Welcome to the Sentinel Innovation and Methods Seminar Series

The webinar will begin momentarily

Please visit www.sentinelinitiative.org for recordings of past sessions and details on upcoming webinars.

Note: closed-captioning for today's webinar will be available on the recording posted at the link above.

Office of Health Information and Data Science

Evaluating the Utility of Synthetic Data for Research

Randi Foraker, PhD, MA, FAHA, FAMIA

Professor of Medicine

Director, Center for Population Health Informatics at the Institute for Informatics

Director, Public Health Data and Training Center at the Institute for Public Health

Director, Center for Administrative Data Research, Washington University School of Medicine

Washington University School of Medicine in St. Louis

MDClone @WUSTL

INSTITUTE FOR INFORMATICS / BJC DATA RESERVOIR

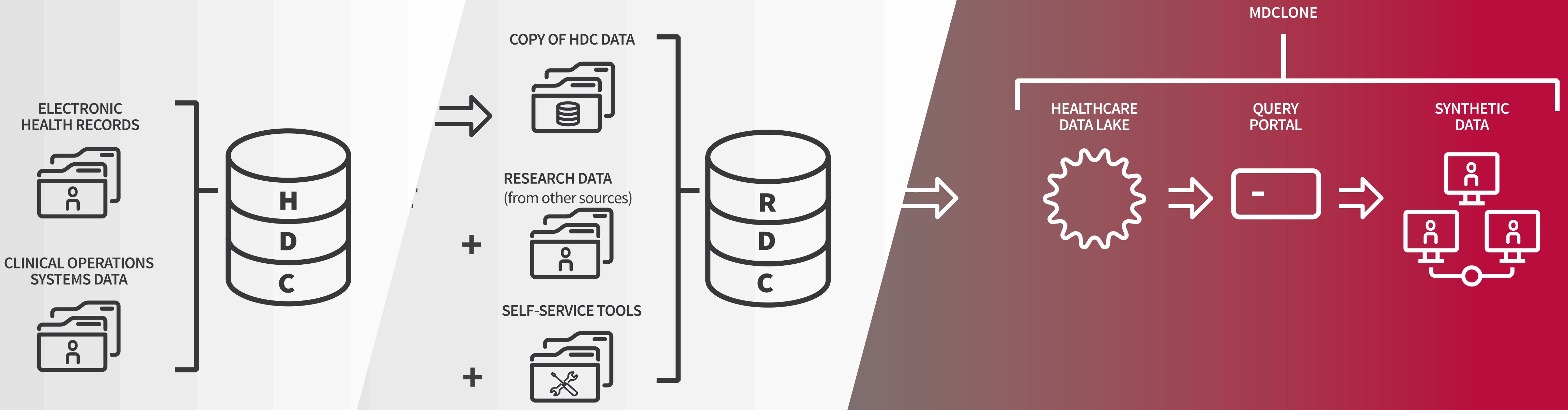
Clinical Data

The HDC offers an integrated view of BJC HealthCare and WUSM physicians data.

Research Data

The RDC augments the HDC's clinical data with WUSM research-oriented data.

Cloud



1

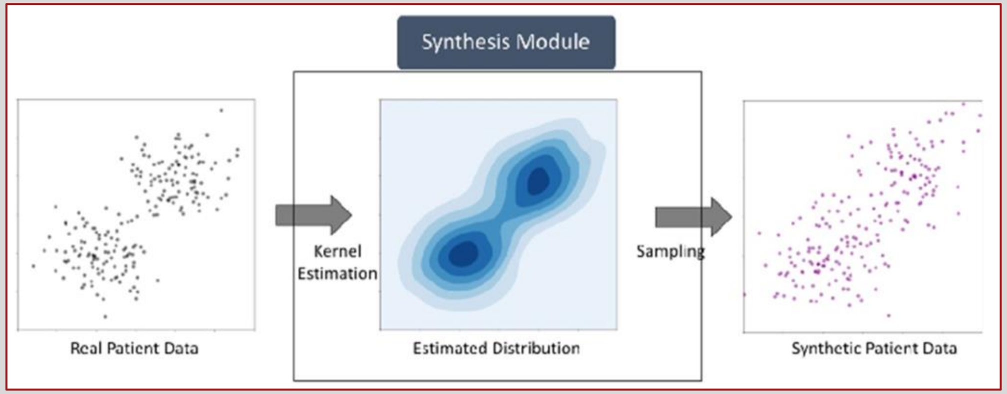
Prediction of head trauma severity. We used logistic regression and area under the receiver operating characteristic curve.

