MEDINFO23

8 – 12 JULY 2023 | SYDNEY, AUSTRALIA

Expanding Information Accessibility: Synthetic Data

Washington University in St. Louis

The Ottawa Hospital | L'Hôpital d'Ottawa

NYU

Randi E. Foraker, Ph.D., M.A.
Adam Wilcox, Ph.D.
Alan Forster, M.D., M.Sc.
Zachary Abrams, Ph.D.
Jon D. Morrow, M.D., M.A., M.B.A.

AIDH IMIA

Session Objectives

- The panellists will explore the current and future state of synthetic data, highlighting their ability to support data sharing, to address privacy and confidentiality, and to advance national and international initiatives.
- Panellists will address the following topics:
  - Statistical validation
  - Privacy validation
  - Enhancing data sharing and collaboration
  - Industry partnerships
  - Use in education and training programs
  - Methodology and process of synthetic data generation

Panellists

Zachary Abrams, Ph.D.
Washington University in St. Louis

Randi E. Foraker, Ph.D.
Washington University in St. Louis

Alan Forster, M.D.
The Ottawa Hospital

Jon D. Morrow, M.D.
New York University

Adam Wilcox, Ph.D.
Washington University in St. Louis

Session Agenda

Dr Morrow	**"Synthetic Data Primer"** Synthetic data background, definitions, methods
Prof Foraker	**"Validation of Synthetic Data"** Statistical validation, privacy, and research use-cases
Dr Forster	**"Creating a Learning System"** Democratisation of data; collaborative partnerships
Prof Abrams Prof Foraker Dr Forster Dr Morrow Prof Wilcox	**Panel discussion** **Q&A**

MEDINFO23

8 – 12 JULY 2023 | SYDNEY, AUSTRALIA

Synthetic Data Primer

NYU

Jon D. Morrow, MD, MA, MBA, FACOG

Clinical Associate Professor

Department of Obstetrics & Gynecology
New York University School of Medicine

AIDH IMIA

Disclosure

Jon D. Morrow, M.D., was previously Senior Vice President at MDClone, Inc., a health IT software company whose synthetic data product was used in some of the referenced studies.

Why use synthetic health data?

- Personal health data are highly protected and sensitive.
- Health data contain a vast amount of valuable information.
- *Data-to-knowledge* work is often done by people not privy to protected information (e.g., non-clinicians, consortium members, and external commercial partners).

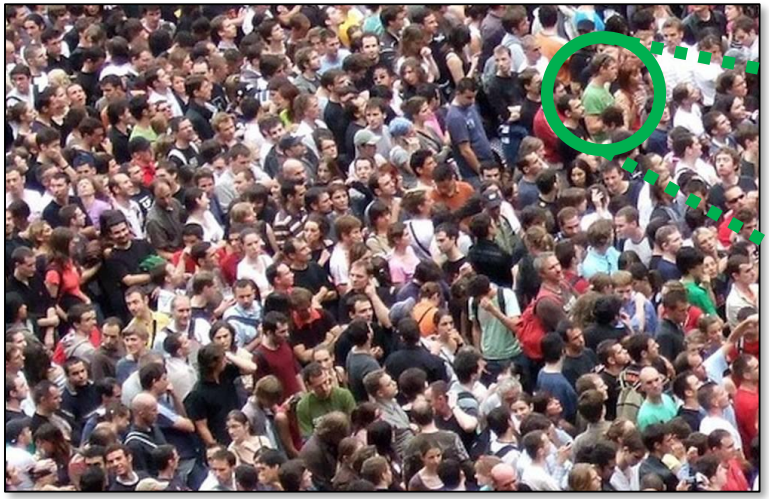
What are synthetic data?

- Maintain the utility, statistical properties, correlations, and higher-order relationships of real data sets without containing or exposing the members of the original set.
- Same format and suitability for analysis as the original.
- No one-to-one correspondence between members of the original and synthetic data sets.

- A form of data anonymisation, but:

Synthetic data ≠ Deidentified data

ORIGINAL



SYNTHETIC

ORIGINAL

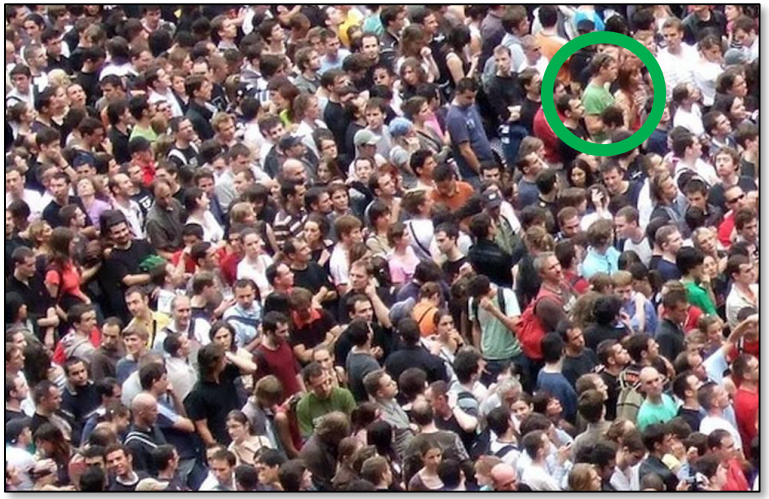
N	2,255
Age, mean ± SD	32.5 ± 6.7 years
Height, mean ± SD	171.1 ± 6.35 cm
Gender	62.1% male 36.9% female 1.0% non-binary
DM	8.2%
BMI ≥ 30 kg/m^2	35.1%
DM among BMI < 30 DM among BMI ≥ 30	6.9% 12.0%

SYNTHETIC

N	2,257
Age, mean ± SD	32.4 ± 6.7 years
Height, mean ± SD	170.9 ± 6.32 cm
Gender	62.0% male 37.1% female 0.9% non-binary
DM	8.2%
BMI ≥ 30 kg/m^2	35.3%
DM among BMI < 30 DM among BMI ≥ 30	7.0% 12.1%

SD = Standard deviation; DM = Diabetes mellitus; BMI = Body mass index.

ORIGINAL



N	2,255
Age, mean ± SD	32.5 ± 6.7 years
Height, mean ± SD	171.1 ± 6.35 cm
Gender	62.1% male 36.9% female 1.0% non-binary
DM	8.2%
BMI ≥ 30 kg/m^2	35.1%
DM among BMI < 30 DM among BMI ≥ 30	6.9% 12.0%

DEIDENTIFIED

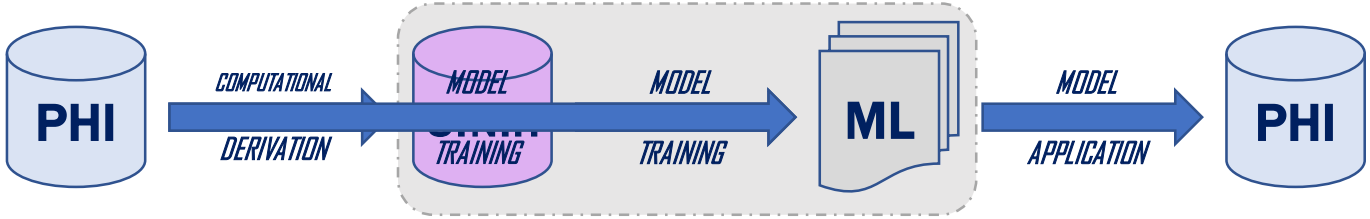
N	2,255
Age, mean ± SD	30.9 ± 5.8 years
Height, mean ± SD	171.1 ± 6.35 cm
Gender	62.1% male 36.9% female 1.0% non-binary
DM	8.2%
BMI ≥ 30 kg/m^2	35.1%
DM among BMI < 30 DM among BMI ≥ 30	6.9% 12.0%

SD = Standard deviation; DM = Diabetes mellitus; BMI = Body mass index.

Simulation vs computation

- **Simulated synthetic data**: Probabilistic synthesis to create large data sets, useful for simulation, systems testing, training, and other uses *(e.g., Synthea®).*
- **Computationally derived synthetic data**: Novel data set, usually (but not necessarily) approximately the same size as the original, populated with new data points to match the original's statistical properties *(e.g., MDClone ADAMS™).*

Synthetic data for AI



- ML models trained on computationally derived synthetic data are equally valid as models trained on the original source data from which the synthetic data were derived.
- This allows ML models to be trained without exposing the original patients' personal information and without compromising patient privacy.

AI = Artificial intelligence; PHI = Protected health information; ML = Machine learning.

MEDINFO23

8 – 12 JULY 2023 | SYDNEY, AUSTRALIA

Validation of Synthetic Data

Washington University in St. Louis

Randi Foraker, PhD, MA, FAHA, FAMIA, FACMI

Director, Center for Population Health

Institute for Informatics (I2)

Washington University in St. Louis

AIDH IMIA

@TheInstituteDH #MEDINFO23

Validation studies: Questions answered

Privacy Studies (n=5)	Statistical Studies (n=6)
Does anyone look the same?	Does it look the same?
Can you automatically identify people?	Does it work the same?
Can you manually identify people?	Does it work better than other approaches?
It is a problem if people are identified?	What can we do differently with synthetic data?
Does it solve the problem of linked data sets?	

