The Impact of Hallucinations in Synthetic Health Data on Prognostic Machine Learning Models

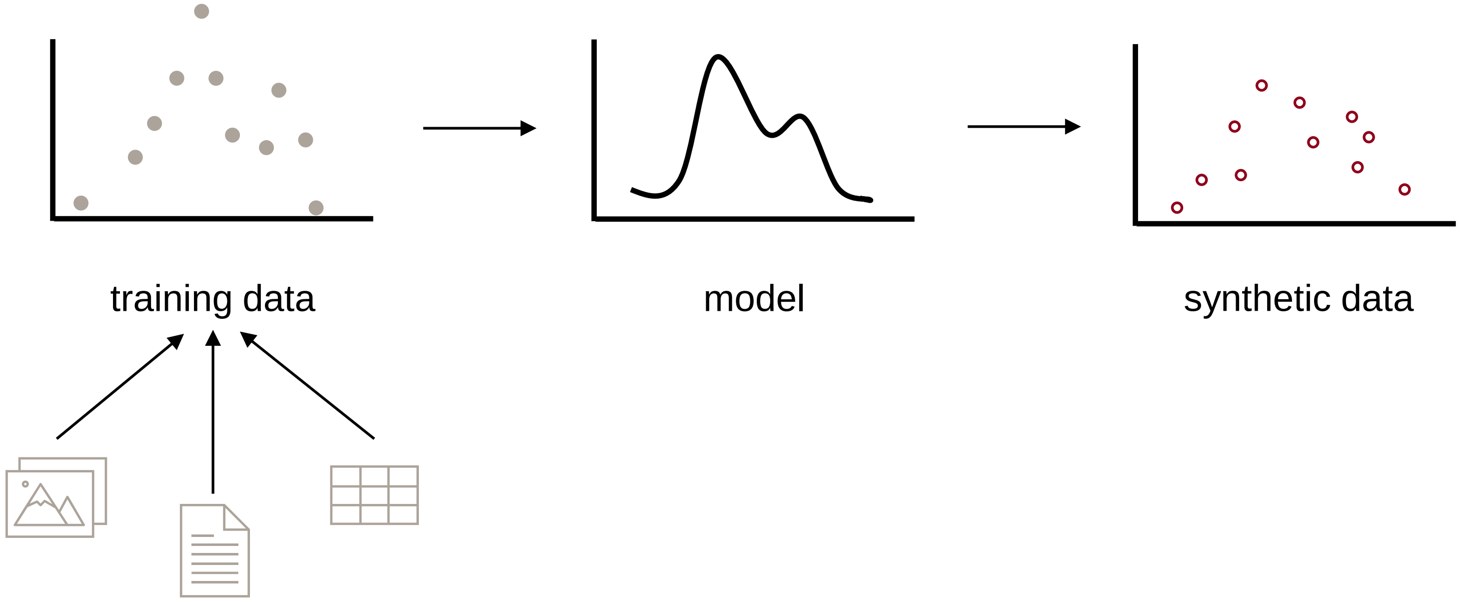
Lisa Pilgram, MD
Postdoctoral Fellow at the Electronic Health Information Laboratory (Khaled El Emam)

