendel*, Tellen D. Bennett, Christopher G. Chute, David A. Eichmann, Justin Guinney, Warren A. Kibbe, Hongfang Liu, Philip R.O. Payne, Emily R. Pfaff, Peter N. Robinson, Joel H. Saltz, Heidi Spratt, Justin Starren, Christine Suver, Adam B. Wilcox, Andrew E. Williams, Chunlei Wu.
- Key liaisons at data partner sites.
- Regulatory staff at data partner sites.
- Individuals at the sites who are responsible for creating the datasets and submitting data to N3C.
- Data Ingest and Harmonization Team: Christopher G. Chute*, Emily R. Pfaff*, Davera Gabriel, Stephanie S. Hong, Kristin Kostka, Harold P. Lehmann, Richard A. Moffitt, Michele Morris, Matvey B. Palchuk, Xiaohan Tanner Zhang, Richard L. Zhu.
- Phenotype Team (Individuals who create the scripts that the sites use to submit their data, based on the COVID and Long COVID definitions): Emily R. Pfaff*, Benjamin Amor, Mark M. Bissell, Marshall Clark, Andrew T. Girvin, Stephanie S. Hong, Kristin Kostka, Adam M. Lee, Robert T. Miller, Michele Morris, Matvey B. Palchuk, Kellie M. Walters.
- Project Management and Operations Team: Anita Walden*, Yooree Chae, Connor Cook, Alexandra Dest, Racquel R. Dietz, Thomas Dillon, Patricia A. Francis, Rafael Fuentes, Alexis Graves, Julie A. McMurry, Andrew J. Neumann, Shawn T. O'Neil, Usman Sheikh, Andréa M. Volz, Elizabeth Zampino.
- Partners from NIH and other federal agencies: Christopher P. Austin*, Kenneth R. Gersing*, Samuel Bozzette, Mariam Deacy, Nicole Garbarini, Michael G. Kurilla, Sam G. Michael, Joni L. Rutter, Meredith Temple-O'Connor.
- Analytics Team (Individuals who build the Enclave infrastructure, help create codesets, variables, and help Domain Teams and project teams with their datasets): Benjamin Amor*, Mark M. Bissell, Katie Rebecca Bradwell, Andrew T. Girvin, Amin Manna, Nabeel Qureshi.
- Publication Committee Management Team: Mary Morrison Saltz*, Christine Suver*, Christopher G. Chute, Melissa A. Haendel, Julie A. McMurry, Andréa M. Volz, Anita Walden.
- Publication Committee Review Team: Carolyn Bramante, Jeremy Richard Harper, Wenndy Hernandez, Farrukh M Koraishy, Federico Mariona, Saidulu Mattapally, Amit Saha, Satyanarayana Vedula.
- Synthetic Data Domain Team including Yujuan Fu, Nisha Mathews, Ofer Mendelevitch.

CONFLICT OF INTEREST STATEMENT

All authors have completed the ICMJE uniform disclosure form at http://www.icmje.org/downloads/coi_disclosure.docx and declare: Authors JAT, ABW, REF, and PROP received financial support from the National Center for Advancing Translational Sciences, National Institutes of Health, through grant number U24TR002306 disbursed to their affiliated institutions for the submitted work; authors JDM and NZ are employees of MDClone; this manuscript underwent National Covid Cohort Collaborative (N3C) publication review described at https://covid.cd2h.org/publication-review; the institution RF and PROP are affiliated with (Washington University in St. Louis) is a customer of MDClone; author JAT became an employee of Philips North America LLC during the 3rd round of manuscript revisions; all authors declare no other relationships or activities that could appear to have influenced the submitted work.

HUMAN SUBJECTS PROTECTIONS

This study was approved by the Washington University and University of Washington Internal Review Boards.

DATA AVAILABILITY

The entirety of code used in this analysis is contained within a single Palantir Foundry Code Workbook using a saved Spark environment to preserve required software versions and dependencies. The code workbook and source data have been stored within the National Covid Cohort Collaborative (N3C) enclave (https://covid.cd2h.org/enclave) so that they may inform and be reused in future validation work. To view National Covid Cohort Collaborative (N3C) Data Enclave & Data Access Requirements, please navigate to the N3C website.

