Intelligence-Based Medicine 1-2 (2020) 100007

Contents lists available at ScienceDirect

Intelligence-Based Medicine

journal homepage: www.journals.elsevier.com/intelligence-based-medicine

Synthea™ Novel coronavirus (COVID-19) model and synthetic data set

Jason Walonoski [a,*], Sybil Klaus [a], Eldesia Granger [a], Dylan Hall [a], Andrew Gregorowicz [a], George Neyarapally [a], Abigail Watson [b], Jeff Eastman [c]

[a] *MITRE, USA*
[b] *Symptomatic, USA*
[c] *MiHIN, USA*

ARTICLE INFO

Keywords:
Synthetic data
COVID-19
Electronic health records

ABSTRACT

March through May 2020, a model of novel coronavirus (COVID-19) disease progression and treatment was constructed for the open-source Synthea patient simulation. The model was constructed using three peer-reviewed publications published in the early stages of the global pandemic, when less was known, along with emerging resources, data, publications, and clinical knowledge. The simulation outputs synthetic Electronic Health Records (EHR), including the daily consumption of Personal Protective Equipment (PPE) and other medical devices and supplies.

For this simulation, we generated 124,150 synthetic patients, with 88,166 infections and 18,177 hospitalized patients. Patient symptoms, disease severity, and morbidity outcomes were calibrated using clinical data from the peer-reviewed publications. 4.1% of all simulated infected patients died and 20.6% were hospitalized. At peak observation, 548 dialysis machines and 209 mechanical ventilators were needed. This simulation and the resulting data have been used for the development of algorithms and prototypes designed to address the current or future pandemics, and the model can continue to be refined to incorporate emerging COVID-19 knowledge, variations in patterns of care, and improvement in clinical outcomes. The resulting model, data, and analysis are available as open-source code on GitHub and an open-access data set is available for download.

Introduction

Synthetic data generation is a proven approach to sharing realistic-but-not-real data without the privacy and security risks associated with real health data.

Synthetic data is not deidentified data. Synthetic data is generated either from models based on aggregated statistics (e.g., modeling and simulation without direct access to any individual data points) or models abstracted from sensitive data (e.g., machine learning models that were trained from, but do not preserve, individual data records). Deidentified data are often modified from real data points using methodologies such as masking or deleting fields and introducing noise.

The assumption that deidentification guarantees privacy or eliminates risk is false [1]. Synthetic data has been widely used as a safe alternative to deidentification. Synthetic data is considered ethically superior to deidentified data, because there is no individual sensitive record underneath any synthetic record that can ever be reidentified [1].

Synthetic clinical data sets can be openly shared to enable innovation, such as from the open-source Synthea ("Synthetic Health") project which publicly provides millions of longitudinal synthetic health records [2]. Synthetic data is being used for software testing and validation (including privacy and security testing), education, academic research, feasibility assessments and algorithm validation, but not yet for clinical discovery and scientific inference [3]. Criticisms of Synthea are that it does not fully account for variations in health care delivery by providers, has limited heterogenous health outcomes after major interventions, but it can be improved and validated [4], and that it does not yet contain sufficient clinical notes [5].

Other synthetic generation techniques, such as Generative Adversarial Networks (GANs) used in medical imaging are experiencing rapid growth (150 articles in the last three years) and the results are filling an important niche in data science [6].

Synthetic data aligns with the Open Science movement which includes open access, open source, and open data among its principles to address the scientific reproducibility problem.

The scientific reproducibility problem is especially severe in health

[*] Corresponding author.
E-mail address: jwalonoski@mitre.org (J. Walonoski).

https://doi.org/10.1016/j.ibmed.2020.100007
Received 25 June 2020; Received in revised form 15 September 2020; Accepted 24 September 2020

Table 1
Study characteristics.

Source	Location	Timing	Total Patients	Admitted Patients	Survivors	Non-Survivors	Outcomes Not Determined at Publication
Zhou et al. [9]	Wuhan	Jan 11, 2020–Jan 31, 2020 (20 days)	191	191	137 (71.73%)	54 (28.27%)	0
Guan et al. [10]	Wuhan	Dec 11, 2019–Jan 31, 2020 (51 days)	1099	1099	64 (5.82%)	15 (1.36%)	1029 (93.63%)
Richardson et al. [11]	New York City	Mar 1, 2020 – Apr 4, 2020 (34 days)	5700	5700	2081 (36.51%)	553 (9.70%)	3066 (53.79%)
Synthetic Data	Massachusetts	Jan 20, 2020–May 26, 2020 (127 days)	124,150 (88,166 infected)	18,177	14,654 (80.61%)	3548 (19.51%)	0

research (especially health machine learning) where data sets and code are more likely to be unavailable. Synthetic data has been identified as a way for researchers to meaningfully release data, code, and results [7].

