THE UNIVERSITY OF ALABAMA AT BIRMINGHAM

SCHOOL OF MEDICINE

De-identified Delirium Data: Finding Delirium to Study Delirium

Co-PIs: Richard Kennedy, John Osborne

Background

- Delirium affects 30-40% of hospitalized older adults
  - Associated with increased risk of death, functional decline, and long-term cognitive impairment
- Development of NLP and ML algorithms to automate delirium identification is hampered by lack of suitable EHR data
- UAB Virtual Acute Care for Elders (ACE) quality improvement program has instituted delirium screening on all inpatient admissions over age 65
- Virtual ACE has determined delirium status on $\geqslant$ 33,000 patients over 6 years, providing a rich set of data for release to other investigators after de-identification

Aims and Hypotheses

- Aim 1: Deidentify an electronic health record (EHR) note corpus for studies of delirium
  - Hypothesis 1: Time to de-identify EHR notes for delirium will be decreased by machine de-identification
- Aim 2: Release a deidentified corpus of EHR notes and structured data for studies of delirium
  - Hypothesis 2: Our de-identified corpus will have greater than 90% power to detect differences between participants with and without delirium for analyses with commonly used ML algorithms

Delirium EHR Note Corpus

- Delirium screening with the Nursing Delirium Screening (NuDESC) scale performed once per nursing shift
  - 0 = no delirium
  - 1 = possible delirium
  - 2 = delirium
- Delirium status could change for participant as delirium develops / resolves during hospitalization
- Manual de-identification of 5,564 EHR text notes across 282 participants
  - 51.8% delirium, 16.2% possible delirium, 30.1% no delirium
  - 36.1% progress notes, 8.2% consult notes

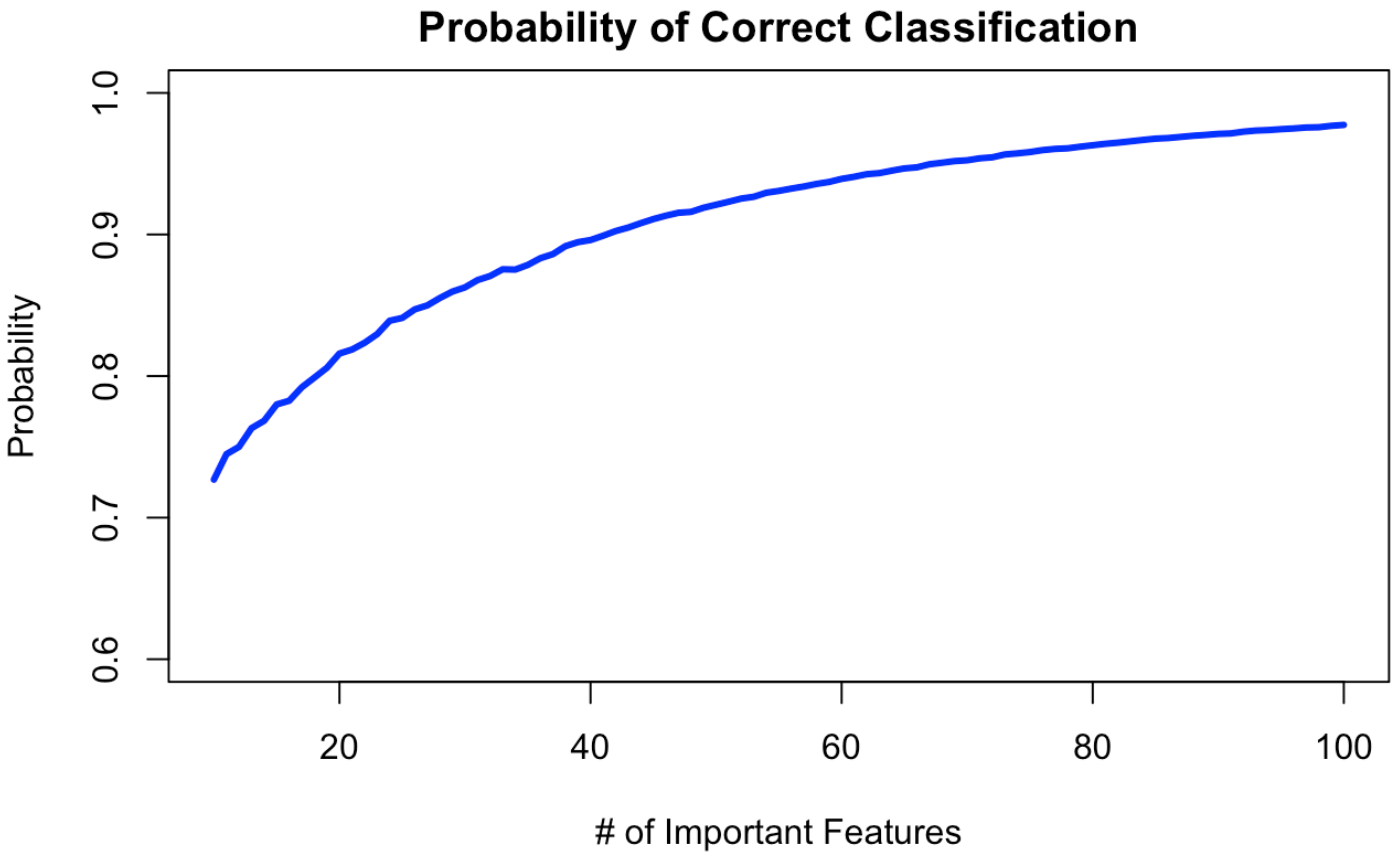
Delirium EHR Note Corpus

- Structured data added to EHR text note data for analysis
- Data set released to Physionet (https://physionet.org) for use by external investigators