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

Performance metrics of sepsis prediction models

Training set		Real	Synthetic	Synthetic
Testing set		Real	Synthetic	Real
Train	Accuracy	0.845	0.869	0.852
	Precision	0.803	0.840	0.812
	Recall	0.704	0.758	0.719
	F1	0.750	0.797	0.763
	AUROC	0.809	0.842	0.818
5-fold cross-validation	Accuracy	0.795	0.802	0.799
	Precision	0.712	0.73	0.723
	Recall	0.637	0.67	0.639
	F1	0.672	0.69	0.678
	AUROC	0.855	0.86	0.847
Test	Accuracy	0.834	0.833	0.834
	Precision	0.811	0.759	0.829
	Recall	0.677	0.678	0.654
	F1	0.738	0.716	0.731
	AUROC	0.887	0.885	0.892

JAMIA Open, Volume 3, Issue 4, December 2020, Pages 557–566,
doi:10.1093/jamiaopen/ooaa060

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)

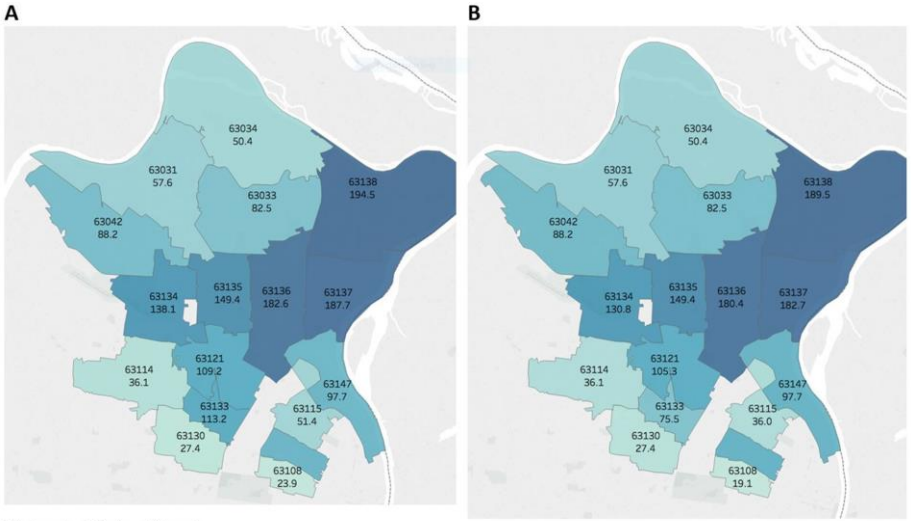
Received 14 August 2020; Revised 14 October 2020; Accepted 20 October 2020

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

Chlamydia rates (per 100 000 persons) by zip code: real (left) versus synthetic (right) data.



*Darker color indicates a higher rate

JAMIA Open, Volume 3, Issue 4, December 2020, Pages 557–566,
doi:10.1093/jamiaopen/ooaa060

JOURNAL OF MEDICAL INTERNET RESEARCH Foraker et al

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[5], PhD; N3C Collaborative[6]

[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 2511
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.

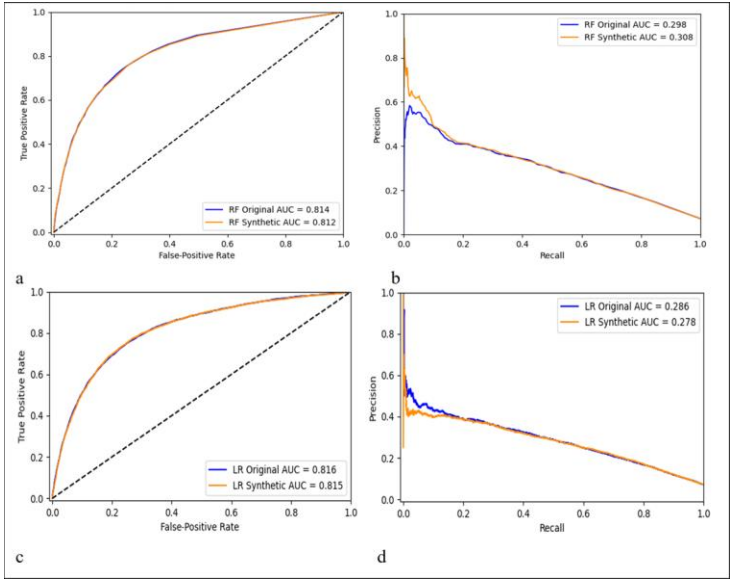
Objective: We aim to compare the results from analyses of synthetic data to those from original data and assess the strengths and limitations of leveraging computationally derived data for research purposes.

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 univariable 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.

Prediction performance for the two models by receiver operating characteristic curves (A, C) and precision-recall curves (B, D) by using original and synthetic data.



J Med Internet Res. 2021;23(10):e30697. doi:10.2196/30697. PMID: 34559671; PMCID: PMC8491642.

JOURNAL OF MEDICAL INTERNET RESEARCH Foraker et al

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 2511
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.

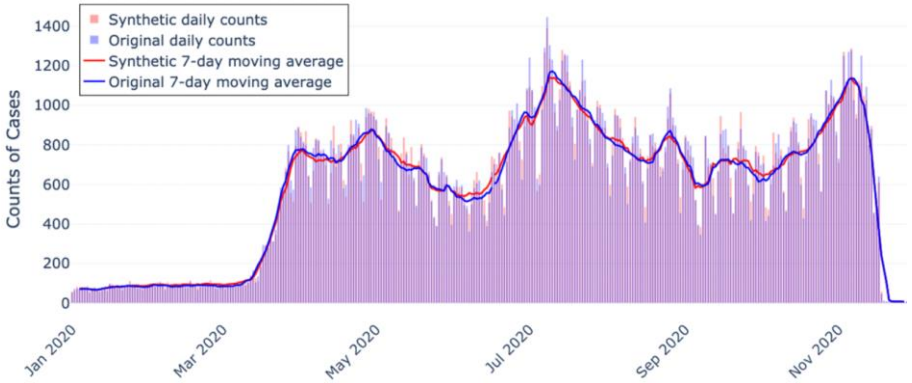
Objective: We aim to compare the results from analyses of synthetic data to those from original data and assess the strengths and limitations of leveraging computationally derived data for research purposes.

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 univariable 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.

All sites positive tests (cases)



J Med Internet Res. 2021;23(10):e30697. doi:10.2196/30697. PMID: 34559671; PMCID: PMC8491642.

SYNTHETIC DATA

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

MEDINFO23

8 – 12 JULY 2023 | SYDNEY, AUSTRALIA

Creating a Learning System

The Ottawa Hospital | L'Hôpital d'Ottawa

Alan Forster, MD, MSc

Executive Vice President of Innovation and Quality

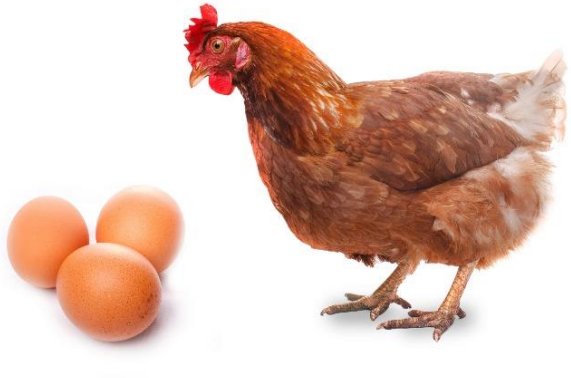
The Ottawa Hospital

AIDH IMIA

Evidence based decision making in healthcare: what is the problem?

Technical

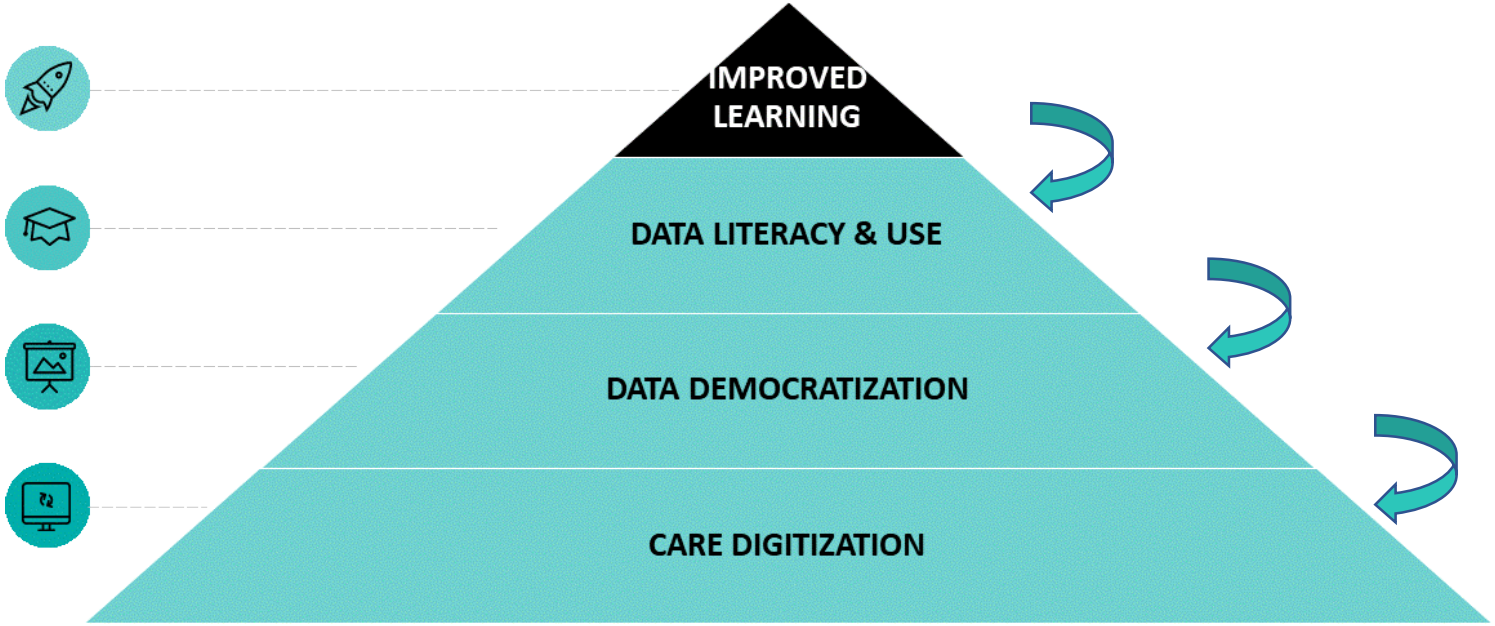
- Privacy
- Complexity
- Quality
- Completeness