When properly constructed and validated, synthetic data used in data analytics and machine learning tasks has been shown to have the same results as real data in several domains without compromising privacy [8]. However, these domains are generally not as complex or as high-stakes as health care responses to a pandemic such as COVID-19, so synthetic health data should always be validated for a researcher's specific use-case prior to utilization. Without access to the unpublished raw data, only peer-reviewed research results with summary statistics, we have attempted to calibrate the synthetic data to those reference statistics (as presented in Methods and Materials and Results). We outline limitations and suggested uses in the Discussion section.

Materials and methods

March through May 2020, a model of novel coronavirus (COVID-19) disease progression and treatment was constructed for the open-source Synthea patient simulation. The model was constructed during the global pandemic using emerging resources, data, publications, and clinical expertise. Therefore, the model does not currently represent knowledge of the virus that has emerged since May 2020, including various clades and associated degrees of severity, and the care pathway represents what was reasonably known at that stage of the pandemic. The simulation outputs synthetic Electronic Health Records (EHR), including

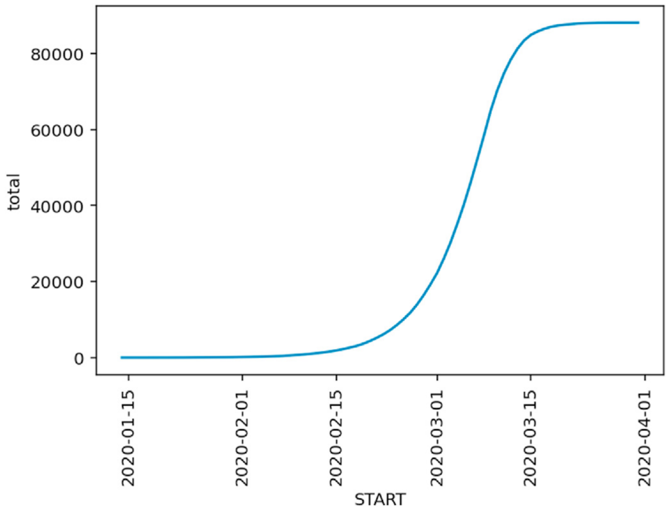

Fig. 1. Infection rate of simulated patients.

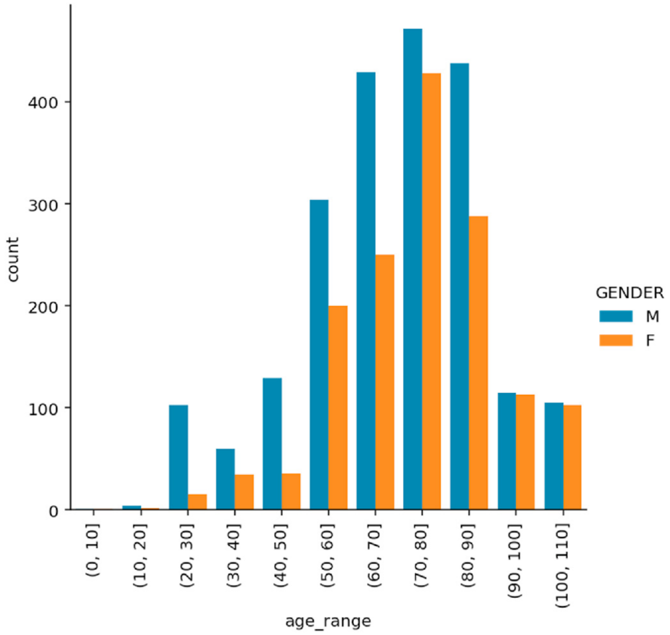

Fig. 2. Mortality of simulated patients by age range and gender.

Table 2
Modules and their Descriptions.

