DATA+AI SUMMIT 2022

Deidentifying 700mm Patient Notes

Lessons Learned from Our Journey at Providence

Lindsay Mico
Director Data Science, Providence

Nadaa Taiyab
Senior Data Scientist, Tegria

ORGANIZED BY databricks

Providence

Deidentification supports the Providence vision of "Health for a Better World"

25.6m
Total patient visits

1,085
Clinics

1
Health plan

$1.7b
Community benefit

52
Hospitals

1,700+
Published research studies

25k
Physicians

2.1m
Covered lives

120k
Caregivers

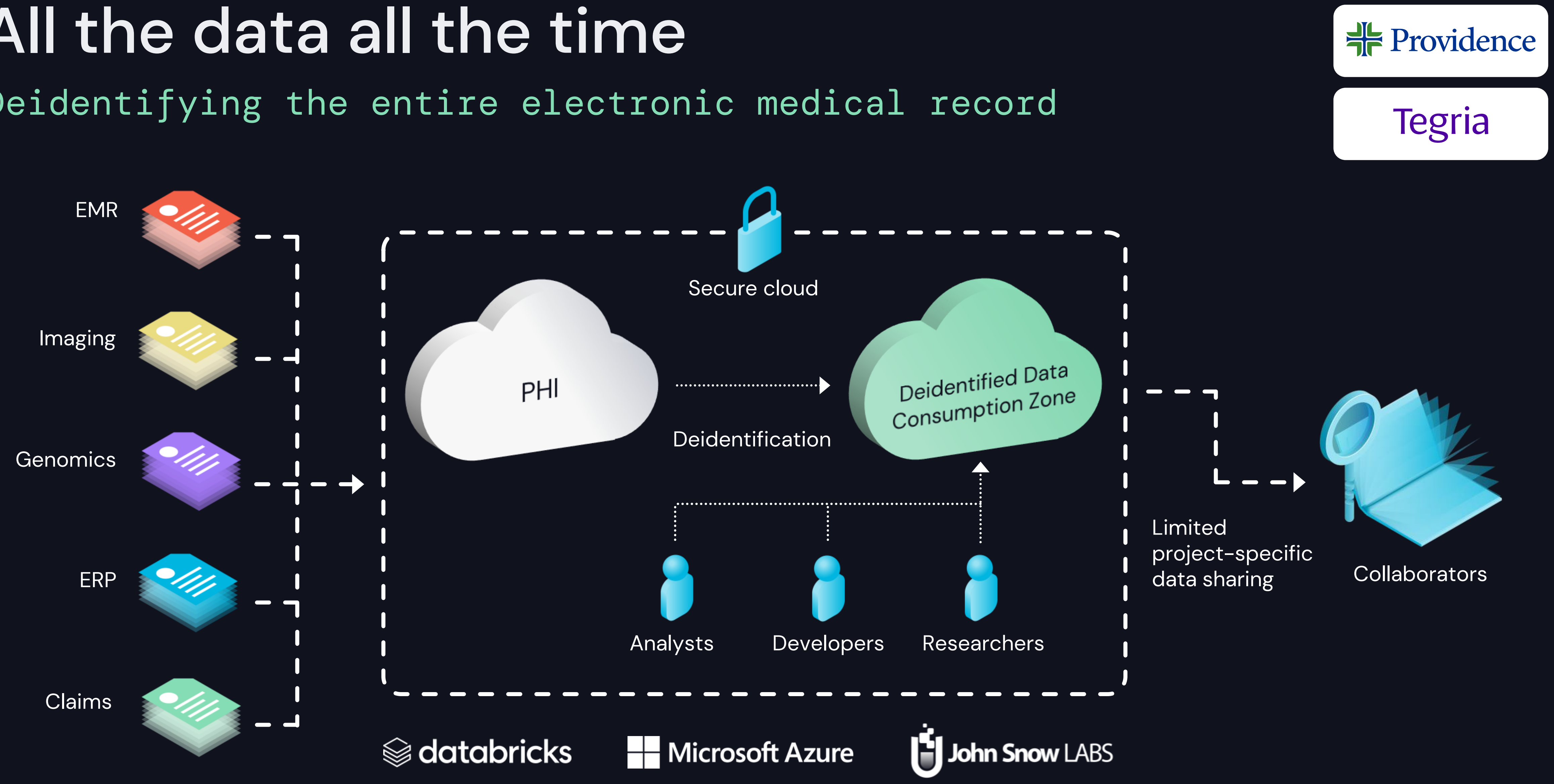
36k
Nurses

17
Supportive housing facilities

High school nursing schools and university

All the data all the time

Deidentifying the entire electronic medical record



Why are we doing this?