2

Sepsis prediction. We demonstrated machine learning approaches of training:testing models on original and synthetic data, respectively.

3

Geospatial analyses. We compared rate differences between zip codes and show the geographic characteristics of illness.



JAMIA Open, 00(0), 2020, 1–10
doi: 10.1093/jamiaopen/ooaa060
Research and Applications

Research and Applications

Spot the difference: comparing results of analyses from real patient data and synthetic derivatives

Randi E. Foraker,[1,2] **Sean C. Yu**,[2] **Aditi Gupta**,[2] **Andrew P. Michelson**,[3] **Jose A. Pineda Soto**,[4] **Ryan Colvin**,[2,4] **Francis Loh**,[5] **Marin H. Kollef**,[3] **Thomas Maddox**,[6] **Bradley Evanoff**,[1] **Hovav Dror**,[7] **Noa Zamstein**,[7] **Albert M. Lai**,[1,2] and **Philip R.O. Payne**[1,2]

[1]Division of General Medical Sciences, Department of Medicine, School of Medicine, Washington University in St. Louis, St. Louis, Missouri, USA, [2]Department of Medicine, Institute for Informatics, School of Medicine, Washington University in St. Louis, St. Louis, Missouri, USA, [3]Division of Pulmonary and Critical Care Medicine, Department of Medicine, School of Medicine, Washington University in St. Louis, St. Louis, Missouri, USA, [4]Division of Critical Care Medicine, Department of Anesthesiology and Critical Care Medicine, Children's Hospital of Los Angeles, Los Angeles, California, USA, [5]School of Medicine, Washington University in St. Louis, St. Louis, Missouri, USA, [6]Healthcare Innovation Lab, BJC Healthcare, School of Medicine, Washington University in St. Louis, St. Louis, Missouri, USA and [7]MDClone Ltd, Beer Sheva, Israel

Corresponding Author: Randi Foraker, PhD, MA, FAHA, FAMIA, Associate Professor, Institute for Informatics (I2), Director, Center for Population Health Informatics at I2, Washington University in St. Louis, School of Medicine, 600 S. Taylor Avenue, Suite 102, Campus Box 8102, St. Louis, MO 63110, USA (http://informatics.wustl.edu)