Module	Description
covid19	Determines exposure and infection rates.
covid19/admission	Contains the daily loop during hospitalization and ICU treatment.
covid19/determine_risk	Determines risk based on comorbidities, severity of disease, and whether or not the patient will survivor.
covid19/infection	Determines whether or not patients will be testing, the testing results, and whether or not they are admitted to the hospital.
covid19/measurements_daily	Records daily lab values.
covid19/measurements_frequent	Records frequent lab values.
covid19/measurements_vitals	Records vital signs.
covid19/medications	Potentially enrolls a critical or severe patient in one of eighteen clinical trials.
covid19/nonsurvivor_lab_values	Sets lab values for patients who will not survive.
covid19/survivor_lab_values	Sets lab values for patients who will survive.
covid19/outcomes	Determines outcomes and complications based on risk and disease severity.
covid19/end_outcomes	Ends complications after recovery (if applicable).
covid19/supplies_hospitalization	Contains the daily supplies used within the hospital for 1 patient, 1 physician, and 1 nurse.
covid19/supplies_icu	Contains the daily supplies used within ICU for 1 patient, 1 physician, and 1 nurse.
covid19/supplies_intubation	Contains the supplies used for intubation.
covid19/symptoms	Determines the symptoms presenting in each patient.
covid19/end_symptoms	Ends symptoms after recovery (if applicable).
covid19/diagnose_blood_clot	Patients will likely develop blot clots during inpatient and ICU stay.
covid19/treat_blood_clot	Blood clots need to be treated once they are developed.
covid19/diagnose_bacterial_infection	Patients may develop a secondary bacterial infection in the ICU.

Table 3
Outcomes.

Outcome	All Patients (n = 88,166)	Hospitalized (n = 18,177)	ICU Admitted (n = 3677)	Required Ventilation (n = 2914)
Home Isolation	0.80	0.02	0.02	0.02
Hospital Admission	0.20	1.00	1.00	1.00
ICU Admission	0.04	0.20	1.00	1.00
Ventilated	0.03	0.16	0.79	1.00
Recovered	0.95	0.80	0.32	0.14
Death	0.04	0.19	0.67	0.85

Table 4
Supply and device usage.

DESCRIPTION	QUANTITY
Alcohol disinfectant	245,937
Antiseptic towelette	1,947,098
Basic endotracheal tube single-use	2914
Carbon dioxide breath analyzer	2914
Disposable air-purifying respirator	446,438
Endotracheal tube holder	2914
Endotracheal tube stylet single-use	2914
Face shield	437,696
Human plasma blood product (product)	143
Isolation gown reusable	48,350
Isolation gown single-use	2,583,654
Laryngoscope blade single-use	2914
Lubricant	2914
Nasogastric tube device	2914
Nitrile examination/treatment glove non-powdered sterile	5,759,164
Operating room gown single-use	2914
Protective glasses device	8742
Suction system	2914
Surgical cap single-use	2914
Syringe device	5828
Viral filter	2914

the daily consumption of Personal Protective Equipment (PPE) and other medical supplies.

The COVID-19 models within Synthea were primarily modeled on three peer-reviewed clinical papers, based on findings from Wuhan, China [9,10] and mortality data from New York City, USA [11]. The characteristics of these studies and the final synthetic data are summarized in Table 1.

Data from the primary reference Tables and Figures used as input into the modeling process are included within this paper along side our results for comparison purposes. Model clinical definitions, commentary, and assumptions are detailed in Appendix A.

The Synthea simulation is divided into "modules" which are summarized in Table 2. The Synthea models are viewable at:

- https://synthetichealth.github.io/module-builder.

and can be downloaded from:

- https://github.com/synthetichealth/synthea/tree/master/src/main/resources/modules

Results

We generated 124K patients and performed some basic analysis to produce Figures and summarize outcomes corresponding to Figures and Tables from our primary data sources and present these for comparison. It is important to note that the model was developed from the primary source tables, and not the raw data from these sources which was unpublished.

Infection rate and mortality

Of the 88,166 infections in the generated population (not all the simulated patients became infected), 18,177 patients were hospitalized. Based on current knowledge this hospitalization rate is high, but at the early stages of the pandemic, we estimated hospitalization rates based on projected outcomes related to patient comorbidities and risk factors without consideration of disease prevention measures. The simulated infection timeline is illustrated in Fig. 1. The mortality graph illustrated in Fig. 2 shows the mortality disparity between age and gender groups.

Outcomes

Outcomes are enumerated in Table 3. The "Outcome" column describes an outcome (e.g. ventilated, recovered, or death), and the other columns are cross correlated with other groups (e.g. all patients, patients who were hospitalized, patients who were admitted to the ICU, and patients who required ventilation).

Regarding cells marked with "1.00" – that indicates that 100% of the patients in that group had that outcome. For example, the cell corresponding to the "ICU Admitted" column and "Hospital Admission" row is "1.00" – indicates that all patients who were admitted to the ICU were also admitted to the hospital (in this case, a prerequisite event in our model).