Deidentification at scale value proposition

Data Sharing

Respond to emerging threats such as pandemics

Medical Research

Make more and richer data available for medical research

Security

Part of Providence's layered approach to security

Tegria

What We Do

Optimize Care

- Care Journey
- Provider Experience
- Patient & Care Team Support

Integrate Technology

- Platform Modernization
- Interoperability
- Cybersecuity
- IT Efficiency

Transform Operations

- Revenue Cycle
- Automation
- Growth

Accelerate Revenue

- Analytics & Insights
- Patient Access Growth & Retention
- Learning & Adoption

How to Deidentify Patient Notes at Massive Scale

How does deidentification of notes work?

Tag PHI entities and then obfuscate them

Patient Note

MRN: 2349874
Account#: FN2340985

Maria Gonzalez is a 38 y.o. female with pregnancy complications. Admitted on 1/15/2022. Lives at 345 N 4th St, Seattle. She asked to be emailed at mgonzal3@hotmal.com.

How does deidentification of notes work?

Tag PHI entities and then obfuscate them

Patient Note

MRN: 2349874
Account#: FN2340985

Maria Gonzalez is a 38 y.o. female with pregnancy complications. Admitted on 1/15/2022. Lives at 345 N 4th St, Seattle. She asked to be emailed at mgonzal3@hotmal.com.

Step 1: Tag

MRN: <MEDICALRECORD>
Account#: <IDNUM>

<PATIENT> is a 38 y.o. female with pregnancy complications. Admitted on <DATE>. Lives at <STREET>, <CITY>. She asked to be emailed at <EMAIL>.

How does deidentification of notes work?

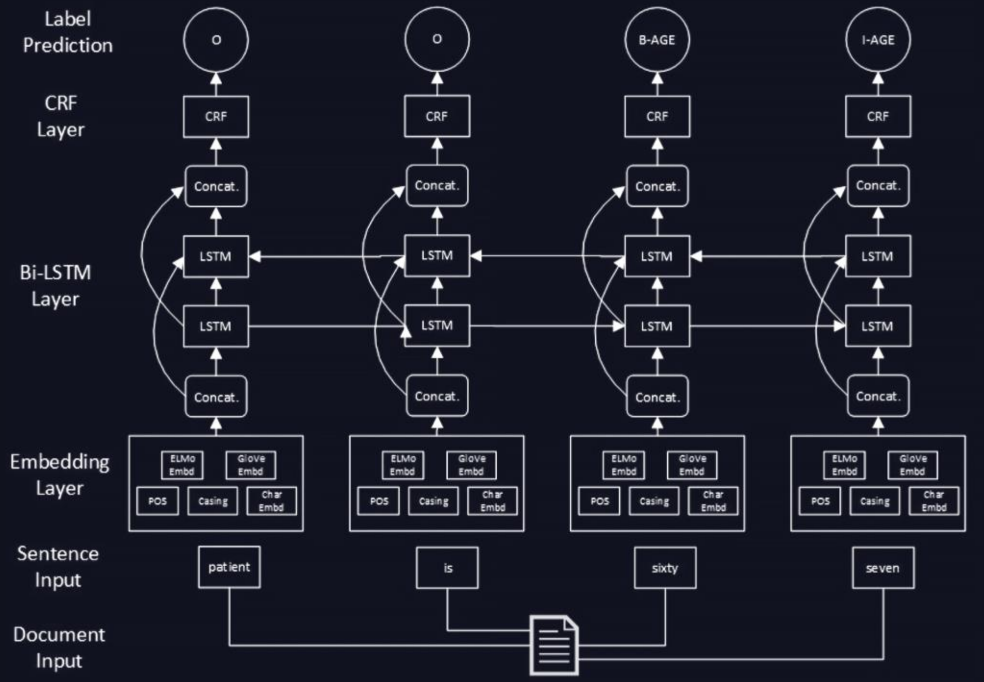
Tag PHI entities and then obfuscate them