Cultural

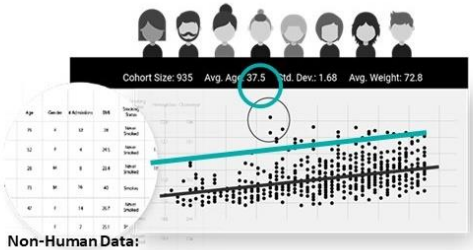
- Decision making
- Priorities
- Action tracking
- Trust

Capability Maturity



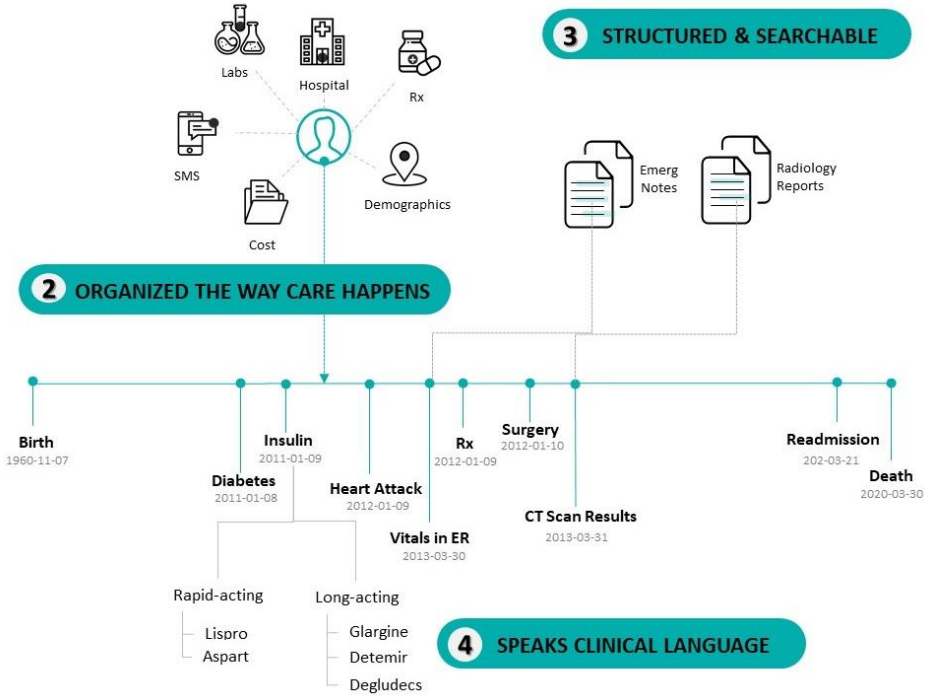
Data Democratization

Human Data:

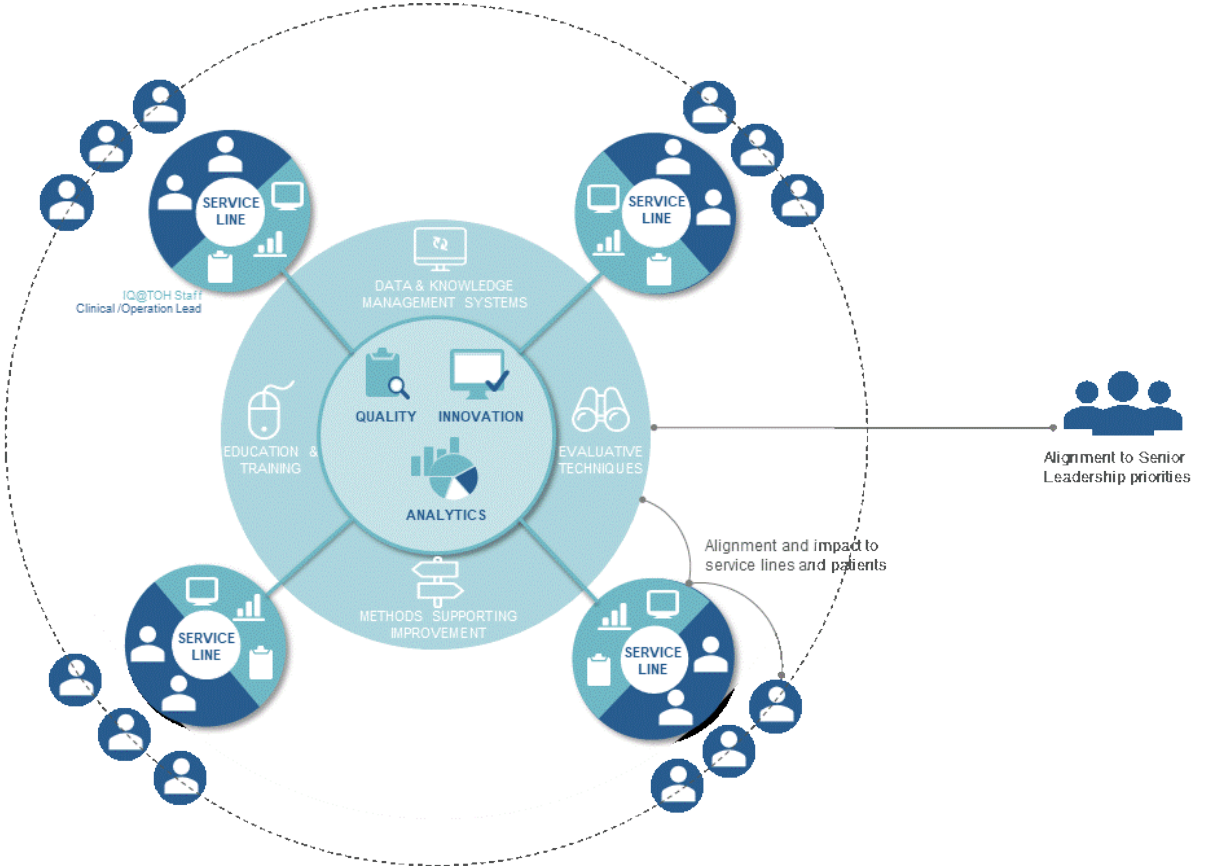


Non-Human Data:

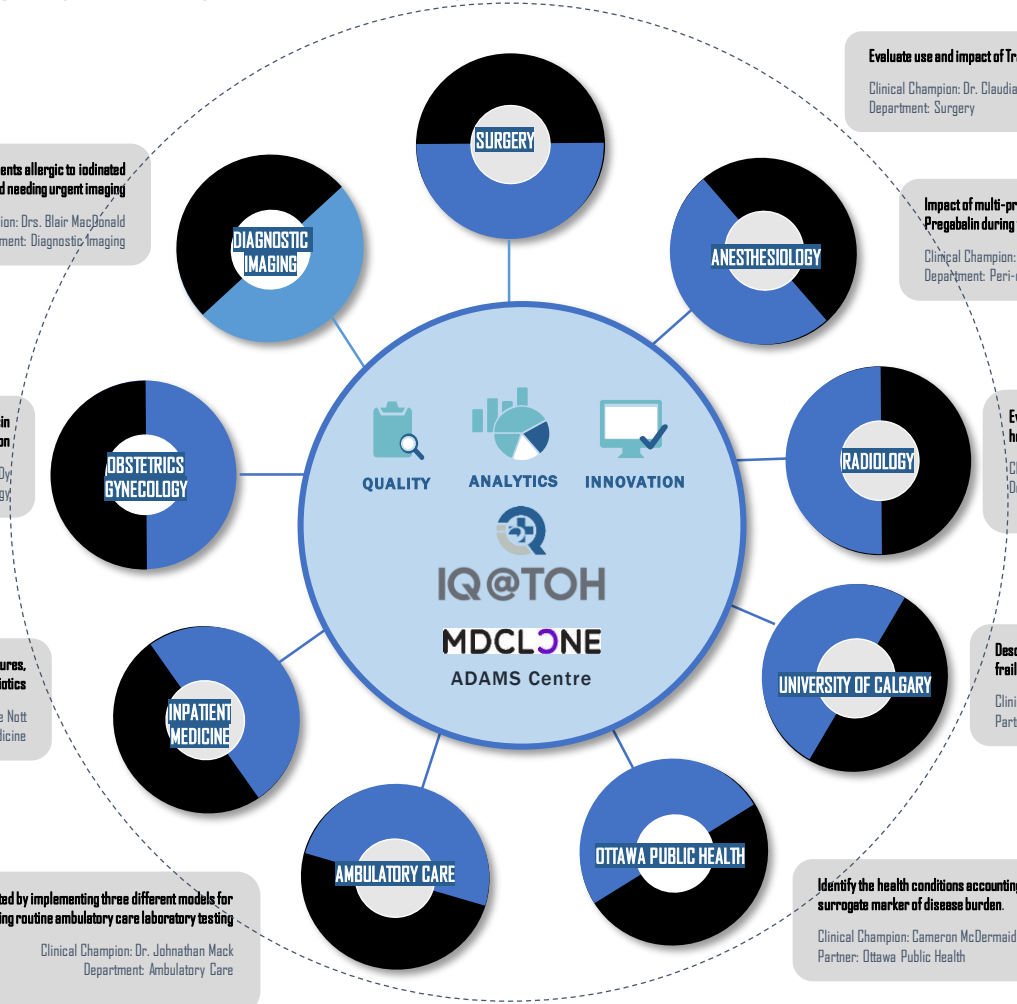
1 ACCESSIBLE & COMPLIANT



A new way of working together



DEVELOP USE CASES TO IMPROVE LEARNING



Evaluate use and impact of Tranexamic acid in surgical patients

Clinical Champion: Dr. Claudia Malic
Department: Surgery

Impact of multi-pronged approach to reduce inappropriate use of Pregabalin during the peri-op period

Clinical Champion: Dr. Sarah Tierney
Department: Peri-op Anesthesiology

Evaluate the value of CT scan of the head in anticoagulated patients with head injury, and compare to findings in Sheba Medical Centre

Clinical Champion: Dr. Santanu Chakraborty
Department: Radiology

Describe demographic and clinical characteristics of high versus low frailty groups, using the CIHI-HFRM

Clinical Champion: Dr. Karen Tang
Partner: University of Calgary

Identify the health conditions accounting for the majority of in-patient hospital care costs as a surrogate marker of disease burden.

