Evaluating Synthea: Comprehensive Analysis of a Leading Synthesized Medical Record Generator

Zach Wagner[1], Clair Blacketer[2,3]
[1]Grassfield High School, Chesapeake, VA, [2]Janssen Research& Development, Raritan, NJ,
[3]Department of Medical Informatics, Erasmus, Rotterdam, NL

Background

Observational healthcare research has grown significantly due to the availability of electronic healthcare record (EHR) data, aiding in policy, healthcare delivery, procedural advancements, and outbreak responses [1]. However, accessing EHR data is challenging due to privacy regulations like the Health Insurance Portability and Accountability Act (HIPAA) in the US and General Data Protection Regulation (GDPR) in the EU. Synthetic data—simulating real populations using AI and ML models—emerges as a promising solution to bypass patient privacy concerns and expedite research [2].

Developed by MITRE in 2017, Synthea is an open-source generator of synthesized EHRs for research and innovation without legal or privacy restrictions. According to the developers, it provides state-level patient information, mimicking real populations in geographic distributions, disease rates, doctor visits, hospitalizations, drug usages, and social demographics. Unlike other methods, Synthea avoids re-identification risks and does not rely on real patient data [3]. Despite its promise, further research is needed to fully understand and improve Synthea's capabilities.

Despite its potential, Synthea faces limitations. Hodges et al. found Synthea's data unreliable without external modification, like the "Medication Diversification Tool" (MDT) for realistic medication distributions [4]. The lack of diversity in Synthea’s modeling capacities hinders its fidelity. Further analysis is required to validate Synthea’s capabilities, especially in tracking chronic diseases, which are significant healthcare concerns. This study aims to expand on prior research by comparing Synthea data generated to emulate the population of California (CA) to real-world data reported by the state.

Methods

A 1,162,848 person sample of Synthea data was generated using version 2.7 of the tool[1]. The only parameter given at the time of generation was that the patients should all be

[1] https://github.com/synthetichealth/synthea

modeled from the state of California. The sample was then converted to the Observational Medical Outcomes Partnership Common Data Model (CDM) using the ETL-Synthea R package version 1.0[2]. General database characterizations were generated using the Achilles R package version 1.7[3].

To compare Synthea-generated data with real California data, we examined demographic breakdowns, hospitalization rates, and chronic disease prevalence rates. Demographic information for the state was obtained from the United States Census Bureau Comparative Demographic Estimates[5]. Hospitalizations were obtained from the California Department of Health Care Access and Information, Healthcare Analytics Branch[6]. Prevalence rates for Coronary Heart Disease (CHD) and Hypertension were obtained from the CDC's Interactive Atlas for Heart Disease and Stroke[7].

For all comparisons, the standardized mean difference (SMD) was used to measure the likeness of the Synthea values compared to the real-world reported California values. Lower SMD indicates closer similarity between synthetic and real values. SMD ranges were defined as follows: 0%-8% for very close similarity, 8%-30% for moderate similarity, and greater than 30% for poor similarity.

Demographic and Country Representation: We compared the age, gender, race, and ethnicity breakdowns between the synthetic and real-world data using Achilles analysis IDs 2, 3, 4 and 5 and the overall census data for the state.

Hospitalizations by Age Group: We compared hospitalization rates across all available persons, stratified year and gender. The percentage of inpatient visits observed per year in the synthetic data was calculated by dividing the number of hospitalizations by the total number of people with observation during the year. These were compared to the overall hospitalization rates reported by the California Department of Health Care Access and Information.

Prevalence of Chronic Disease: The final step was to compare the prevalence rates for various two different chronic diseases at the county level. We identified the ICD10CM codes used in CDC's Interactive Atlas for Heart Disease and Stroke to define CHD and Hypertension. These source codes were then mapped to standard concepts, and we calculated the number of adults in the synthetic sample aged >18 in 2021 with at least one record of any of the identified standard concepts, following the CDC's methods and stratified by county. These counts were then divided by the total adults in each county to get a crude prevalence rate for CHD and Hypertension in the Synthea data. The crude rates were then compared to the rates reported by the CDC for California.

[2] https://github.com/ohdsi/ETL-Synthea

[3] https://github.com/ohdsi/Achilles

Results

Demographic Data Comparison: The demographic data comparison between real and synthetic data showed high accuracy in gender comparisons. Figure 1 highlights the synthetic generation of characteristics for ethnicity, age, race, and gender compared to the real California population. The SMD for males and females was 1.94%, indicating very close similarity. Birth year data also tracked closely, except for the oldest age group, which had a high SMD of 116.96%. Excluding this group, the average SMD was 7.76%, suggesting very close similarity. Race comparisons showed moderate consistency with close similarity for Black (1.39%) and Asian (5.18%) populations, but discrepancies for White (34.13%) and other populations (53.20%). Ethnicity comparisons were highly comparable, with an SMD of 2.76% for both Hispanic or Latino and not Hispanic or Latino categories.