Patient Note	Step 1: Tag	Step 2: Obfuscate
MRN: 2349874 Account#: FN2340985 Maria Gonzalez is a 38 y.o. female with pregnancy complications. Admitted on 1/15/2022. Lives at 345 N 4th St, Seattle. She asked to be emailed at mgonzal3@hotmal.com.	MRN: <MEDICALRECORD> Account#: <IDNUM> <PATIENT> is a 38 y.o. female with pregnancy complications. Admitted on <DATE>. Lives at <STREET>, <CITY>. She asked to be emailed at <EMAIL>.	MRN: 4369094 Account#: ZQ4895023 Zoe Kennedy is a 38-y.o. female with pregnancy complications. Admitted on 12/1/2021. Lives at E. Beverly St, Renton. She asked to be emailed at zemak9@snowmail.com.

Deep learning for the win

We are using Spark-nlp for Healthcare from John Snow Labs

NER Architecture



Char-CNN-BiLSTM

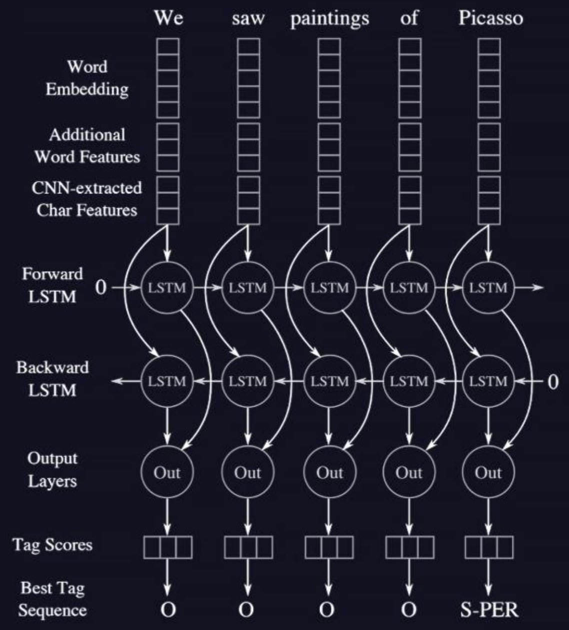


Figure 1: The (unrolled) BLSTM for tagging named entities. Multiple tables look up word-level feature vectors. The CNN (Figure 2) extracts a fixed length feature vector from character-level features. For each word, these vectors are concatenated and fed to the BLSTM network and then to the output layers (Figure

- Named entity recognition (NER) models tag patient identifiers.
- The models are “state of the art,” meaning they are the best available on the market.

Simple, scalable, easy-to-use code

GEEK OUT!

Patient note with PHI

Pre-processing

Tokenization

Embeddings

NER models

Regex

```
documentAssembler = DocumentAssembler()\
     .setInputCol("NOTE_TEXT")\
     .setOutputCol("document")

sentenceDetector = SentenceDetector()\
     .setInputCols(["document"])\
     .setOutputCol("sentence")

tokenizer = Tokenizer()\
      .setInputCols(["sentence"])\
      .setOutputCol("token")

embeddings = WordEmbeddingsModel.load("embeddings_clinical")\
      .setInputCols(["sentence", "token"])\
      .setOutputCol("embeddings")

ner1 = MedicalNerModel.load("ner_deid_subentity_augmented")\
       .setInputCols(["sentence", "token", "embeddings"]) \
       .setOutputCol("ner1")

ner_converter1 = NerConverter() \
       .setInputCols(["sentence", "token", "ner1"]) \
       .setOutputCol("entity1")

ner2 = MedicalNerModel.load('ner_deid_subentity_generic') \
       .setInputCols(["sentence", "token", "embeddings"]) \
       .setOutputCol("ner2")

ner_converter2 = NerConverter()\
       .setInputCols(["sentence", "token", "ner2"])\
       .setOutputCol("entity2")\

entityRuler = EntityRulerApproach() \
       .setInputCols(["document", "token"]) \
       .setOutputCol("entity3") \
       .setPatternsResource('./custom_regex') \
       .setEnablePatternRegex(True)
```