Table 1: Data Set Composition

Domain	Measure(s)
Collected as part of Virtual ACE	
Demographics	Age, race, gender, education, marital status, residence
Delirium	Nursing Delirium Screening Scale[11]
Delirium	Confusion Assessment Method[50]
Global Cognition	Six-Item Screener[51]
ADLs	Katz Index[52]
Collected from EHR	
Clinical Text	Physician, nursing and consultant notes; radiology summaries; chief complaints
Laboratory Values	LOINC Codes
Medication	RXNorm Codes
Billing Codes	ICD-9-CM and ICD-10-CM
Procedures	CPT Codes
Admission Service	Medical / surgical classification

Predicted Classification Performance

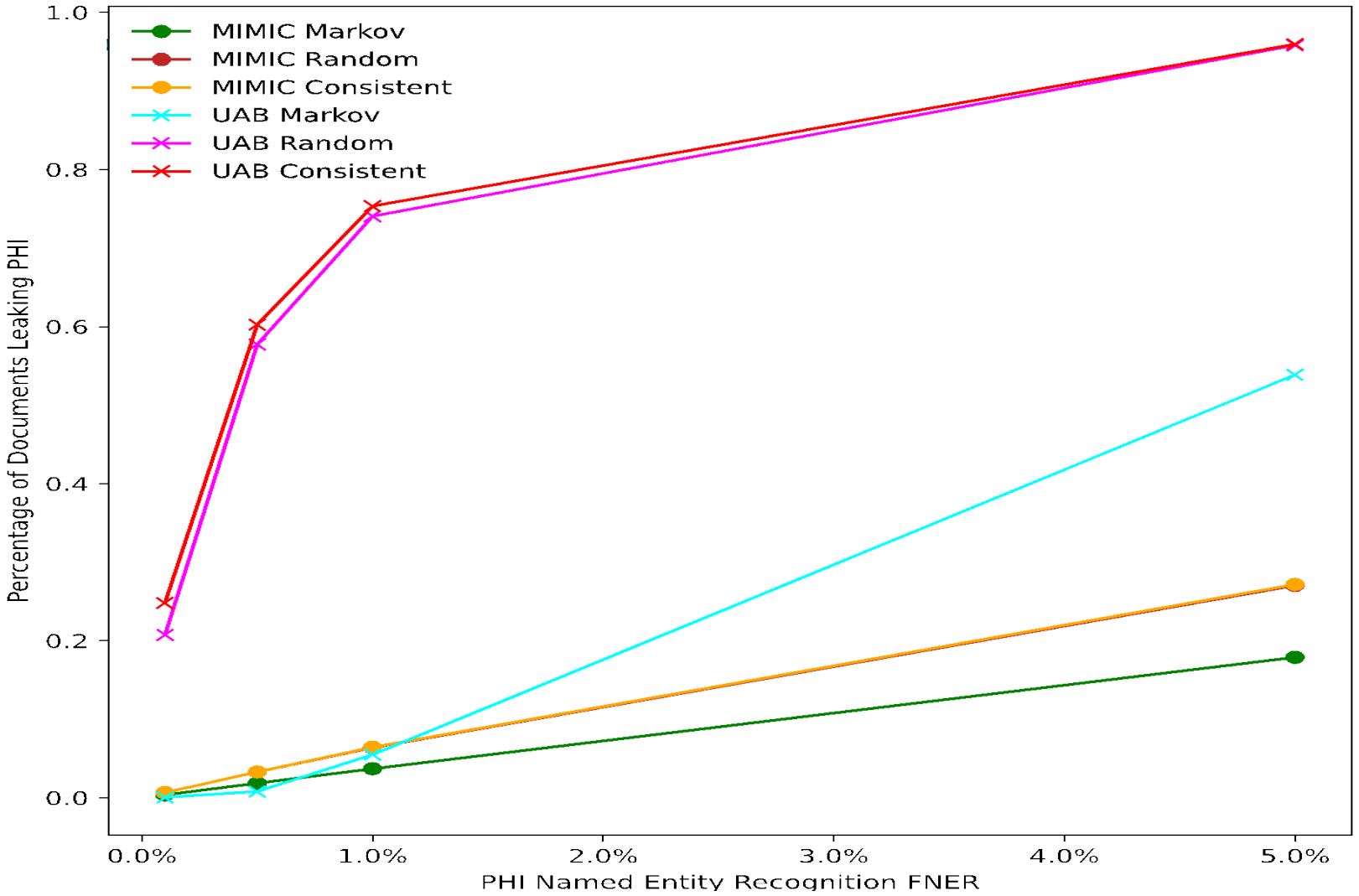


Calculations assume 10,000 features (words) present in EHR notes, and an effect size of 0.2 (small effect) for important features

Automated De-Identification

- Implemented 3 Hiding in Plain Sight (HIPS) strategies using BRAT annotation format to replace PHI
  - Consistent (same text used as replacement every time)
  - Random (different text used as replacement every time)
  - Markov (different text used as replacement 50% of time)
- Delirium corpus used to simulate data leakage for each replacement strategy with different false negative rates
  - UAB discharge note corpus and MIMIC dataset used for validation
- Software release: https://github.com/uabnlp/BRATsynthetic

Automated De-Identification



Summary

- We have successfully de-identified a corpus of EHR notes during hospitalization containing both episodes of delirium and episodes without delirium
  - This is the first widely available corpus of notes associated with delirium, which will facilitate development of NLP algorithms
- This corpus also contains structured data that can be used for modeling associations with delirium
- We have also developed and released software for automating de-identification of EHR notes using novel HIPS strategy to decrease probability of data leakage

Acknowledgments

- Tobias O'Leary
- Akhil Nadimpalli
- Ahana Chatterjee
- Mojisola Fasokun

- National Institutes of Health