Clinical Champion: Cameron McDermaid
Partner: Ottawa Public Health

Describe the clinical and financial opportunities created by implementing three different models for providing routine ambulatory care laboratory testing

Clinical Champion: Dr. Johnathan Mack
Department: Ambulatory Care

Impact of multi-pronged approach to reduce overuse of urine cultures, urinalysis and antibiotics

Clinical Champion: Drs. Derek Macfadden, Caroline Nott
Department: In-Patient Medicine

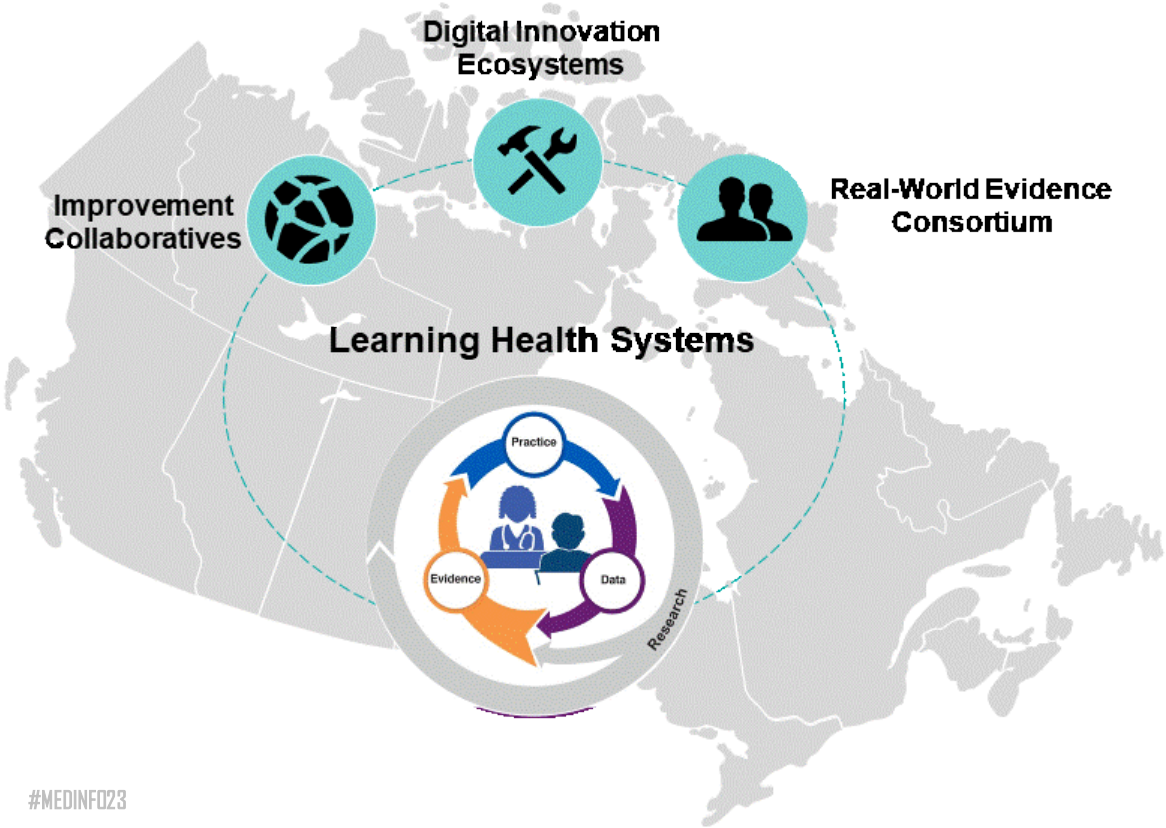
Evaluate use and impact Carbetocin in patients undergoing C-section

Clinical Champion: Dr. Jessica Dy
Department: Obstetrics and Gynecology

Define the business case for automatically substituting Provist in patients allergic to iodinated contrast media and needing urgent imaging

Clinical Champion: Drs. Blair MacDonald
Department: Diagnostic Imaging

Collaboration at scale



MEDINFO23

8 – 12 JULY 2023 | SYDNEY, AUSTRALIA

Panel Discussion

Moderator:
Randi E. Foraker, Ph.D.
Washington University in St. Louis

Zachary Abrams, Ph.D.
Washington University in St. Louis

Alan Forster, M.D.
The Ottawa Hospital

Jon D. Morrow, M.D.
New York University

Adam Wilcox, Ph.D.
Washington University in St. Louis

MEDINFO23

8 – 12 JULY 2023 | SYDNEY, AUSTRALIA

Washington University in St. Louis

The Ottawa Hospital | L'Hôpital d'Ottawa

NYU

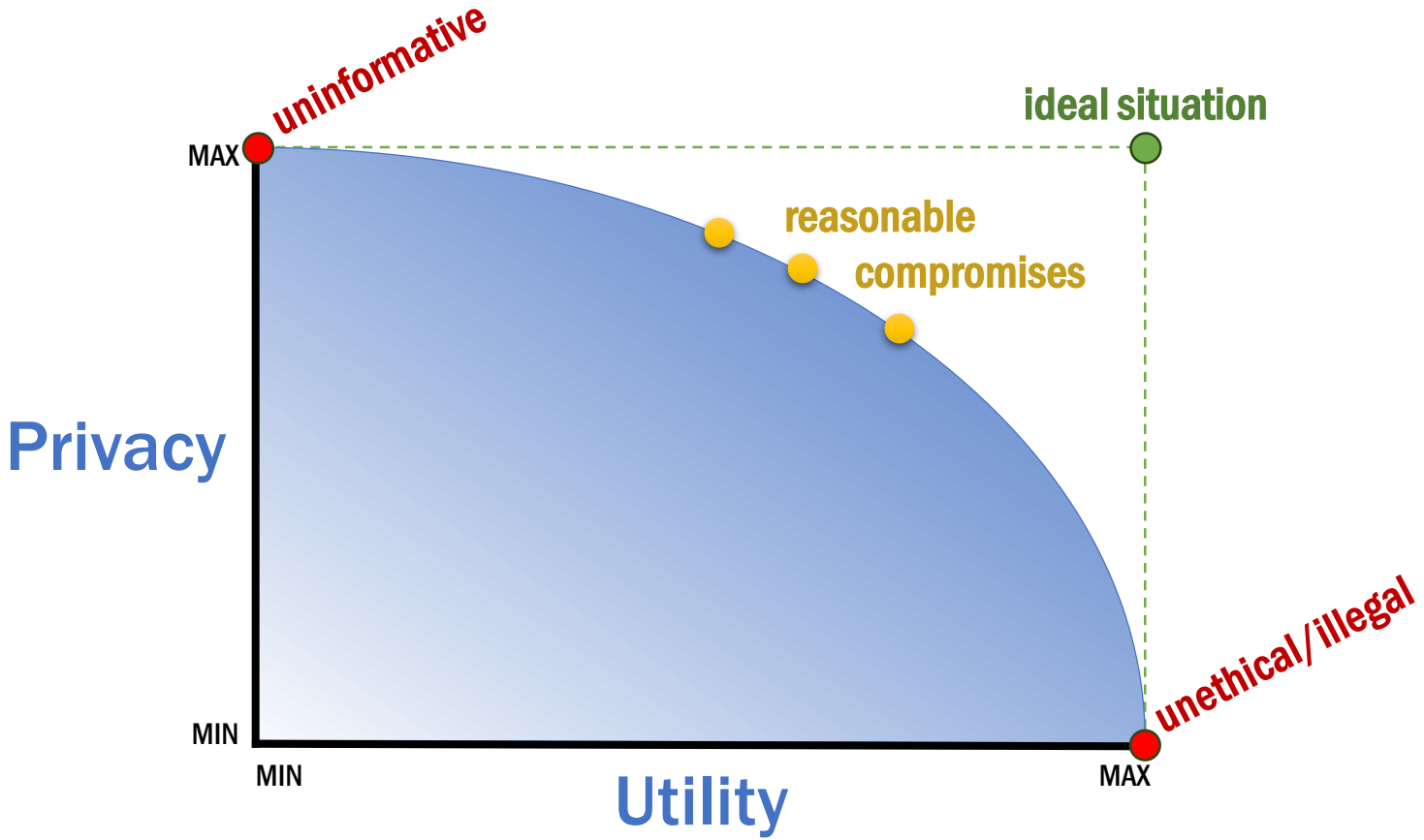
Thank you!

randi.foraker@wustl.edu

@TheInstituteDH #MEDINFO23

AIDH IMIA

RESERVE SLIDES:
Jon


uninformative
ideal situation
MAX
reasonable
compromises
Privacy
MIN
unethical/illegal
MIN
MAX
Utility

Formal definition of synthetic data

- "Synthetic data are microdata records created to **improve data utility** while **preventing disclosure** of confidential respondent information."
- "Synthetic data is created by **statistically modeling original data** and then using those models to generate **new data values** that reproduce the **original data's statistical properties.**"

Philpott D. *A Guide to Federal Terms and Acronyms.* Bernan Press, 2017:184.