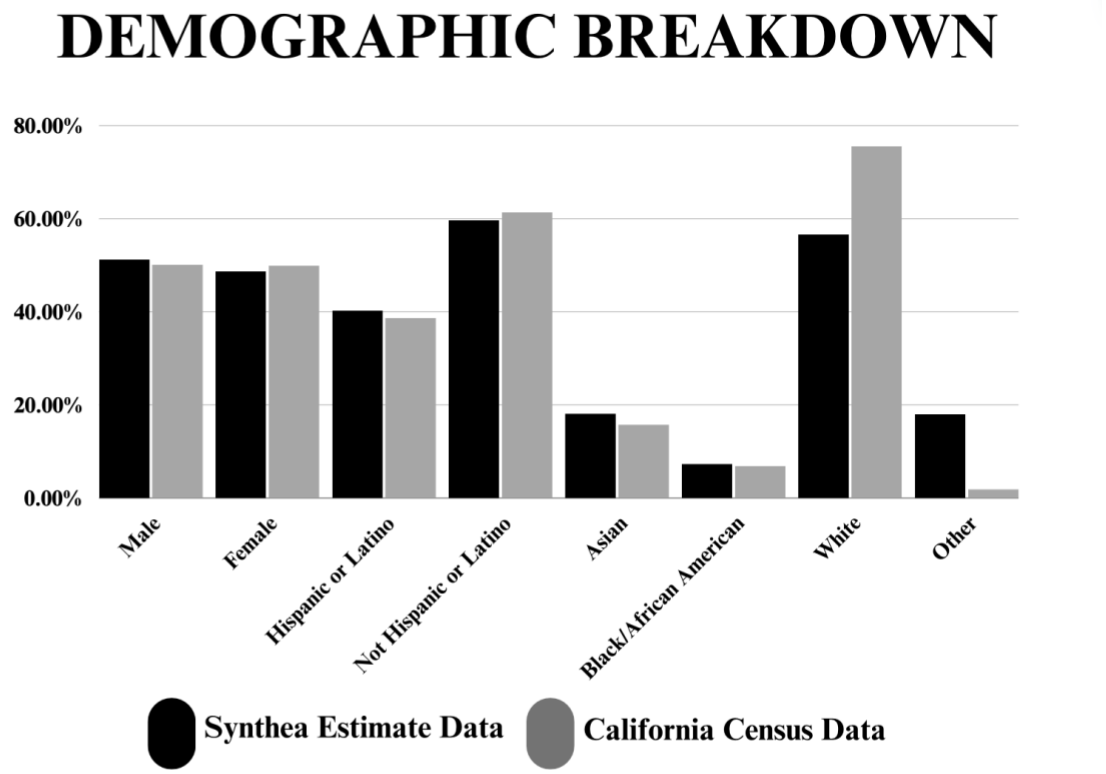

Figure 1. Demographic comparison between Synthea and the California census Data

Hospitalization Data Comparison: The distribution of hospitalization rates across gender and the overall population was compared and detailed in Figure 2. The distribution of inpatient visits stratified by gender showed fairly accurate results, with an average SMD of 9.64% for female visits. Overall hospitalization rates across the two data pools also showed moderate similarities. Notable differences were seen in 2020, with a difference of over 10 percentage points. Excluding 2020, the average SMD was 9.21%, indicating moderate similarity. For 2020, the SMD was poor at 33.64%.