Agenda

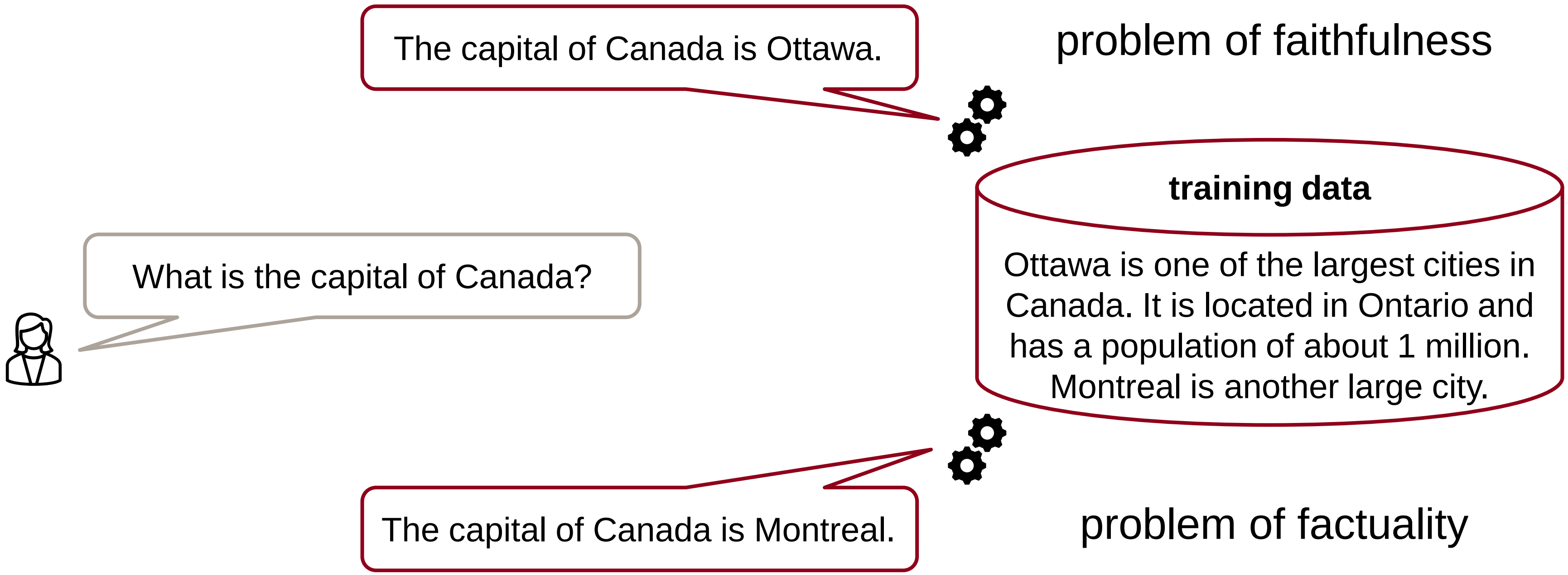
Hallucinations in Generative Modeling	1	What are generative models and hallucinations?
Methodology	2	How to generate synthetic health data and measure hallucinations?
Results	3	When do hallucinations occur and how do they impact prognostic ML models?
Conclusions	4	Understanding the implications and limitations of our results

HALLUCINATIONS IN GENERATIVE MODELING

Generative Modeling

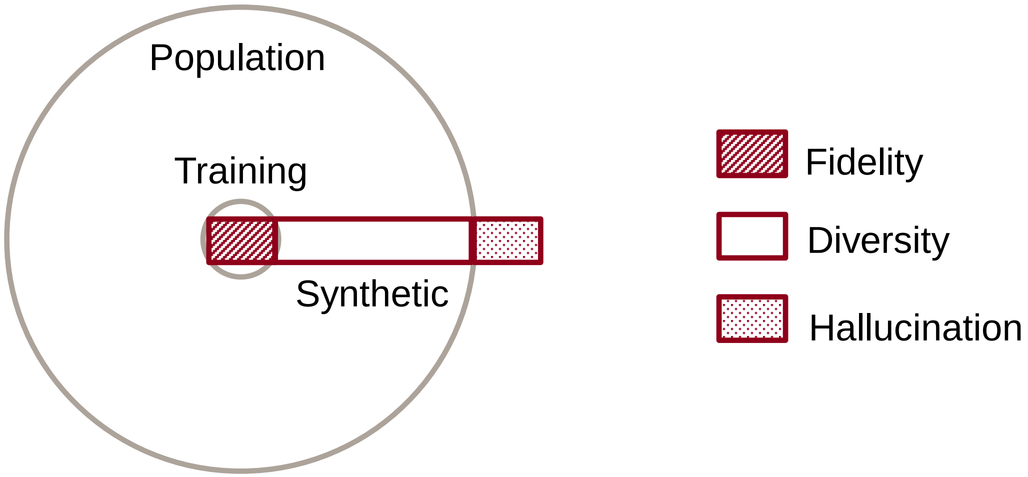


Hallucinations in Text Generation



Hallucinations in Tabular Synthetic Health Data

- Problem of factuality: Hallucinated patients are synthetic patients that are non-existent (or implausible) in the reference population.