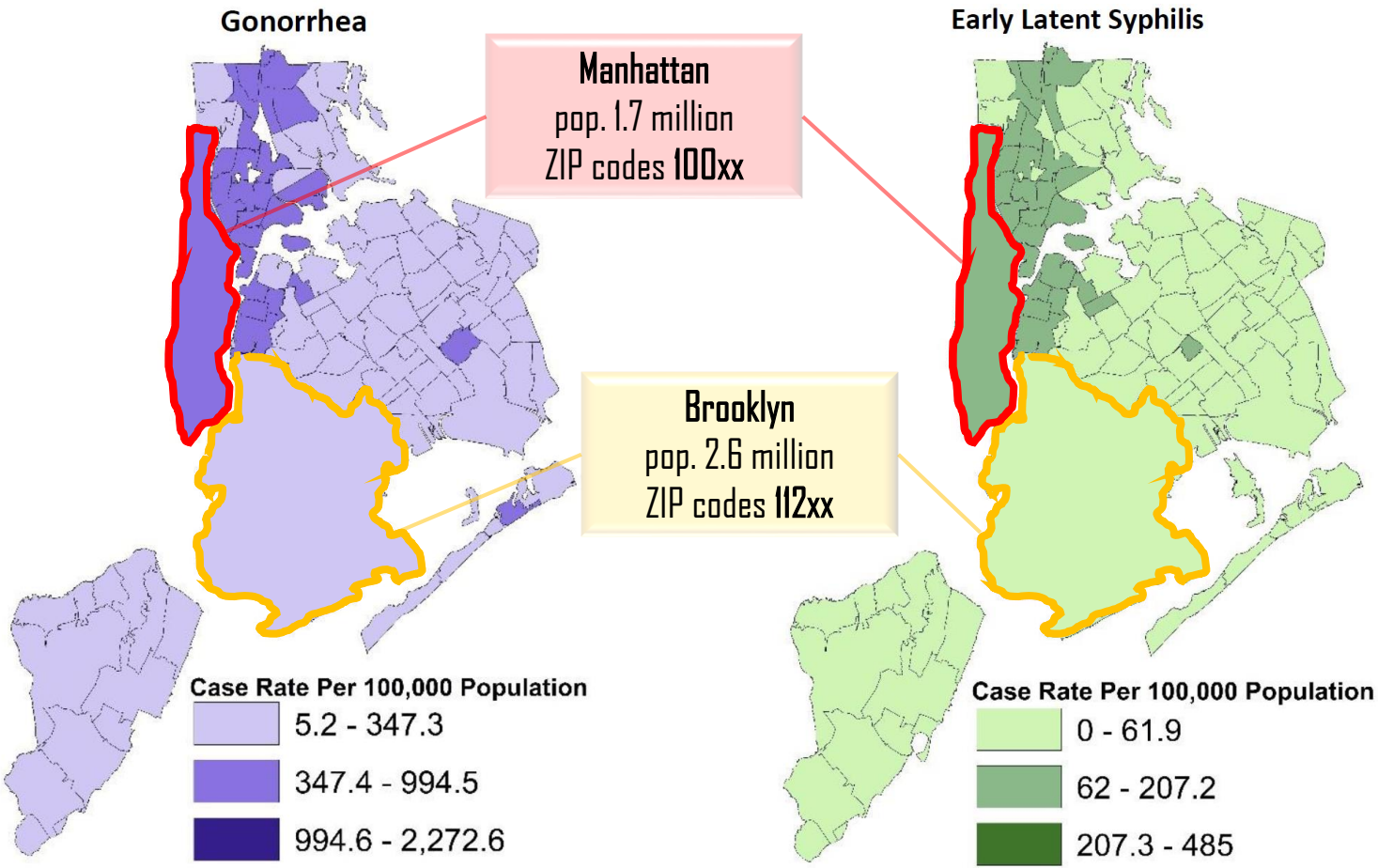
ORIGINAL

N	2,255
Age, mean ± SD	32.5 ± 6.7 years
Height, mean ± SD	171.1 ± 6.35 cm
Gender	62.1% male 36.9% female 1.0% non-binary
DM	8.2%
BMI ≥ 30 kg/m^2	35.1%
DM among BMI < 30 DM among BMI ≥ 30	6.9% 12.0%

DEIDENTIFIED

N	2,255
Age, mean ± SD	30.9 ± 5.8 years
Height, mean ± SD	171.1 ± 6.35 cm
Gender	62.1% male 36.9% female 1.0% non-binary
DM	8.2%
BMI ≥ 30 kg/m^2	35.1%
DM among BMI < 30 DM among BMI ≥ 30	6.9% 12.0%

SD = Standard deviation; DM = Diabetes mellitus; BMI = Body mass index.



Maps: NYC Dept. of Health and Mental Hygiene. NYC STI Surveillance Report, 2021.

Computational derivation overview

- First, derive statistical models from the original data set.
- Then, sample novel synthetic data points to fit the models.

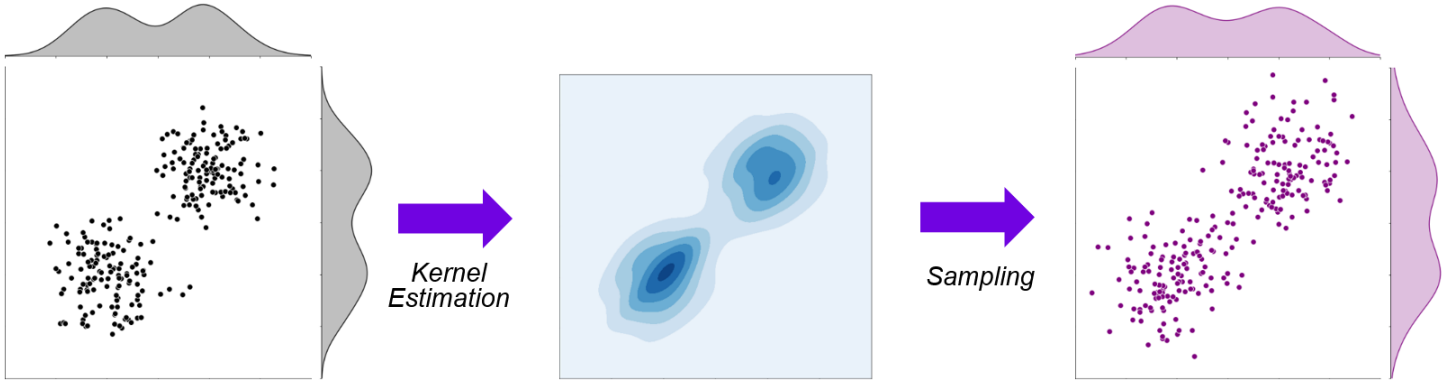


Figure adapted from: Foraker RE, Yu SC, Gupta A, et al. Spot the difference: comparing results of analyses from real patient data and synthetic derivatives. *JAMIA Open* 2020;3:557-66.

Computational derivation methodology

- For **categorical variables**, to mitigate the potential for an inference attack due to the finite number of categories:
  - Group synthetic individuals who share identical categories.
  - If any group contains $<\kappa$ members, censor some discrete values until all groups contain $\geq\kappa$ members.
  - Create clusters of $<\kappa$ individuals, minimizing the scaled Euclidean distance between data points.
  - Replace each cluster's numeric variables with an alternate matrix with similar statistical properties, preserving statistical characteristics for every pair of variables within each cluster.
  - The alternate matrix is selected randomly from the unlimited number of alternatives, resulting in an irreversible transformation.
- To protect against a difference attack, slightly alter the population size.

Thomas JA, Foraker RE, Zamstein N, Morrow JD, Payne PRO, Wilcox A. 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 (N3). *J Amer Med inform Assoc* 2022;29:1350-65.