Received 14 August 2020; Revised 14 October 2020; Accepted 20 October 2020

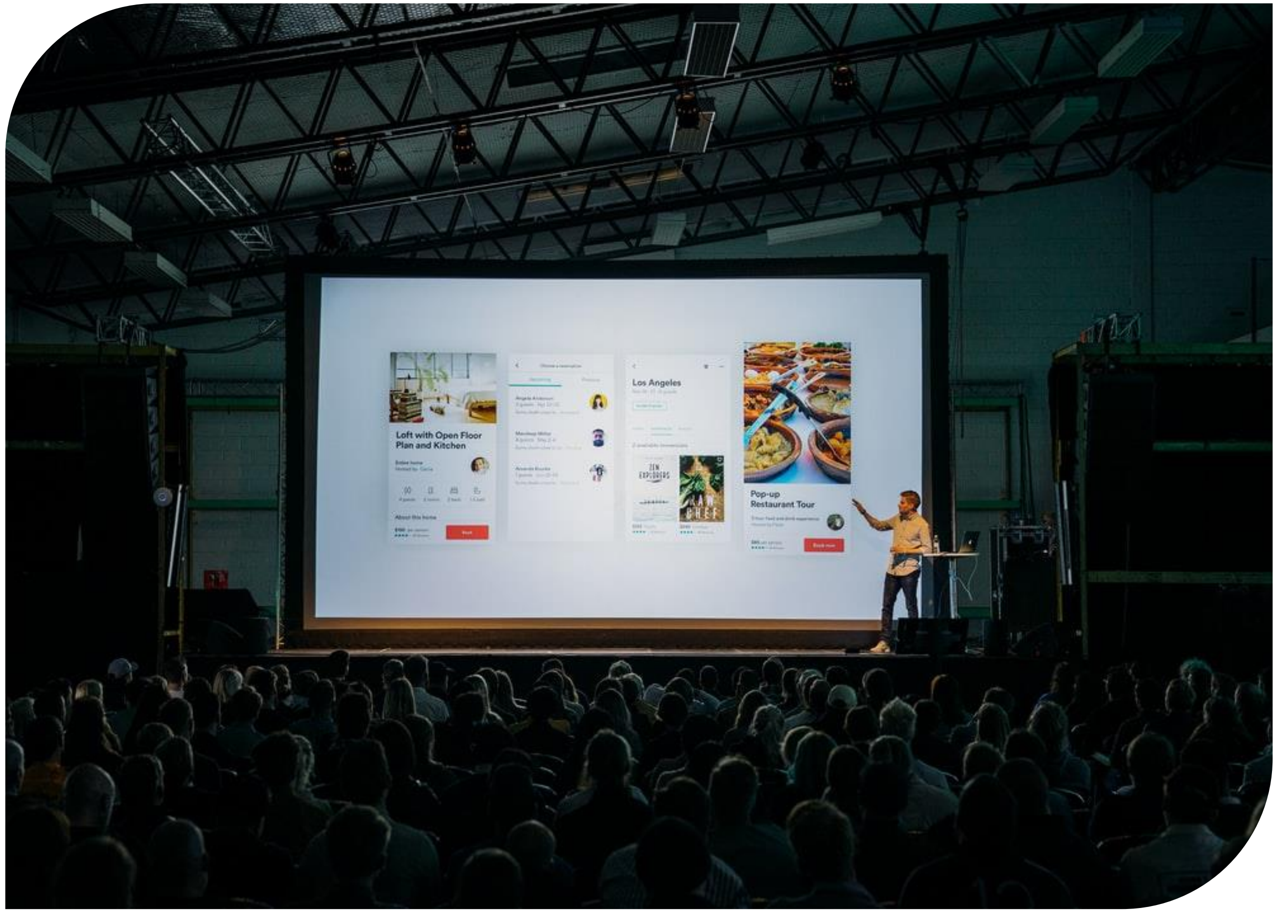
ABSTRACT

Background: Synthetic data may provide a solution to researchers who wish to generate and share data in support of precision healthcare. Recent advances in data synthesis enable the creation and analysis of synthetic derivatives as if they were the original data; this process has significant advantages over data deidentification.
Objectives: To assess a big-data platform with data-synthesizing capabilities (MDClone Ltd., Beer Sheva, Israel) for its ability to produce data that can be used for research purposes while obviating privacy and confidentiality concerns.
Methods: We explored three use cases and tested the robustness of synthetic data by comparing the results of analyses using synthetic derivatives to analyses using the original data using traditional statistics, machine learning approaches, and spatial representations of the data. We designed these use cases with the purpose of conducting analyses at the observation level (Use Case 1), patient cohorts (Use Case 2), and population-level data (Use Case 3).
Results: For each use case, the results of the analyses were sufficiently statistically similar ($P > 0.05$) between the synthetic derivative and the real data to draw the same conclusions.
Discussion and conclusion: This article presents the results of each use case and outlines key considerations for the use of synthetic data, examining their role in clinical research for faster insights and improved data sharing in support of precision healthcare.

Key words: synthetic data, protected health information, precision health care, electronic health records and systems, data analysis

Risk stratification

- Direct high-risk surgical intervention among heart failure (HF) patients
- Computationally-derived data on 26,575 HF patients seen in 2018
- Used 27 features to predict 1-year mortality
- Cross-validation methods were used in supervised deep- and machine-learning approaches
    - Deep neural networks
    - Random forest
    - Logistic regression



frontiers in Digital Health

ORIGINAL RESEARCH
published: 07 December 2020
doi: 10.3389/fdgth.2020.576945

The Use of Synthetic Electronic Health Record Data and Deep Learning to Improve Timing of High-Risk Heart Failure Surgical Intervention by Predicting Proximity to Catastrophic Decompensation