The Impact of Hallucinations in Synthetic Health Data on Prognostic ML Models

1. What is the hallucination rate (HR) during tabular synthetic health data generation (SDG)?

2. Does the magnitude of the HR in synthetic health data affect the performance of downstream prognostic ML models?

METHODOLOGY

Identification of Hallucinated Patients

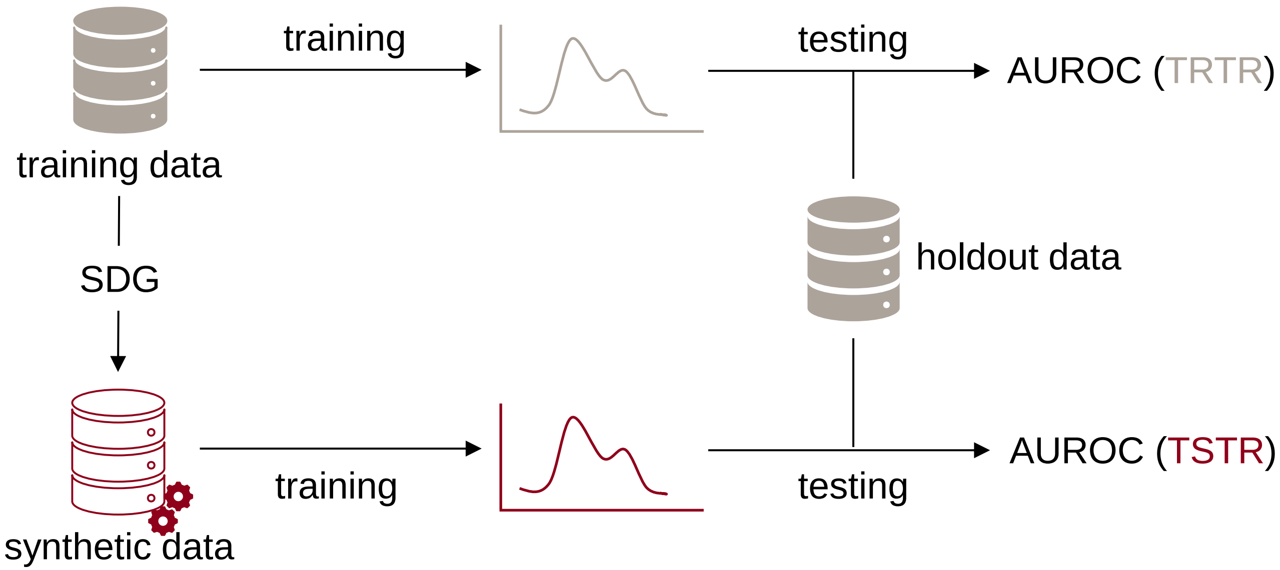
- Hallucinations are synthetic patients (i.e., x_s) that are non-existent in the population, meaning that they have a non-zero (i.e., $\tau = 0$) distance from all population records (i.e., x_r).

$$\min d(x_s, x_r) > \tau$$

- The hallucination rate (HR) is the proportion of hallucinated patients among all synthetic patients.

Downstream Utility: Prognostic AI/ML Modeling

- Train-synthetic-test-real (TSTR) is when a prognostic AI/ML model is trained on the synthetic data and then tested on unseen real records.
- Train-real-test-real (TRTR) is when a prognostic AI/ML model is trained on the (real) training data and then tested on unseen real records.



Study Set Up

- 12 real world health datasets
  - 6,354 population variants with varying complexity by changing the number and type of variables included
- 1 SDG model ("generator"): Sequential Trees (ST)
  - 1 trained SDG model per population variant
  - 10 synthetic health datasets per trained SDG model
- 1 prognostic AI/ML model: light gradient boosting machine
  - AUROC (TRTR)
  - $AUROC_{avg}$(TSTR)
- Mixed-effect models were used to estimate the fixed effect of HR on AI/ML model performance with the specific health dataset as random effect.