RESERVE SLIDES:
Randi

https://informatics.wustl.edu/mdclone/

Lessons learned

- Synthetic data benefits:
  - Enhancing AI/ML training opportunities
  - Accelerating research
  - Facilitating data sharing
  - Expanding data access to community partners / health dashboards
- Next steps:
  - Expanding upon 'information gain' analyses
  - Establishing an STL data hub

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

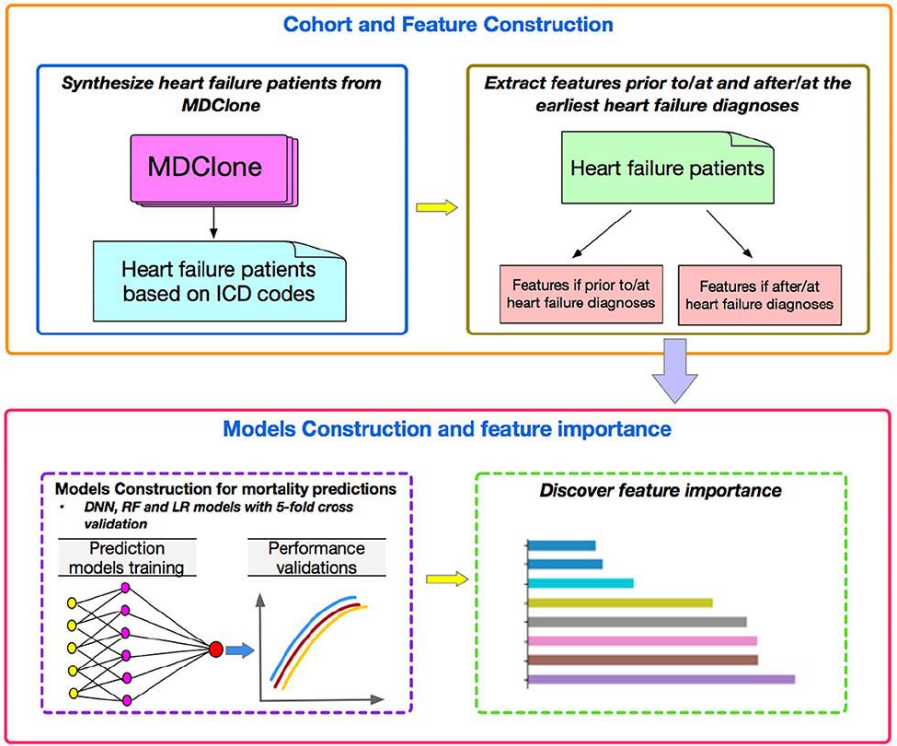
Received: *27 June 2020*
Accepted: *13 November 2020*
Published: *07 December 2020*

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.

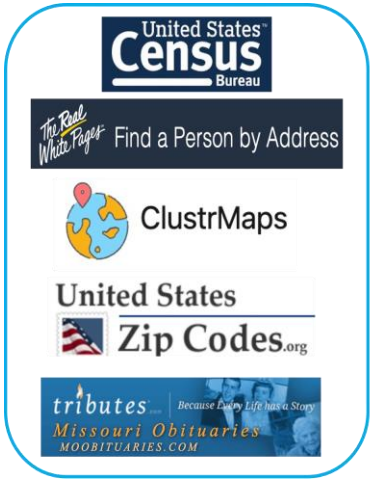
Methods: We developed *synthetic* data from EHR data of 26,575 HF patients admitted to a single institution during the decade ending on 12/31/2018. Twenty-seven clinically-relevant features were synthesized and utilized in supervised deep learning and machine learning algorithms (i.e., deep neural networks [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.

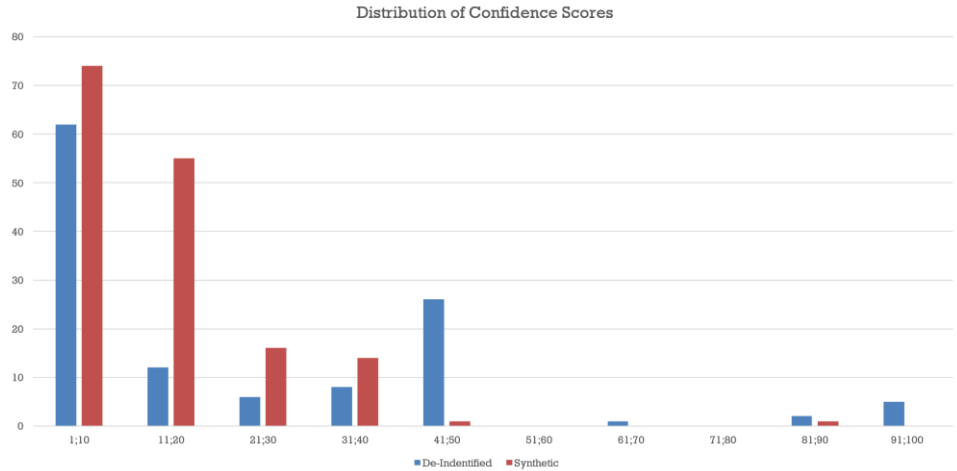


RESERVE SLIDES:
Zach

- We used a traditional adversarial approach to assess the privacy preserving nature of synthetic data as compared to de-identified data.

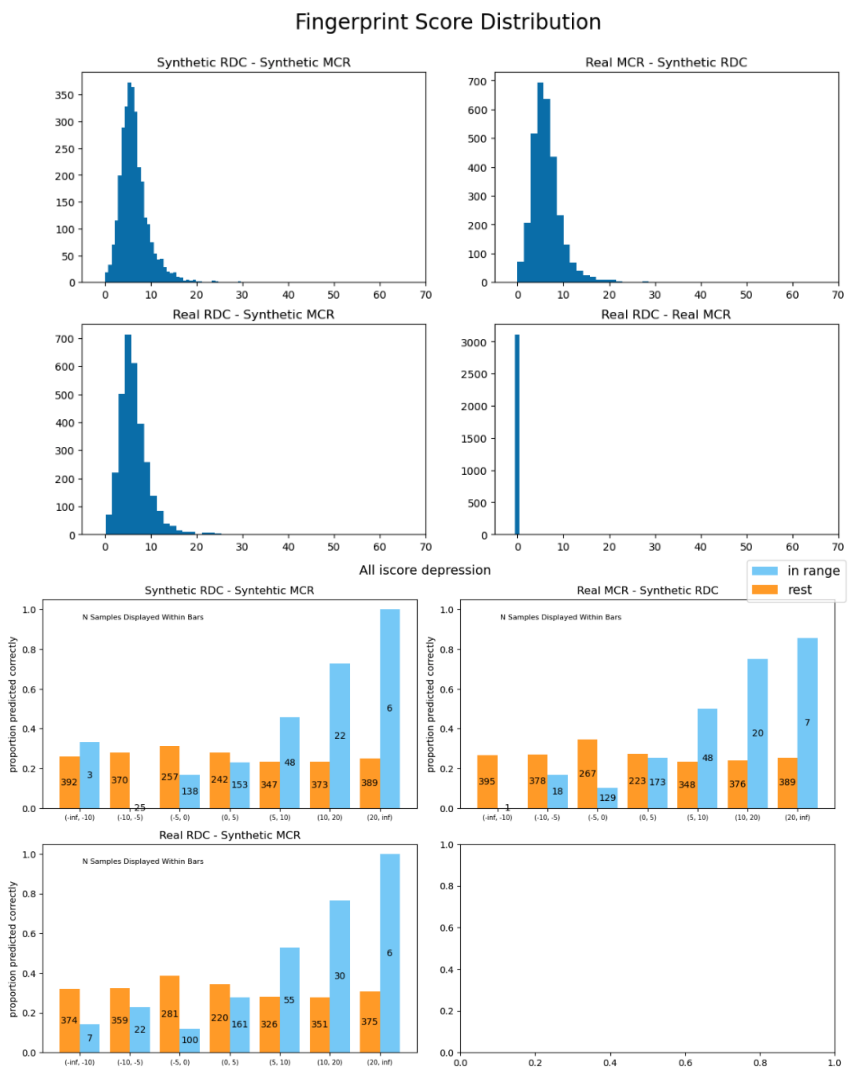


- Our results indicate that synthetic data cannot be confidently re-identified to the same level as de-identified data.



- We measured fingerprinting scores between sets of real and synthetic data to measure the amount of variability introduced via synthetic data.

- Measured the amount of information that could reliably be learned about real data from synthetic data.