Merge tags

Obfuscation

Build ML pipeline

Run ML pipeline

```
chunk_merge = ChunkMergeApproach()\
       .setInputCols("entity1", "entity2", "entity3")\
       .setOutputCol("entity")\
       .setMergeOverlapping(True)

deid = DeIdentification()\
       .setInputCols(["sentence", "token", "entity"])\
       .setOutputCol("deid") \
       .setMode("obfuscate")

from pyspark.ml import Pipeline

nlpPipeline = Pipeline(stages=[
       documentAssembler,
       sentenceDetector,
       tokenizer,
       embeddings,
       ner1,
       ner_converter1,
       ner2,
       ner_converter2,
       entityRuler,
       chunk_merge,
       deid
       ])

empty_data = spark.createDataFrame([[""]]).toDF("NOTE_TEXT")
model = nlpPipeline.fit(empty_data)

result = model.transform(df_spark)
```

Deidentified patient note

Runs on Databricks

Built on Apache Spark and Spark ML

Cluster configuration

Databricks runtime version

10.5 ML (includes Apache Spark 3.2.1, Scala 2.12)

Worker and driver type

Standard_L32s_v2. 256 GB Memory, 32 Cores

Min-max workers

20 - 40

Pipeline throughput

Number notes

99,773

Number tokens

39,519,209

Run time

26.53 minutes

The hard part is meeting HIPAA

HIPAA privacy rule deidentification methods

Safe harbor
$164.514(b)(2)

Removal of 18 types of identifiers

No actual knowledge, residual information can identify individual

Expert determination
$164.514(b)(1)

Apply statistical or scientific principles

Very small risk that anticipated recipient could identify individual

Expert determination

Most clinical content is unstructured, which requires expert determination to deidentify

The law provides no quantitative criteria for certifying a dataset as having “very small risk” of reidentification.

How to evaluate reidentification risk?

1. Sampling

2. Annotation

3. Evaluation

Sampling

Statistical techniques for sound sampling

Drew a stratified random sample of 683 notes, stratified by:

- Gender
- Race
- Location

Used a power analysis to determine the sample size

Let's annotate some notes!

This is where the humans get involved to create a labeled dataset

Note samples	Model tags	Annotation
GONZALEZ,	PATIENT	PATIENT
maria	✘	**PATIENT ✓**
Admitted		O
May 1, 2021	DATE	DATE
Medical Record #:		O
437590234	MEDICALRECORD	MEDICALRECORD

The model missed "maria." I'll tag it as a patient name!

Measure what matters

Text	Actual	Predicted	Match
JONES, JENNIFER	PATIENT NAME	PATIENT NAME	✓
56789-4056	ZIP CODE	IDNUM	✓
940 Beverly Ave.	STREET		✗

For HIPAA, “what matters” is removing identifiers to protect the patient’s identity.

The zip code is incorrectly tagged as an ID. But since anything tagged will be removed, we will count it as a match!

Evaluation Metrics

% PHI Entities Found (Recall)

$$\frac{\textit{Entities Found}}{\textit{Total Entities}} = \frac{23}{25} = 92\%$$

% PHI Prevalence Pre-DeID

$$\frac{\textit{Notes with PHI Entities}}{\textit{Total Notes}} = \frac{8}{10} = 80\%$$

% PHI Prevalence Post-DeID

$$\frac{\textit{Notes with Missed Entities}}{\textit{Total Notes}} = \frac{2}{10} = 20\%$$

PHI Entities include patient name, phone/fax, email, street address, city, zip dates, id numbers, and organization names

Notes	PHI Entities
1	NAME, DATE
2	STREET, IDNUM, DATE
3	
4	NAME, NAME, DATE, DATE
5	PHONE, EMAIL, NAME, NAME, DATE, DATE
6	
7	STREET, DATE, DATE
8	MEDICALRECORD
9	DATE, STREET, CITY, EMAIL
10	NAME, NAME