Aixia Guo[1]*, *Randi E. Foraker*[1,2], *Robert M. MacGregor*[3], *Faraz M. Masood*[3], *Brian P. Cupps*[3] and *Michael K. Pasque*[3]

[1] Institute for Informatics (I²), Washington University School of Medicine, St. Louis, MO, United States, [2] Department of Internal Medicine, Washington University School of Medicine, St. Louis, MO, United States, [3] Department of Surgery, Washington University School of Medicine, St. Louis, MO, United States

OPEN ACCESS

Edited by:
Juan Liu,
Huazhong University of Science and Technology, China

Reviewed by:
Liang Zhang,
Xidian University, China
Zhibo Wang,
University of Central Florida, United States
Kongtao Chen,
University of Pennsylvania, United States

***Correspondence:**
Aixia Guo
aixia.guo@wustl.edu

Specialty section:
This article was submitted to Health Informatics, a section of the journal Frontiers in Digital Health

Received: 27 June 2020
Accepted: 13 November 2020
Published: 07 December 2020

Citation:
Guo A, Foraker RE, MacGregor RM, Masood FM, Cupps BP and Pasque MK (2020) The Use of Synthetic Electronic Health Record Data and Deep Learning to Improve Timing of High-Risk Heart Failure Surgical Intervention by Predicting Proximity to Catastrophic Decompensation. Front. Digit. Health 2:576945. doi: 10.3389/fdgth.2020.576945

Objective: Although many clinical metrics are associated with proximity to decompensation in heart failure (HF), none are individually accurate enough to risk-stratify HF patients on a patient-by-patient basis. The dire consequences of this inaccuracy in risk stratification have profoundly lowered the clinical threshold for application of high-risk surgical intervention, such as ventricular assist device placement. Machine learning can detect non-intuitive classifier patterns that allow for innovative combination of patient feature predictive capability. A machine learning-based clinical tool to identify proximity to catastrophic HF deterioration on a patient-specific basis would enable more efficient direction of high-risk surgical intervention to those patients who have the most to gain from it, while sparing others. *Synthetic* electronic health record (EHR) data are statistically indistinguishable from the original protected health information, and can be analyzed as if they were original data but without any privacy concerns. We demonstrate that *synthetic* EHR data can be easily accessed and analyzed and are amenable to machine learning analyses.

Methods: We developed *synthetic* data from EHR data of 26,575 HF patients admitted to a single institution during the decade ending on 12/31/2018. Twenty-seven clinically-relevant features were synthesized and utilized in supervised deep learning and machine learning algorithms (i.e., deep neural networks [DNN], random forest [RF], and logistic regression [LR]) to explore their ability to predict 1-year mortality by five-fold cross validation methods. We conducted analyses leveraging features from prior to/at and after/at the time of HF diagnosis.

Results: The area under the receiver operating curve (AUC) was used to evaluate the performance of the three models: the mean AUC was 0.80 for DNN, 0.72 for RF, and 0.74 for LR. Age, creatinine, body mass index, and blood pressure levels were especially important features in predicting death within 1-year among HF patients.

MDClone @N3C

Original Paper

The National COVID Cohort Collaborative: Analyses of Original and Computationally Derived Electronic Health Record Data

Randi Foraker[1,2], MA, PhD; Aixia Guo[2], PhD; Jason Thomas[3], BS; Noa Zamstein[4], MSc, PhD; Philip RO Payne[1,2], PhD; Adam Wilcox[3], PhD; N3C Collaborative[5]

[1]Division of General Medical Sciences, School of Medicine, Washington University in St. Louis, St. Louis, MO, United States
[2]Institute for Informatics, School of Medicine, Washington University in St. Louis, St. Louis, MO, United States
[3]Department of Biomedical and Medical Education, School of Medicine, University of Washington, Seattle, WA, United States
[4]MDClone Ltd, Beer Sheva, Israel
[5]See Acknowledgments

Corresponding Author:
Randi Foraker, MA, PhD
Division of General Medical Sciences
School of Medicine
Washington University in St. Louis
600 S. Taylor Avenue, Suite 102
Campus Box 8102
St. Louis, MO, 63110
United States
Phone: 1 314 273 2211
Fax: 1 314 273 1390
Email: randi.foraker@wustl.edu

Abstract

Background: Computationally derived ("synthetic") data can enable the creation and analysis of clinical, laboratory, and diagnostic data as if they were the original electronic health record data. Synthetic data can support data sharing to answer critical research questions to address the COVID-19 pandemic.