Supply and device usage

The amount of supplies consumed for this particular simulation run are enumerated in Table 4. The simulation ran for 88K infected patients, of which 18,177 were admitted to the hospital. A discussion on the assumptions made about supply consumption and device usage are documented in Appendix A. Supply models are documented in Appendix B: Supply and Device Lists.

For this simulation, at peak 548 hemodialysis machines were needed, with 209 mechanical ventilators, as illustrated in Fig. 3.

Major complications among survivors and Non-Survivors

Complications among survivors and non-survivors are listed in Tables 5 and 6 for synthetic patients and the reference data, respectively. There are discrepancies between the these outcomes (compare the "percent" columns in each table) because the outcomes in the model are not fixed by percentages from Table 6, but are based on risk-factors that determine severity and mortality including gender, age, and comorbidities that differ from the reference population.

For comparison, Table 6 reproduces reference data from the "Outcomes" portion of Table 2 from Ref. [9].

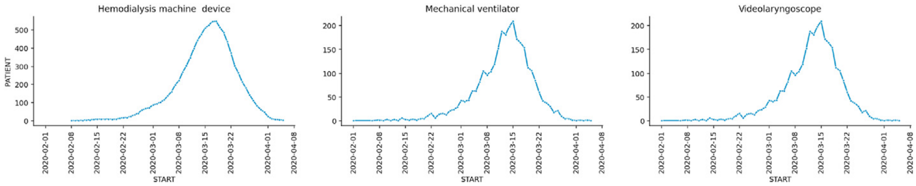

Fig. 3. Device usage over time.

Table 5
Major complications among survivors and non-survivors (synthetic data).

outcome	total (n = 18,177)	percent of inpatient	Survivors (n = 14,654)	percent survivors	Nonsurvivors (n = 3548)	percent non survivors
Sepsis	6945	0.381551	3419	0.233315	3526	0.993799
Respiratory Failure	8710	0.478519	5237	0.357377	3473	0.978861
ARDS	2401	0.131909	85	0.005800	2316	0.652762
Heart Failure	1434	0.078783	124	0.008462	1310	0.369222
Septic Shock	1746	0.095924	0	0.000000	1746	0.492108
Coagulopathy	1389	0.076310	91	0.006210	1298	0.365840
Acute Cardiac Injury	1288	0.070761	20	0.001365	1268	0.357384
Acute Kidney Injury	1252	0.068784	8	0.000546	1244	0.350620

Major Lab Values

Major lab values were modeled according the temporal changes in laboratory markers documented in Fig. 2 from Ref. [9]. These temporal changes in laboratory values are illustrated here in Fig. 4.

Patient timelines

Patient timelines were modified from the illustration of common patient timelines and statistics from Fig. 1 ("Clinical courses of major symptoms and outcomes and duration of viral shedding from illness onset in patients hospitalized with COVID-19") from Ref. [9] and are paralleled here in Figs. 5 and 6. Fig. 5 shows patients who were hospitalized, some of which are later admitted to the ICU. Fig. 6 shows only the ICU patients.

The average of length of stay for the synthetic patients is detailed in Table 7, with reference ranges from Ref. [9] that are inclusive of both survivors and non-survivors.

Patient symptoms

Patient symptoms were modeled from Table 1 "Clinical Characteristics of the Study Patients, According to Disease Severity and Presence or Absense of the Primary Composite End Point" from Ref. [10], except for Loss of Taste which was based upon the findings from Ref. [12]. The symptoms in Ref. [10] are based on disease severity (severe and non-severe) while our findings are broken down by survivor and non-survivor, which are overlapping but not identical populations. The reference data is listed in Table 8 and the simulation results are listed in Table 9 (all infected synthetic patients) and Table 10 (synthetic patients admitted into the ICU).

Discussion

> *"No plan of operations extends with any certainty beyond the first contact with the main hostile force." – Field Marshall Helmuth Karl Bernhard Graf von Moltke*
>
> *"All models are wrong, some are useful." – George P.E. Box*

Synthea is an open-source modeling and simulation platform for disease progression and treatment. If we take the George Box quote above to be true, that all models are wrong, then Synthea is wrong. And if we take Field Marshall Moltke's notion of "*no plan survives contact with the enemy*" as true and expand the scope to modeling and simulation, then we might say that "*no model survives contact with reality.*" Which is all to say that our model of novel coronavirus is flawed as all models are, and as a model, it cannot not survive contact with reality. Nevertheless, we hope it is useful.