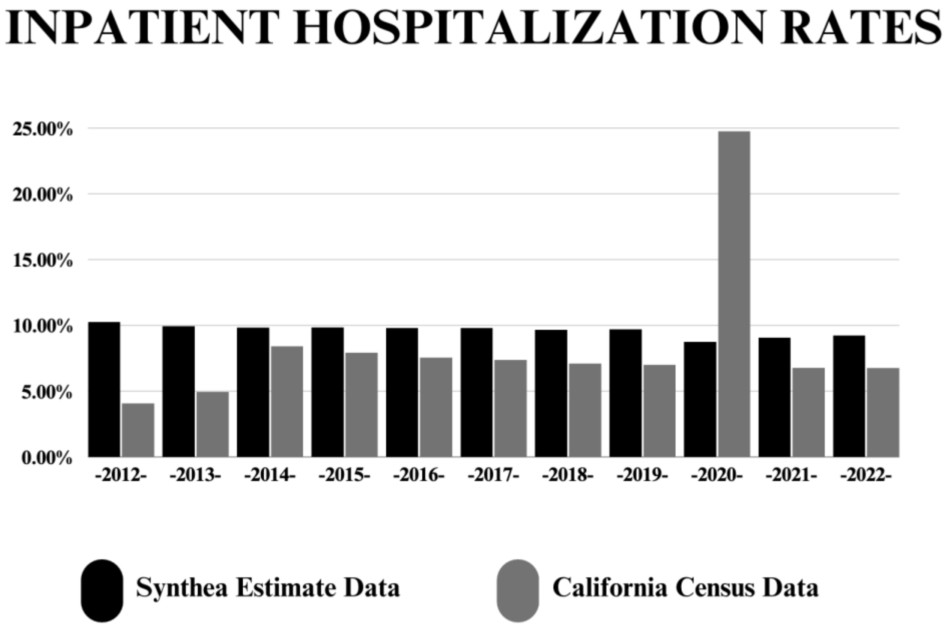

Figure 2. Comparison of hospitalization rates between Synthea and real-world California data, stratified by year

Disease Prevalence Data Comparison: The comparability of chronic disease prevalence rates was the most pertinent result. Figure 3 reveals the percent differences between data from California reports and those generated by Synthea. Synthea underestimated the actual values, but the SMD between most counties was not excessively high. The average SMD for Coronary Heart Disease (CHD) prevalence rates indicated very high similarity at 3.59%. Hypertension modeling showed moderate accuracy with an average SMD of 8.9%, suggesting moderate similarity despite some large gaps in certain counties.

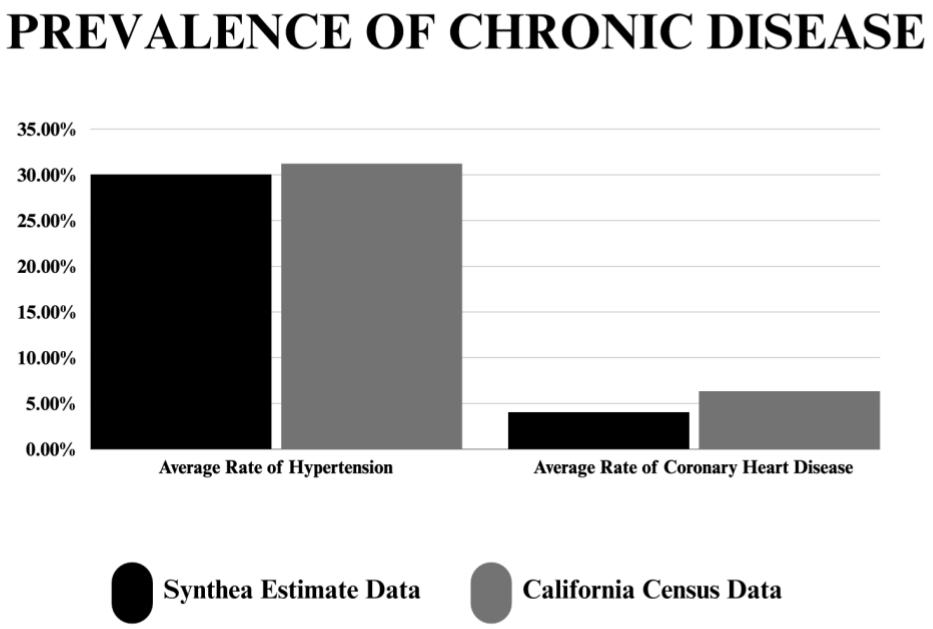

Figure 3. Comparison of the prevalence of hypertension and coronary heart disease between Synthea and real-world California data

Conclusion

Our study bridges the gap in understanding how similar Synthea data are to the real world. As shown in the study by Hodges et al., further research is needed to develop tools that promote realism in synthetic data generations[4]. With continued advancements, Synthea has the potential to significantly improve observational healthcare research, inform treatment rollout policies, manage chronic diseases, and predict future outcomes. High-fidelity low-cost data is essential for these methods, necessitating continued study and improvement of models like Synthea.

References:

1. Gonzales A, Guruswamy G, Smith SR. Synthetic data in health care: A narrative review. PLOS Digital Health2023;2:e0000082. doi:https://doi.org/10.1371/journal.pdig.0000082
2. Azizi Z, Zheng C, Mosquera L, et al. Can synthetic data be a proxy for real clinical trial data? A validation study. BMJ Open 2021;11:e043497.
3. Walonoski J, Kramer M, Nichols J, et al. Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. Journal of the American Medical Informatics Association 2017;25:230–8. doi:https://doi.org/10.1093/jamia/ocx079.

4. Hodges R, Tokunaga K, LeGrand J. A novel method to create realistic synthetic medication data. JAMIA Open 2023;6. doi:https://doi.org/10.1093/jamiaopen/ooad052
5. U.S. Census Bureau. Comparative Demographic Estimates. 2022.
6. Department of Health Care Access and Information, Healthcare Analytics Branch. Hospital Inpatient - Characteristics by Patient County of Residence. 2023.
7. Centers for Disease Control and Prevention. Interactive Atlas of Heart Disease and Stroke.