RESULTS

1 The Hallucination Rate in Synthetic Health Data

- Mean HR was 88.5% (SD 20.7%) in SDG via ST.
- Odds for hallucinations were higher with increasing complexity in SDG via ST. The large majority of health datasets (90.1%) were highly complex.

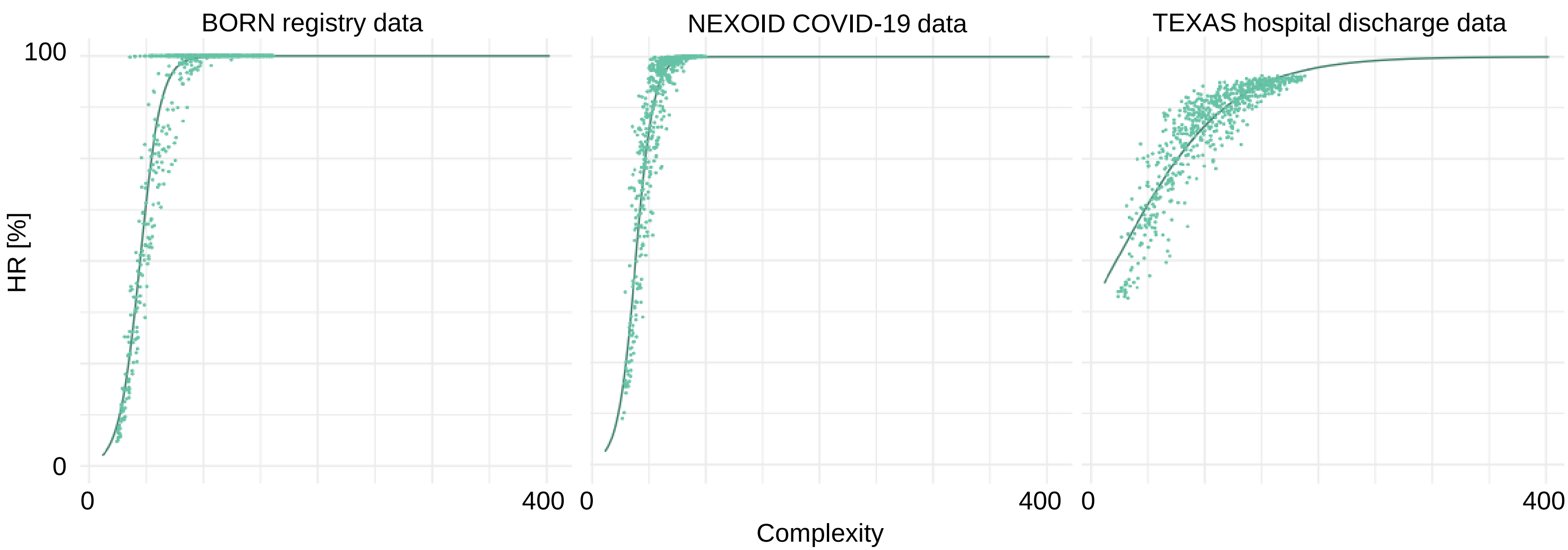


Figure 1. Exemplar mixed-effect model with the health dataset as random effect and complexity as fixed effect and HR as outcome for the ST SDG model. 3 out of 12 health datasets are shown as examples.

2 The Impact of Hallucinations on Prognostic ML Models

- AUROC (TRTR) – AUROC (TSTR) was 0.05 on average in SDG via ST.
- AUROC (TSTR) did not change with increasing HR in SDG via ST.