REFERENCES

1. Azzopardi-Muscat N, Kluge HHP, Asma S, *et al.* A call to strengthen data in response to COVID-19 and beyond. *J Am Med Inform Assoc* 2021; 28 (3): 638–9.
2. Subbian V, Solomonides A, Clarkson M, *et al.* Ethics and informatics in the age of COVID-19: challenges and recommendations for public health organization and public policy. *J Am Med Inform Assoc* 2021; 28 (1): 184–9.
3. Haendel MA, Chute CG, Bennett TD, *et al.*; N3C Consortium. The National COVID Cohort Collaborative (N3C): rationale, design, infrastructure, and deployment. *J Am Med Inform Assoc* 2021; 28 (3): 427–43.
4. The National COVID Cohort Collaborative: Clinical Characterization and Early Severity Prediction | medRxiv. https://www.medrxiv.org/content/10.1101/2021.01.12.21249511v3 Accessed March 1, 2021.
5. HIPAA Privacy Rule and its Impacts on Research. https://privacyruleandresearch.nih.gov/pr_08.asp Accessed March 17, 2021.
6. CFR 164.514—Other Requirements Relating to Uses and Disclosures of Protected Health Information. Content Details—CFR-2011-title45-vol1-sec164-514. https://www.govinfo.gov/app/details/CFR-2011-title45-vol1/CFR-2011-title45-vol1-part164 Accessed March 17, 2021.
7. Raab GM, Nowok B, Dibben C. Guidelines for Producing Useful Synthetic Data. *arXiv:171204078 [stat]* Published Online First: 11 December 2017.http://arxiv.org/abs/1712.04078 Accessed March 17, 2021.
8. Snoke J, Raab GM, Nowok B, *et al.* General and specific utility measures for synthetic data. *J R Stat Soc A* 2018; 181 (3): 663–88.
9. Mukherjee S, Xu Y, Trivedi A, *et al.* privGAN: Protecting GANs from membership inference attacks at low cost. *arXiv:200100071 [cs, stat]* Published Online First: 13 December 2020. http://arxiv.org/abs/2001.00071 Accessed March 17, 2021.