To our knowledge, the Synthea COVID-19 data has been useful in several online challenges, hackathons, and conferences [13–17]. In these venues, the data has spurred innovation and exchange of ideas about software solutions, enabled software development and testing, and has been used as the basis for some prediction modeling.

Those prediction models are likely not suitable for application in the delivery of clinical care, however they do enable a machine-learning team to begin to explore realistic data, develop their ideas and solution, build a processing pipeline – all before they are able to gain secure access to restricted data sets of real COVID-19 patients and outcomes. It also provides learning opportunities to teams that would otherwise be unable to gain access to such data sets and lowers the barrier to entry to participating in AI and ML activities in healthcare.

In the future, when more COVID-19 real-world data sets become available, including EHR data and associated outcomes, it will be possible to tune the model weights and probabilities to match real cohorts and diverse variations in care. For example, using data from one region during a particular month of the pandemic, the model could be calibrated to generate data more representative of that cohort (including infection rates, disease severity, treatments, and outcomes).

Finally, another major weakness of this model is that Synthea does not currently restrict or limit the care or supplies based on capacity, so the resulting data represents an upper bound. The models could be modified to account for this but would require additional pathways to be modeled (for example, what occurs when a required ventilator is unavailable).

Conclusions

The COVID-19 pandemic has brought unprecedented sharing of data, knowledge, techniques, equipment and supplies across the globe. Nevertheless, there is a role for the distribution of realistic-but-not-real synthetic data sets among the community of health innovation (AI, ML, software, other) both with the academic and the practitioner.

This synthetic data fills in data availability gaps, lowers the innovation barrier to entry, and aligns with the Open Science movement which includes open access, open source, and open data among its principles.

Table 6
Major complications among survivors and non-survivors (reference data).

outcome	total (n = 191)	percent	Survivors (n = 137)	percent survivors	Nonsurvivors (n = 54)	percent non survivors
Sepsis	112	0.59	58	0.42	54	1.00
Respiratory Failure	103	0.54	50	0.36	53	0.98
ARDS	59	0.31	9	0.07	50	0.93
Heart Failure	44	0.23	16	0.12	28	0.52
Septic Shock	38	0.20	0	0.00	38	0.70
Coagulopathy	37	0.19	10	0.07	27	0.50
Acute Cardiac Injury	33	0.17	1	0.01	32	0.59
Acute Kidney Injury	28	0.15	1	0.01	27	0.50