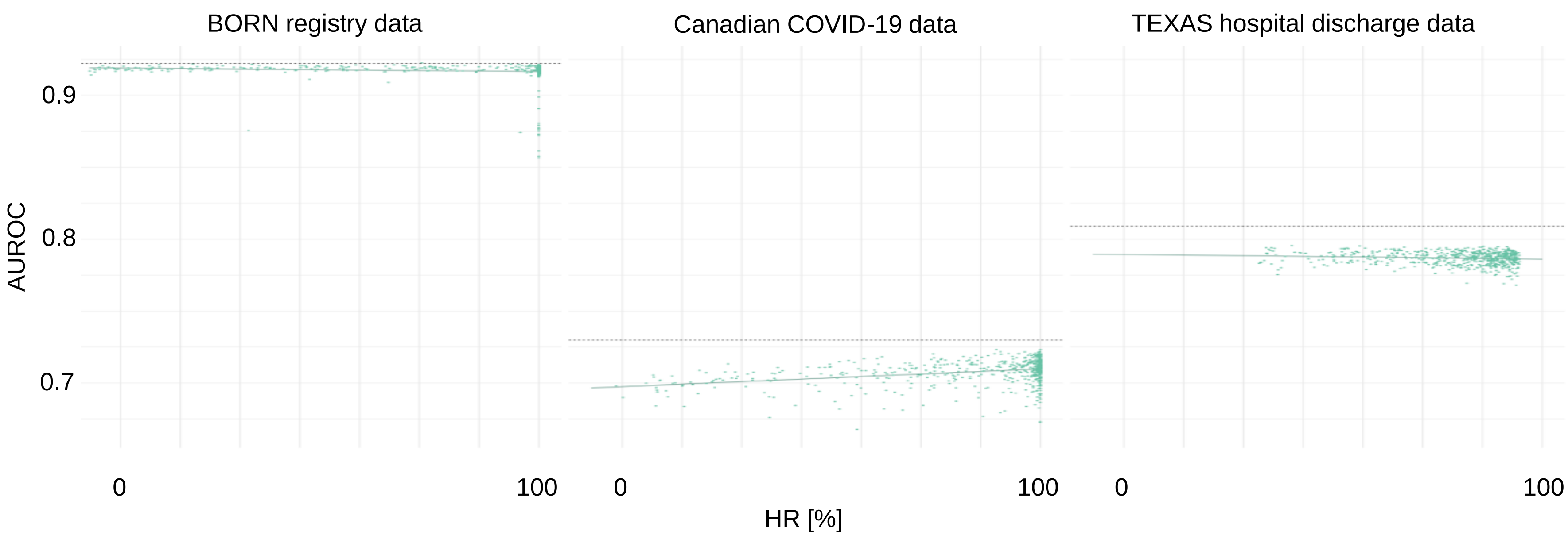


Figure 2. Exemplar mixed-effect model with the health dataset as random effect and HR as fixed effect and TSTR as outcome for the ST SDG model. 3 out of 12 health datasets are shown as examples. TRTR is indicated as dashed line.

CONCLUSIONS

The Impact of Hallucinations in Synthetic Health Data on Prognostic ML Models

1. What is the hallucination rate (HR) during tabular synthetic health data generation (SDG)?

   Hallucinated patients can make up 100% of the synthetic dataset if the health dataset is highly complex.

2. Does the magnitude of the HR in synthetic health data affect the performance of downstream prognostic ML models?

   Hallucinated patients do not necessarily impact prognostic ML model performance.

Does 1 or 2 vary across different SDG or prognostic ML models?

Limitations

- The results were from one SDG model, other SDG models may present with different HR and different impact on prognostic ML modeling.
- Most health care datasets were highly complex. Different complexity and correlational structures are very likely to impact the HR.
- Definition of hallucinated patients that are based on distribution shifts or correlational structures can produce very different results.
- The HR did not impact prognostic ML models but can still erode trust and may be a challenge for other downstream task (e.g. inference, propensity-score matching)

Scan this QR code to connect with me on LinkedIn

Thank you for listening !

And special thanks to …

… Samer El Kababji
… Dan Liu
… Khaled El Emam

Lisa Pilgram, MD
Postdoctoral Fellow at the Electronic Health Information Laboratory (Khaled El Emam)