Objective: We aim to compare the results from analyses of synthetic data to those from original data and assess the strengths and limitations of leveraging computationally derived data for research purposes.

Methods: We used the National COVID Cohort Collaborative's instance of MDClone, a big data platform with data-synthesizing capabilities (MDClone Ltd). We downloaded electronic health record data from 34 National COVID Cohort Collaborative institutional partners and tested three use cases, including (1) exploring the distributions of key features of the COVID-19–positive cohort; (2) training and testing predictive models for assessing the risk of admission among these patients; and (3) determining geospatial and temporal COVID-19–related measures and outcomes, and constructing their epidemic curves. We compared the results from synthetic data to those from original data using traditional statistics, machine learning approaches, and temporal and spatial representations of the data.

Results: For each use case, the results of the synthetic data analyses successfully mimicked those of the original data such that the distributions of the data were similar and the predictive models demonstrated comparable performance. Although the synthetic and original data yielded overall nearly the same results, there were exceptions that included an odds ratio on either side of the null in multivariable analyses (0.97 vs 1.01) and differences in the magnitude of epidemic curves constructed for zip codes with low population counts.

Conclusions: This paper presents the results of each use case and outlines key considerations for the use of synthetic data, examining their role in collaborative research for faster insights.

(J Med Internet Res 2021;23(10):e30697) doi: 10.2196/30697

KEYWORDS

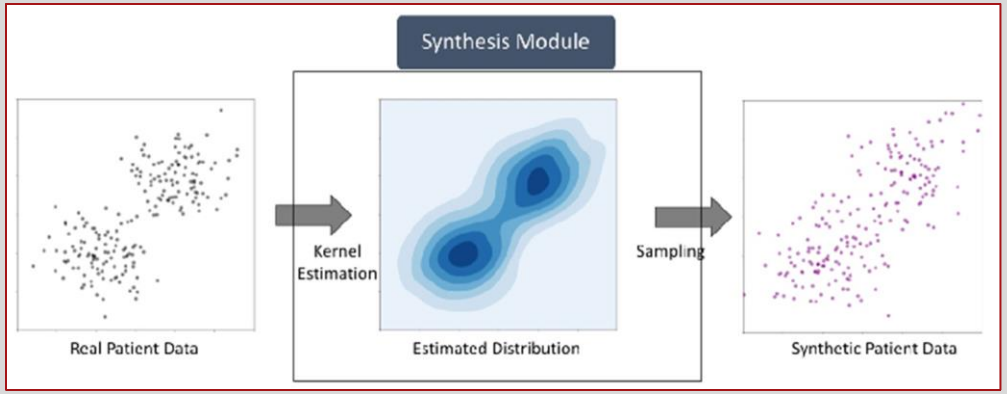
synthetic data; protected health information; COVID-19; electronic health records and systems; data analysis

Foraker R, Guo A, Thomas J, Zamstein N, Payne PR, Wilcox A, 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

1 **Data distributions.** Explored key features off the COVID-19 positive cohort.

2 **Machine learning.** Trained and tested machine learning models assessing the risk of admission among COVID-19 positive patients.

3 **Time-dependent analyses.** Constructed epidemic curves to examine spatial and temporal aspects of the data.



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], Philip R.O. Payne[2,3], Adam B. Wilcox[1,5], the N3C Consortium*

[1]Department of Biomedical Informatics & Medical Education, University of Washington, Seattle, WA, USA
[2]Division of General Medical Sciences, School of Medicine, Washington University in St. Louis, St. Louis, MO, USA
[3]Institute for Informatics, School of Medicine, Washington University in St. Louis, St. Louis, MO, USA
[4]MDClone Ltd., Beer Sheva, Israel
[5]UW Medicine, Seattle, WA, USA
*N3C Consortium authors listed in acknowledgements

Corresponding Author
Jason A. Thomas, BS
PhD Candidate
Department of Biomedical Informatics & Medical Education
University of Washington
Box 358047
Seattle, WA 98195
c: 208.559.6123
thomasjt@uw.edu

Keywords: data utility, data sharing, synthetic data, COVID-19, SARS-CoV-2, electronic health records

ABSTRACT

Objective: To evaluate whether synthetic data derived from a national COVID-19 data set could be used for geospatial and temporal epidemic analyses.
Materials and Methods: Using an original data set (n=1,854,968 SARS-CoV-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=5,819), 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 data sets. 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.