10. Beaulieu-Jones BK, Wu ZS, Williams C, *et al.* Privacy-preserving generative deep neural networks support clinical data sharing. *Circ Cardiovasc Qual Outcomes* 2019; 12 (7): e005122.
11. Foraker R, Mann DL, Payne PRO. Are synthetic data derivatives the future of translational medicine? *JACC Basic Transl Sci* 2018; 3 (5): 716–8.
12. Petti S, Flaxman A. Differential privacy in the 2020 US census: what will it do? Quantifying the accuracy/privacy tradeoff. *Gates Open Res* 2019; 3: 1722.
13. Price WN, Cohen IG. Privacy in the age of medical big data. *Nat Med* 2019; 25 (1): 37–43.
14. Wu L, He H, Zaïane OR. Utility of privacy preservation for health data publishing. In: *Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems*. 2013: 510–1; Porto, Portugal. doi:10.1109/CBMS.2013.6627853.
15. Muniz-Terrera G, Mendelevitch O, Barnes R, *et al.* Virtual cohorts and synthetic data in dementia: an illustration of their potential to advance research. *Front Artif Intell* 2021; 4: 613956.
16. CDC. Transitioning from CDC's Indicators for Dynamic School Decision-Making (released September 15, 2020) to CDC's Operational Strategy for K-12 Schools through Phased Mitigation (released February 12, 2021) to Reduce COVID-19. Centers for Disease Control and Prevention. 2020. https://www.cdc.gov/coronavirus/2019-ncov/community/schools-childcare/indicators.html Accessed March 21, 2021.
17. CDC. Operational Strategy for K-12 Schools through Phased Mitigation. Centers for Disease Control and Prevention. 2020. https://www.cdc.gov/coronavirus/2019-ncov/community/schools-childcare/operation-strategy.html Accessed February 16, 2021.
18. State-By-State Summary of Public Health Criteria in Reopening Plans. National Governors Association. https://www.nga.org/coronavirus-reopening-plans/ Accessed March 21, 2021.
19. Benaim AR, Almog R, Gorelik Y, *et al.* Analyzing medical research results based on synthetic data and their relation to real data results: systematic comparison from five observational studies. *JMIR Med Inform* 2020; 8 (2): e16492.
20. Zhang Z, Yan C, Mesa DA, *et al.* Ensuring electronic medical record simulation through better training, modeling, and evaluation. *J Am Med Inform Assoc* 2020; 27 (1): 99–108.
21. Teixeira da Silva JA, Tsigaris P, Erfanmanesh M. Publishing volumes in major databases related to Covid-19. *Scientometrics* 2021; 126 (1): 831–42.
22. Guerrini CJ, Majumder MA, Lewellyn MJ, *et al.* Citizen science, public policy. *Science* 2018; 361 (6398): 134–6.
23. Katapally TR. A global digital citizen science policy to tackle pandemics like COVID-19. *J Med Internet Res* 2020; 22 (5): e19357.
24. Roche J, Bell L, Galvão C, *et al.* Citizen science, education, and learning: challenges and opportunities. *Front Sociol* 2020; 5: 613814.
25. Juran JM, Godfrey AB, eds. *Juran's Quality Handbook*. 5th ed. New York: McGraw Hill; 1999.
26. Chen J, Chun D, Patel M, *et al.* The validity of synthetic clinical data: a validation study of a leading synthetic data generator (Synthea) using clinical quality measures. *BMC Med Inform Decis Mak* 2019; 19 (1): 44.
27. Foraker RE, Yu SC, Gupta A, *et al.* Spot the difference: comparing results of analyses from real patient data and synthetic derivatives. *JAMIA Open* 2020; 3 (4): 557–66.
28. El Emam K, Mosquera L, Jonker E, *et al.* Evaluating the utility of synthetic COVID-19 case data. *JAMIA Open* 2021; 4 (1): ooab012.
29. Wang Z, Myles P, Tucker A. Generating and evaluating cross-sectional synthetic electronic healthcare data: preserving data utility and patient privacy. *Comput Intell* 2021; 37 (2): 819–51.
30. Hittmeir M, Ekelhart A, Mayer R. On the utility of synthetic data: an empirical evaluation on machine learning tasks. In: *Proceedings of the 14th International Conference on Availability, Reliability and Security*. Canterbury, CA, UK: Association for Computing Machinery; 2019: 1–6. doi:10.1145/3339252.3339281.
31. Emam KE. Seven ways to evaluate the utility of synthetic data. *IEEE Secur Priv* 2020; 18: 56–9.
32. Foraker R, Guo A, Thomas J, *et al.*; N3C Collaborative. The national COVID cohort collaborative: analyses of original and computationally derived electronic health record data. *J Med Internet Res* 2021; 23 (10): e30697.
33. CDC. COVID Data Tracker. Centers for Disease Control and Prevention. 2020. https://covid.cdc.gov/covid-data-tracker Accessed March 23, 2021.
34. Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. *Lancet Infect Dis* 2020; 20 (5): 533–4.