RESERVE SLIDES:
Alan

NEXT STEPS

FINDINGS

IMPORTANCE

Postoperative Pregabalin Use

NEXT STEPS

FINDINGS

Prescribe

PREGABILIN

To Reduce

Caused some incidences of

Respiratory Distress

Cardiac Arrest

Delirium

Clinical Interventions taken

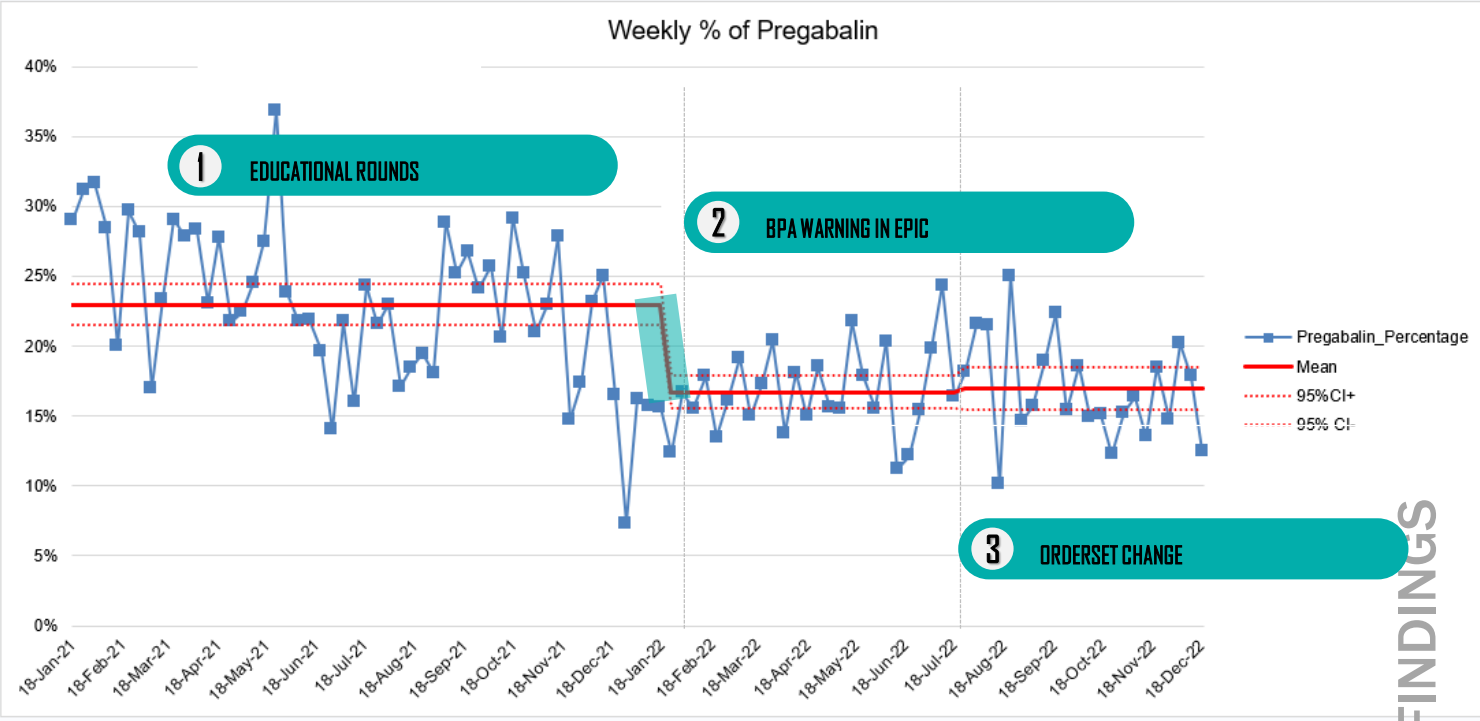
1. EDUCATIONAL ROUNDS
2. BPA WARNING IN EPIC
3. ORDERSET CHANGE

IMPORTANCE

NEXT STEPS

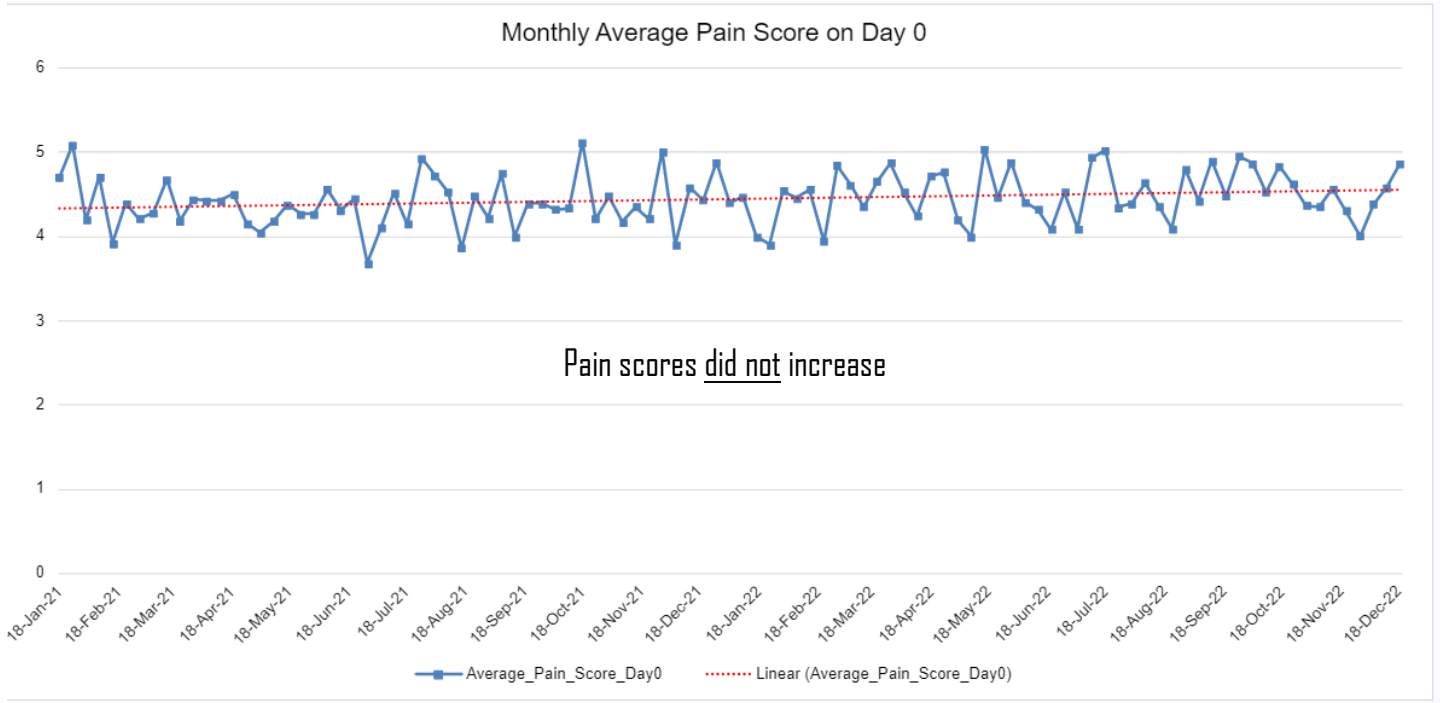
IMPORTANCE

FINDINGS



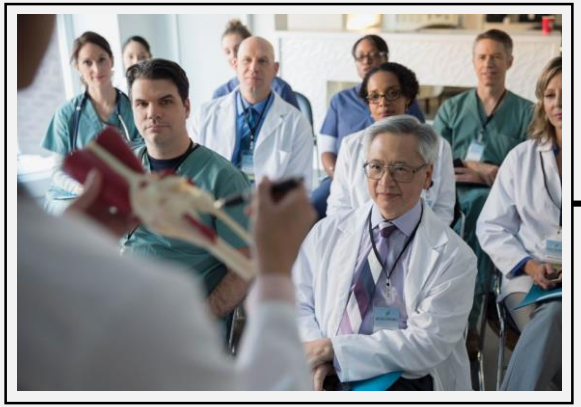
There is already a downward trend of pregabalin-use before the technology intervention, therefore cannot attribute impact to BPA

NEXT STEPS



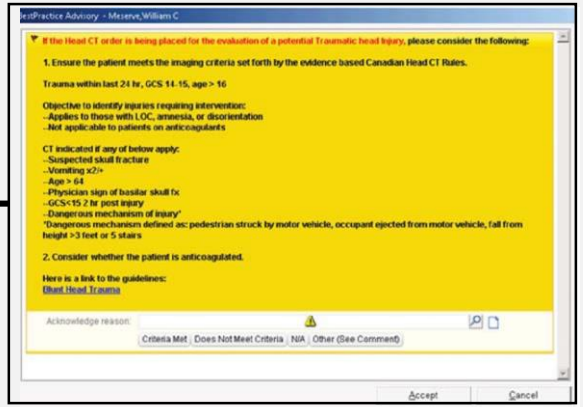
Pain scores did not increase

IMPORTANCE



EDUCATION

Vs.



BPA/DECISION SUPPORT

Key Takeaway: Need to be more thoughtful in implementing decision support, and it is helpful to measure patient impact

Next Step: Work with decision support team and clinicians

NEXT STEPS

FINDINGS

IMPORTANCE

NEXT STEPS

FINDINGS

IMPORTANCE

Radiocontrast Switching in ED

NEXT STEPS

FINDINGS

NO ALLERGY

RADIOCONTRAST MEDIA

HISTORY OF ALLERGY

13 HOUR STERIOD & ANTIHISTAMINE PROTOCOL

↑ Length of Stay

↑ Infection Rates

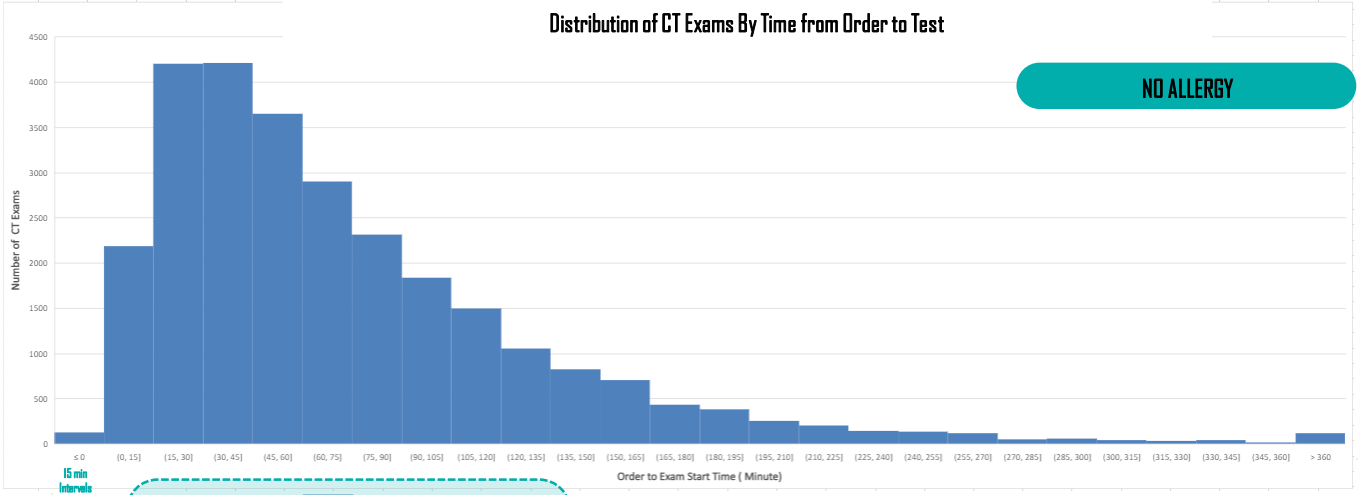
IMPORTANCE

Current evidence does not support the use of steroids because they are not proven superior to contrast switching
(switch to a different low-osmolar/iso-osmolar contrast agent)