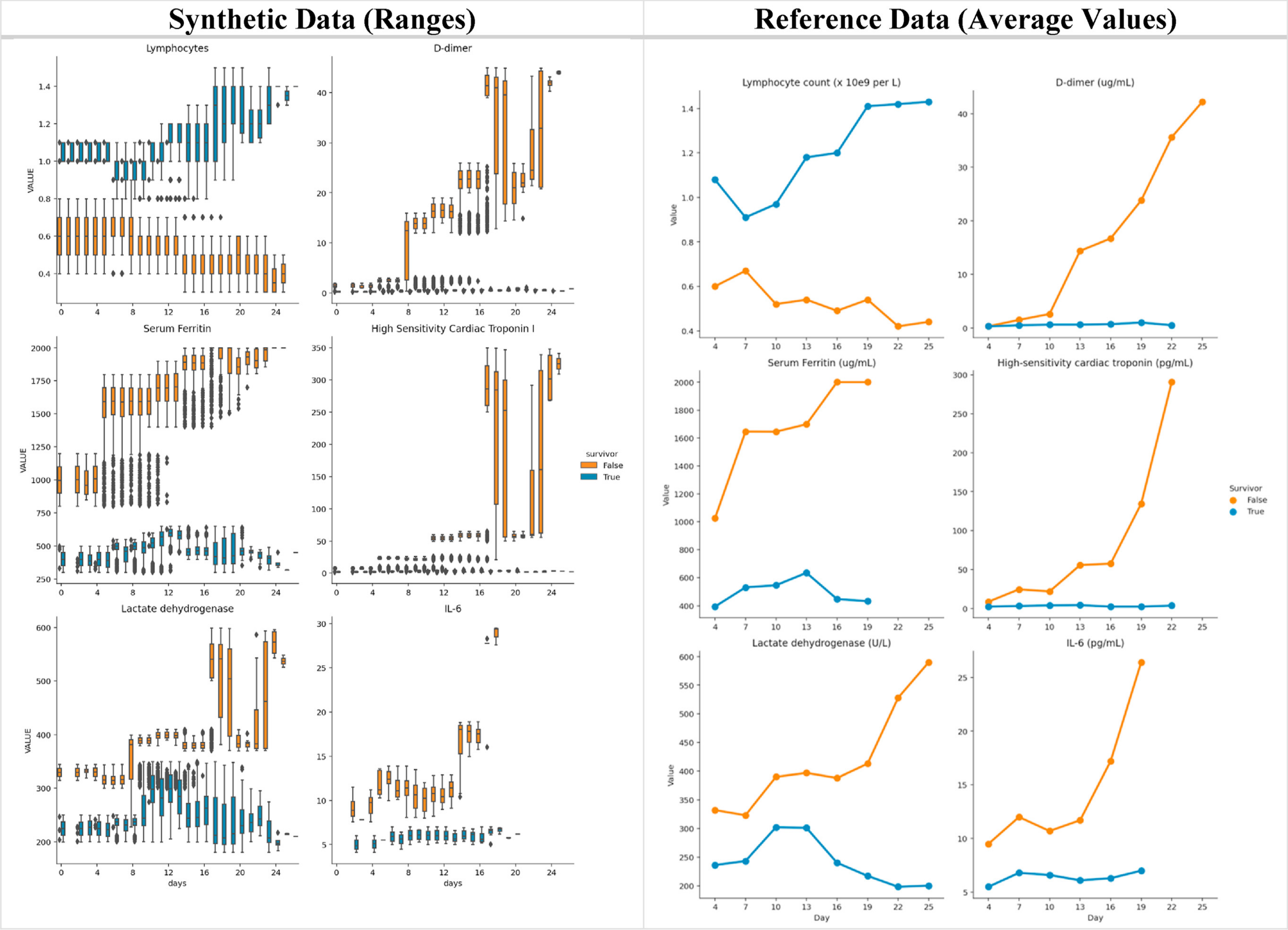

Fig. 4. Major lab values. Synthetic data (left) and reference data (right).

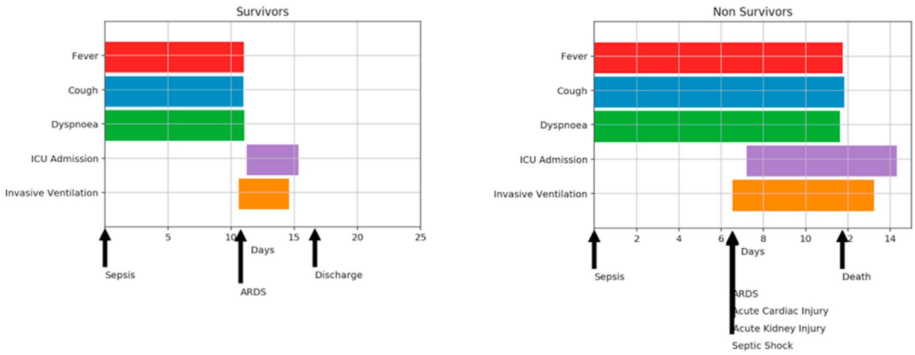

Fig. 5. Synthetic hospitalized patient times. Survivors (left) and non-survivors (right).

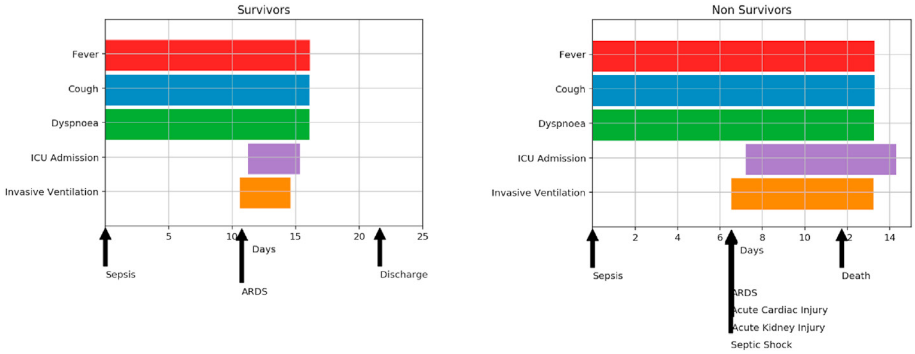

Fig. 6. Synthetic ICU patient timelines. Survivors (left) and non-survivors (right).

Table 7
Length of stay.

Type	Patients	Average Stay (days)	Reference Range (days, average and range) [9]	Total days
inpatient	18,177	15.03	11 (5–15)	273,224
ICU	3677	6.08	8 (2–12)	22,379

Table 8
Symptoms by disease severity (reference data).