35. Roser M, Ritchie H, Ortiz-Ospina E, *et al.* Coronavirus pandemic (COVID-19). *Our World in Data*. Published Online First: 5 March 2020. https://ourworldindata.org/coronavirus Accessed March 23, 2021.
36. Vaidya J, Shafiq B, Jiang X, *et al.* Identifying inference attacks against healthcare data repositories. *AMIA Jt Summits Transl Sci Proc* 2013; 2013: 262–6.
37. Sweeney L. k-anonymity: a model for protecting privacy. *Int J Uncertain Fuzziness Knowl Based Syst* 2002; 10 (05): 557–70.
38. Emam KE, Mosquera L, Bass J. Evaluating identity disclosure risk in fully synthetic health data: model development and validation. *J Med Internet Res* 2020; 22 (11): e23139.
39. Erez L. United States Patent: 10977388—Computer system of computer servers and dedicated computer clients specially programmed to generate synthetic non-reversible electronic data records based on real-time electronic querying and methods of use thereof. 2021. https://patft.uspto.gov/netacgi/nph-Parser?Sect1=PTO1&Sect2=HITOFF&p=1&u=/netahtml/PTO/srchnum.html&r=1&f=G&l=50&d=PALL&s1=10977388.PN Accessed December 3, 2021.
40. Francis P, Probst Eide S, Munz R, Diffix: high-utility database anonymization. In: Schweighofer E, Leitold H, Mitrakas A, *et al.*, eds. *Privacy Technologies and Policy*. Cham: Springer International Publishing; 2017: 141–58. doi:10.1007/978-3-319-67280-9_8.
41. Observational Health Data Sciences and Informatics. OMOP CDM v5.3.1. https://ohdsi.github.io/CommonDataModel/cdm531.html Accessed March 26, 2021.
42. Waskom M. Team the seaborn development. mwaskom/seaborn. Zenodo 2020. doi:10.5281/zenodo.592845.
43. Choosing color palettes—seaborn 0.11.1 documentation. https://seaborn.pydata.org/tutorial/color_palettes.html Accessed March 24, 2021.
44. Jenny B, Kelso NV. Color oracle. *Color Oracle: Design for the Color Impaired*. 2011. https://colororacle.org/index.html
45. Fort D, Cimino J, Wilcox A. *Every Needle in a Haystack: Finding Fingerprints in a Safe Harbor Dataset Using a Single Common Lab Test*. San Francisco, CA, USA: AMIA Summit on Clinical Research Informatics; 2012.
46. Washington State Department of Health. Guidelines for Working with Small Numbers. 2001. https://doh.wa.gov/data-statistical-reports/data-guidelines
47. McCallister E, Grance T, Scarfone K. *Guide to Protecting the Confidentiality of Personally Identifiable Information (PII): Recommendations of the National Institute of Standards and Technology*. Special Publication 800-122, National Institute of Standards and Technology. 2010.
48. Ehlers A, Dyson RL, Hodgson CK, *et al.* Impact of daylight saving time on the clinical laboratory. *Acad Pathol* 2018; 5: 2374289518784222.doi:10.1177/2374289518784222
49. Thomas JA, Wilcox AB, Joo EJ. Readmissions: Data Quality and Prediction. *2018 National Library of Medicine Training Conference—Poster*. Published Online First: 2018. https://osf.io/vk65x/ Accessed April 15, 2021.
50. Adamson AS, Smith A. Machine learning and health care disparities in dermatology. *JAMA Dermatol* 2018; 154 (11): 1247–8.
51. Gijsberts CM, Groenewegen KA, Hoefer IE, *et al.* Race/ethnic differences in the associations of the framingham risk factors with carotid IMT and cardiovascular events. *PLoS One* 2015; 10 (7): e0132321.
52. Grother P, Ngan M, Hanaoka K. *Face Recognition Vendor Test Part 3: Demographic Effects*. Gaithersburg, MD: National Institute of Standards and Technology; 2019. doi:10.6028/NIST.IR.8280.
53. Kessler MD, Yerges-Armstrong L, Taub MA, *et al.*; Consortium on Asthma among African-ancestry Populations in the Americas (CAAPA). Challenges and disparities in the application of personalized genomic medicine to populations with African ancestry. *Nat Commun* 2016; 7: 12521.

54. Klein R, Proctor S, Boudreault M, *et al.* Healthy people 2010 criteria for data suppression. *Healthy People 2010 Stat Notes* 2002; 24: 1–12.

55. Ghorbani A, Natarajan V, Coz D, *et al.* DermGAN: synthetic generation of clinical skin images with pathology. In: *Machine Learning for Health Workshop*. PMLR; 2020: 155–70. http://proceedings.mlr.press/v116/ghorbani20a.html Accessed March 29, 2021.

56. Bellovin SM, Dutta PK, Reitinger N. Privacy and synthetic datasets. *Stan Tech L Rev* 2019; 22: 1–52.

57. Comaniciu D, Ramesh V, Meer P. Real-time tracking of non-rigid objects using mean shift. In: *Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No. PR00662)*, Vol. 2. 2000: 142–9; Hilton Head, SC. doi:10.1109/CVPR.2000.854761.

58. Kaloskampis I, Joshi C, Cheung C, *et al.* Synthetic data in the civil service. *Significance* 2020; 17 (6): 18–23.