NEXT STEPS

IMPORTANCE

FINDINGS



NEXT STEPS

IMPORTANCE

FINDINGS

Distribution of CT Exams By Time from Order to Test

HISTORY OF ALLERGY

~ 400 VISITS ANNUALLY

New contrast will

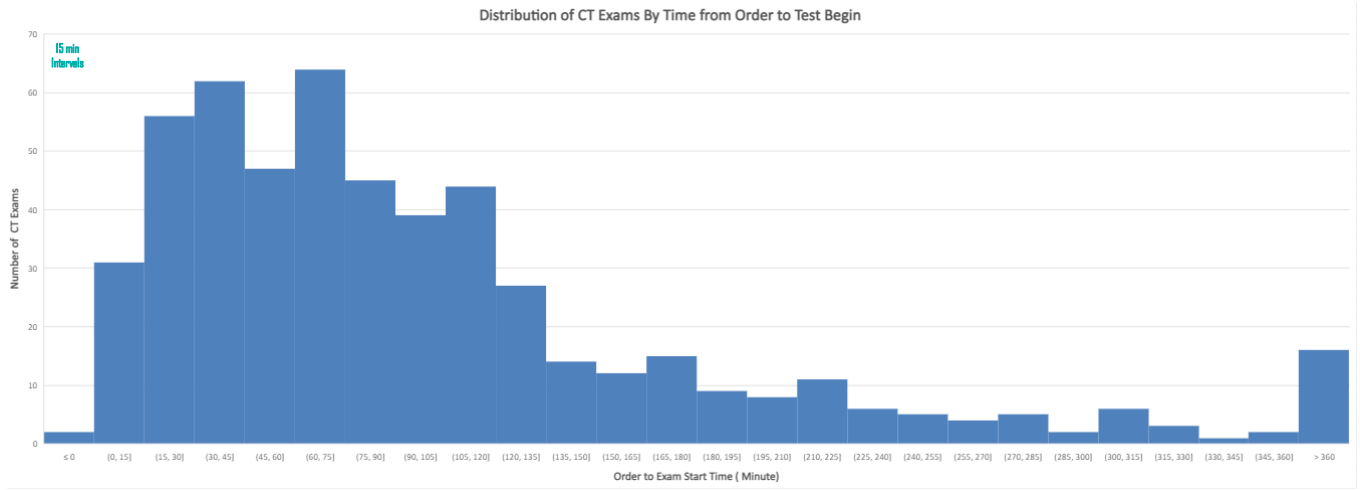
SAVE 2 HRS

Equates to

800 NURSNG HOURS SAVED

Minus cost of new contrast

~ 28K SAVED PER YEAR



Key Takeaway: The better option for patients and staff is also better for the bottom line

Next Step: Work with clinical leaders to implement, monitor benefits

FINDINGS

IMPORTANCE

NEXT STEPS

NEXT STEPS

FINDINGS

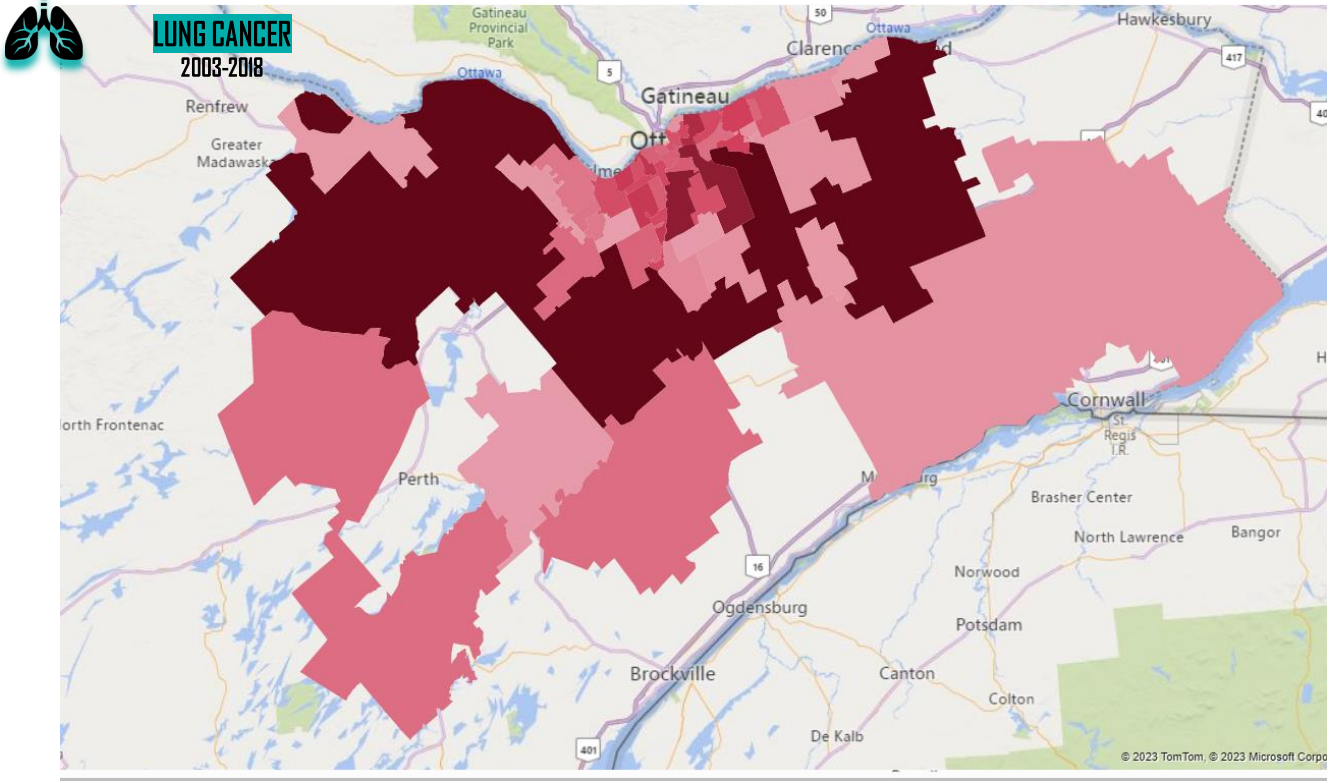
IMPORTANCE

Causes of Hospitalizations in Ottawa

NEXT STEPS

FINDINGS

IMPORTANCE



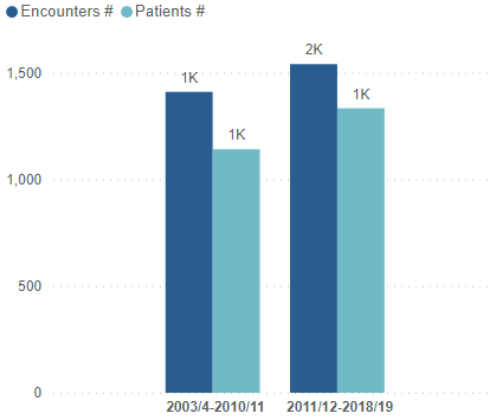
NEXT STEPS

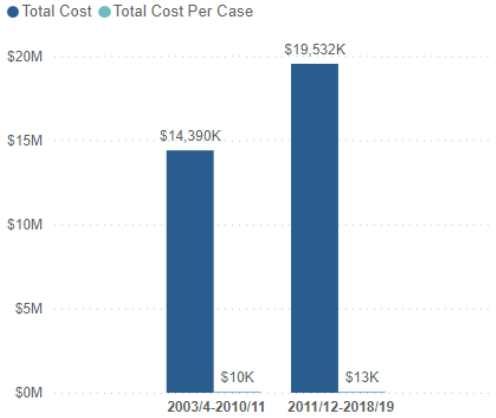
IMPORTANCE

FINDINGS

Cancers of the Urinary Tract

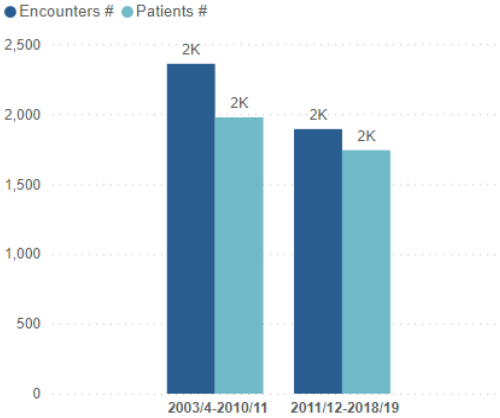
↑ Patients, Cases & Cost

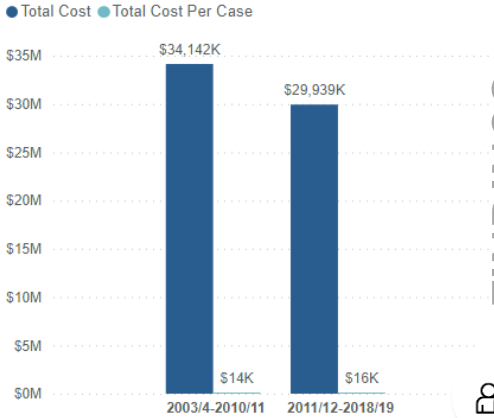




Cancers of the Lung

↓ Patients, Cases & Cost





- Extend analysis to determine impact of EMR, covid-19, and other factors
- Assess other diseases
- Further analysis to determine possible explanations for decreasing cases and lung cancer trends
- Focused efforts within FSAs to address possible SDOH → impact of screening, smoking behaviour, other environmental risks

NEXT STEPS

FINDINGS

IMPORTANCE