	Symptoms	All Patients Percentage	All Patients (n = 1099)	Severe Percentage	Severe (n = 173)	Non Severe Percentage	Non Severe (n = 926)
0	Conjunctival Congestion	0.008	9	0.023	4	0.005	5
1	Nasal Congestion	0.048	53	0.035	6	0.051	47
2	Headache	0.136	150	0.150	26	0.134	124
3	Cough	0.678	745	0.705	122	0.673	623
4	Sore Throat	0.139	153	0.133	23	0.140	130
5	Sputum Production	0.337	370	0.353	61	0.334	309
6	Fatigue	0.381	419	0.399	69	0.378	350
7	Hemoptysis	0.009	10	0.023	4	0.006	6
8	Shortness of Breath	0.187	205	0.376	65	0.151	140
9	Nausea	0.050	55	0.023	12	0.046	43
10	Diarrhea	0.038	42	0.058	10	0.035	32
11	Muscle Pain	0.149	164	0.173	30	0.145	134
12	Joint Pain	0.149	164	0.173	30	0.145	134
13	Chills	0.115	126	0.150	26	0.108	100
14	Loss of Taste	0.64	130 (n = 374)	n/a	n/a	n/a	n/a

Table 9
Symptoms by Mortality for all infected synthetic patients.

	Symptoms	All Patients Percentage	All Patients (n = 88,166)	Survivor Percentage	Survivor (n = 84,618)	Non Survivor Percentage	Non Survivor (n = 3548)
0	Conjunctival Congestion	0.008371	738	0.007718	653	0.023895	87
1	Nasal Congestion	0.046257	4078	0.046697	3951	0.036254	132
2	Headache	0.137602	12,131	0.136747	11,570	0.158473	577
3	Cough	0.677892	59,763	0.676346	57,225	0.713540	2598
4	Sore Throat	0.140585	12,394	0.140623	11,898	0.138149	503
5	Sputum Production	0.336468	29,663	0.336004	28,429	0.346608	1262
6	Fatigue	0.383961	33,850	0.383529	32,450	0.393299	1432
7	Hemoptysis	0.009698	855	0.009006	762	0.026092	95
8	Shortness of Breath	0.198662	17,514	0.191434	16,197	0.367481	1338
9	Nausea	0.051032	4499	0.050467	4270	0.063444	231
10	Diarrhea	0.038929	3432	0.038176	3230	0.057127	208
11	Muscle Pain	0.150726	13,288	0.149310	12,633	0.184839	673
12	Joint Pain	0.150726	13,288	0.149310	12,633	0.184839	673
13	Chills	0.116629	10,282	0.115295	9755	0.148036	539
14	Loss of Taste	0.506375	44,642	0.506105	42,821	0.512497	1866

Table 10
Symptoms by Mortality for synthetic patients admitted to the ICU.

	Symptoms	All Patients Percentage	All Patients (n = 3677)	Survivor Percentage	Survivor (n = 1179)	Non Survivor Percentage	Non Survivor (n = 2498)
0	Conjunctival Congestion	0.022573	83	0.017797	21	0.024820	62
1	Nasal Congestion	0.037259	137	0.040678	48	0.035629	89
2	Headache	0.157193	578	0.140678	166	0.164932	412
3	Cough	0.717161	2637	0.722881	853	0.714572	1785
4	Sore Throat	0.132445	487	0.117797	139	0.139311	348
5	Sputum Production	0.345662	1271	0.339831	401	0.348279	870
6	Fatigue	0.403046	1482	0.422034	498	0.394315	985
7	Hemoptysis	0.026380	97	0.022881	27	0.028022	70
8	Shortness of Breath	0.375578	1381	0.394915	466	0.366693	916
9	Nausea	0.065271	240	0.069492	82	0.063251	158
10	Diarrhea	0.054392	200	0.054237	64	0.054444	136
11	Muscle Pain	0.178406	656	0.173729	205	0.180544	451
12	Joint Pain	0.178406	656	0.173729	205	0.180544	451
13	Chills	0.147131	541	0.144915	171	0.148118	370
14	Loss of Taste	0.527604	1940	0.539831	637	0.521617	1303

The Synthea COVID-19 data has been used in many online challenges, hackathons, and conferences, and we hope it will be useful to academics, students, and practitioners.

The synthetic data is available here: https://synthea.mitre.org/downloads.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ibmed.2020.100007.