Predicting COVID-19 Regional Case Loads

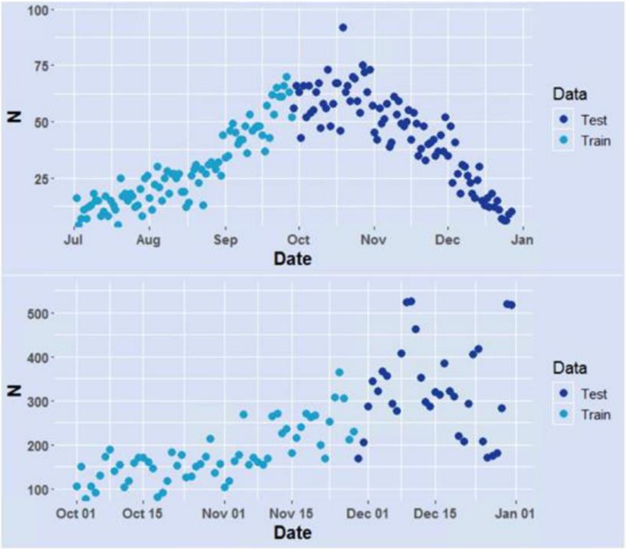


Figure 1: Total number of daily cases. Top: MA, Bottom: NC.

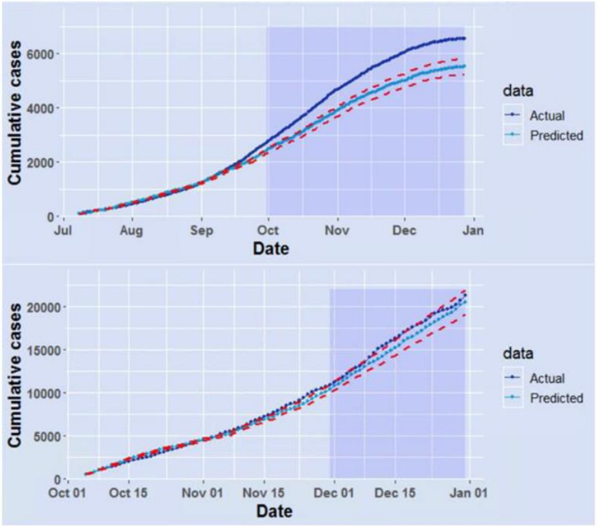


Figure 2: Training set and prediction window (light blue) for cumulative cases. Confidence intervals are indicated with dashed red lines.

Introduction: Need to predict future disease trajectories given current cases, hospitalizations, and deaths

Methods: Predicted daily infection rates and attempted to locate a “peak” of cases using a delayed elasticity model

Results: Our method correctly predicted the change in slope for the epidemic curves in 2 states (MA and NC)

Discussion: We demonstrated that emergency department incidence rates were able to be used to predict a peak in the curve (MA) and a flattening of the curve where a peak did not exist (NC)

SYNTHETIC DATA

"...empowers researchers to produce valid results, over a short period of time, while protecting patient privacy."

- Synthetic data are useful for research; statistical validation and privacy-preserving evaluations should persist
- Inconsistent data quality and biases in the source data are significant issues that are not unique to synthetic data

Randi Foraker, PhD, MA, FAHA, FAMIA

Professor, General Medical Sciences

Director, Center for Population Health Informatics, Institute for Informatics

Director, Public Health Data and Training Center, Institute for Public Health

Director, Center for Administrative Data Research, Washington University in St. Louis

Washington University School of Medicine

Campus Box 8102 | Sixth Floor, 4444 Forest Park Avenue | St. Louis, MO 63108

randi.foraker@wustl.edu