References

[1] Gallagher T, Dube K, McLachlan S. Ethical issues in secondary use of personal health information. IEEE future directions: technology policy & ethics. http://sites.ieee.org/futuredirections/tech-policy-ethics/may2018/ethical-issues-in-secondary-use-of-personal-health-information/; 2018.

[2] Walonoski J, Kramer M, Nichols J, Quina A, Moesel C, Hall D, Duffett C, Dube K, Gallagher T, McLachlan S. Synthea: an approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. J Am Med Inf Assoc March 2018;25(3):230–8. https://doi.org/10.1093/jamia/ocx079.

[3] Ahalt SC, Chute CG, Fecho K, Glusman G, Hadlock J, Taylor CO, Biomedical Data Translator Consortium. Clinical data: sources and types, regulatory constraints, applications. Clinical and translational science 2019;12(4):329–33. https://doi.org/10.1111/cts.12638.

[4] Chen J, Chun D, Patel M, Chiang E, James J. The validity of synthetic clinical data: a validation study of a leading synthetic data generator (Synthea) using clinical quality measures. BMC Med Inf Decis Making 2019;19(1):44. https://doi.org/10.1186/s12911-019-0793-0.

[5] Orenstein EW, Rasooly IR, Mai MV, Dziorny AC, Phillips W, Utidjian L, Bonafide CP. Influence of simulation on electronic health record use patterns among pediatric residents. J Am Med Inf Assoc: JAMIA 2018;25(11):1501–6. https://doi.org/10.1093/jamia/ocy105.

[6] Yi X, Walia E, Babyn P. Generative adversarial network in medical imaging: a review. Med Image Anal 2019;58. https://doi.org/10.1016/j.media.2019.101552.

[7] McDermott, M.B.A., Wang, S., Marinsek, N., Ranganath, R., Ghassemi, M., Foschini, L. Reproducibility in machine learning for health. ICLR 2019 reproducibility in machine learning workshop. https://arxiv.org/abs/1907.01463.

[8] Patki N, Wedge R, Veeramachaneni K. The synthetic data vault. IEEE International Conference on Data Science and Advance Analytics Montreal, CA 2016. https://dai.lids.mit.edu/wp-content/uploads/2018/03/SDV.pdf.

[9] Zhou Fei, Yu Ting, Du Ronghui, Fan Guohui, Liu Ying, Liu Zhibo, Xiang Jie, Wang Yeming, Song Bin, Gu Xiaoying, Guan Lulu, Wei Yuan, Li Hui, Wu Xudong, Xu Jiuyang, Tu Shengjin, Zhang Yi, Chen Hua, Cao Bin. Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. Lancet 2020;395(10229):1054–62. https://doi.org/10.1016/S0140-6736(20)30566-3. ISSN 0140–6736.

[10] Guan WJ, Ni ZY, Hu Y, Liang WH, Ou CQ, He JX, Liu L, Shan H, Lei CL, Hui D, et al. Clinical characteristics of coronavirus disease 2019 in China. N Engl J Med 2020. https://doi.org/10.1056/NEJMoa2002032. PMID: 32109013.

[11] Richardson S, Hirsch JS, Narasimhan M, et al. Presenting characteristics, comorbidities, and outcomes among 5700 patients hospitalized with COVID-19 in the New York city area. J Am Med Assoc April 22, 2020. https://doi.org/10.1001/jama.2020.6775.

[12] Spinato G, Fabbris C, Polesel J. Alterations in smell or Taste in mildly symptomatic outpatients with SARS-CoV-2 infection. J Am Med Assoc April 22, 2020. Published online, https://jamanetwork.com/journals/jama/fullarticle/2765183.

[13] Pandemic Response Hackathon. Hack COVID-19: project roundup. Convened by datavant. https://datavant.com/pandemic-response-hackathon/.

[14] COVID19 geomapping: geomapping COVID19 data from hospitals using FHIR. https://devpost.com/software/covid19-on-fhir.

[15] MIT COVID19 Challenge. https://covid19challenge.mit.edu. http://www.hsraanz.org/wp-content/uploads/Covid-Data-sets.pdf.

[16] FHIR Dev Days. Microsoft hack-on-FHIR and the Synthea COVID-19 dataset. https://www.devdays.com/us/; June 15 – 18, 2020.

[17] PrecisionFDA: VHA innovation ecosystem and precisionFDA COVID-19 risk factor modeling challenge. https://precision.fda.gov/challenges/11; June 2, 2020 – July 3, 2020.