PHI Entities*: 23 Found and 2 Missed

Top Level Results

(Two models + custom regex)

% PHI entities tagged

98%

% PHI prevalence in notes before

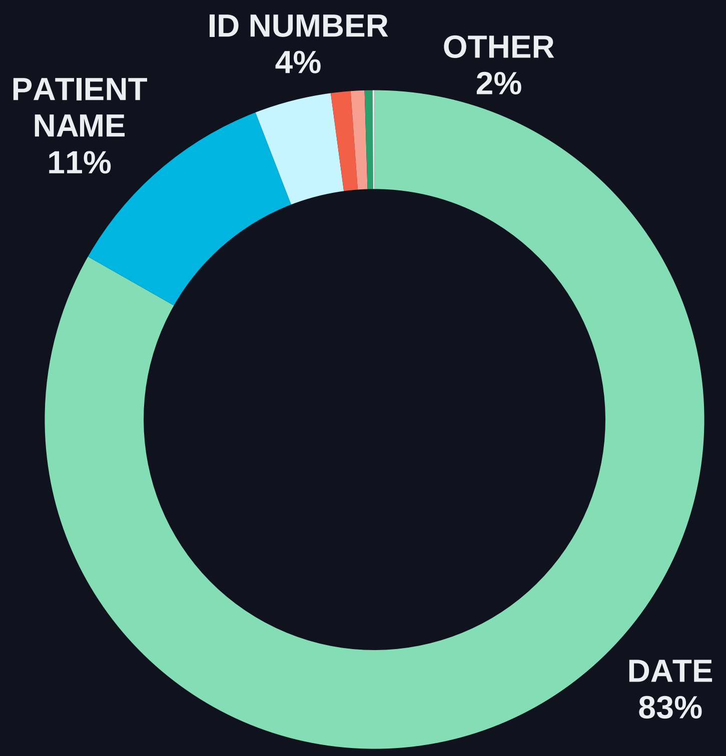
75%

% PHI prevalence in notes after

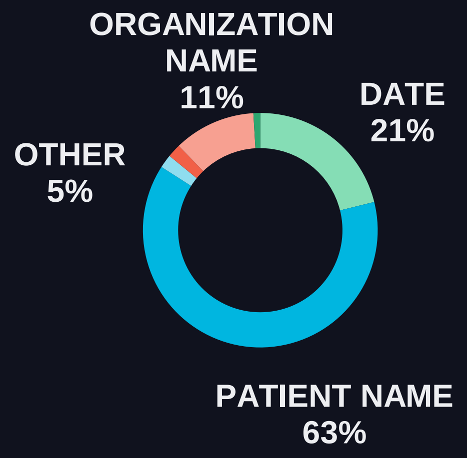
13%

Results by Entities

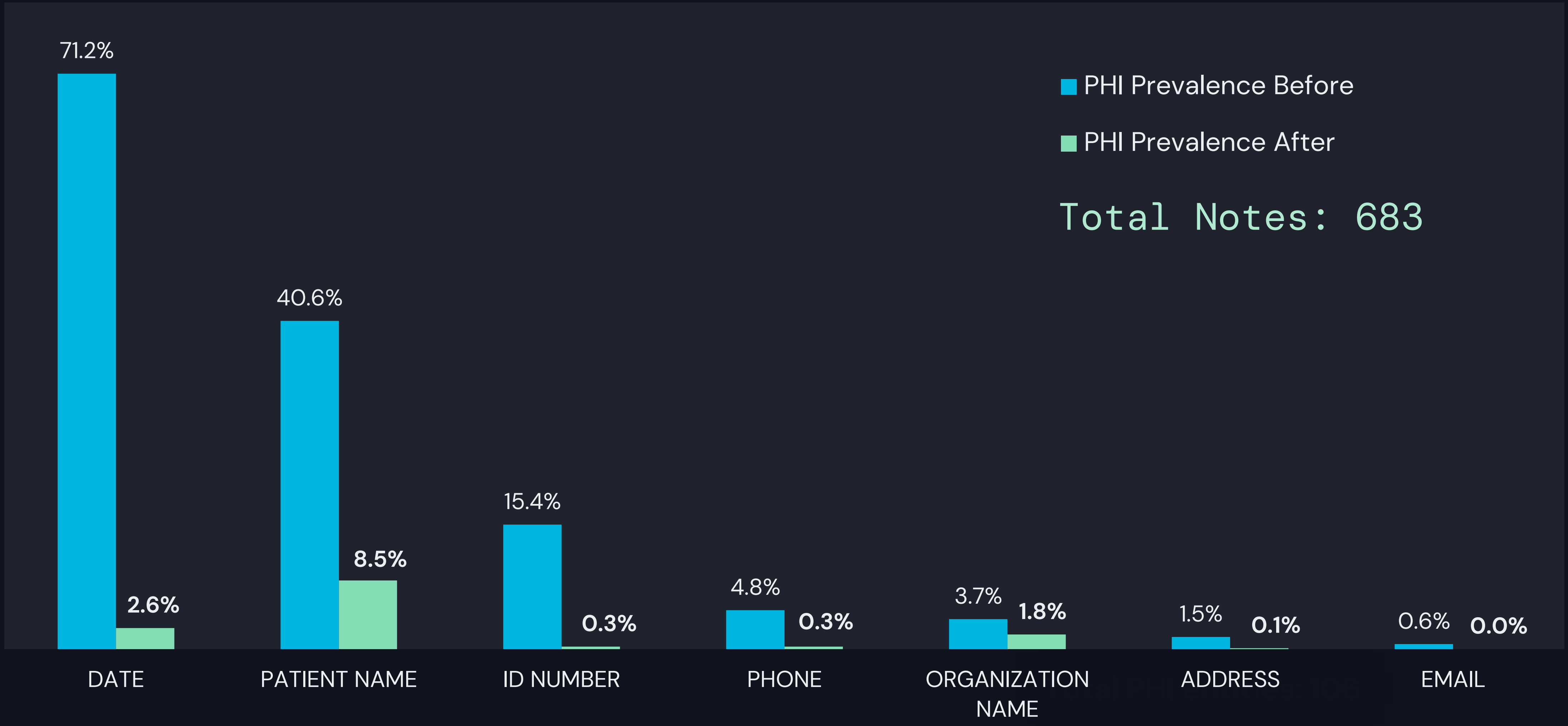
PHI entities before (4,537)



PHI entities after (106)



Results by PHI Prevalence in Notes



Equity Analysis

Effectiveness Metric

Prevalence After/ Prevalence Before

0 = No notes with PHI after Deid

1 = All notes still retain PHI after Deid

Effectiveness by Race

Race	Sample Size (notes)	All Entities	Patient Name
Caucasian	532	0.17	0.21
Asian	42	0.33	0.37
Other Minority	109	0.13	0.13

Underperforming on Asians. Need larger sample to test if this is significant.

Effectiveness by Sex

Sex	Sample Size (notes)	All PHI Entities	Patient Name
Male	281	0.15	0.18
Female	402	0.19	0.23

Need further analysis to understand if this difference is significant.

Conclusion

We can help you

Our vision is "health for a better world"

Deidentification at scale

- Secure cloud architecture
- Annotation resources
- Building the deidentification pipeline
- Expert determination

Contact us if you would like help deidentifying data at scale:

Lindsay Mico
Lindsay.Mico@providence.org

(Amar) Nadaa Taiyab
Amar.Taiyab@tegria.com

Appendix

Power Analysis

$$n = \frac{Z^2 \left(1 - \frac{\alpha}{2}\right) \times p(1-p)}{E^2} = \frac{1.96^2 \times 0.8(1-0.8)}{0.03^2} = \mathbf{683\ samples}$$

n → sample size

α → (1 - confidence level) = (1 - 0.95) = 0.05

Z → Z statistic for confidence level = 1.959964

p → expected proportion = 0.8

E → margin of error = 0.3

GEEK OUT!

Rules for “Partial” Matches

<table>
<tr><th>Annotated Data</th><th>Model Prediction</th><th>Match</th></tr>
<tr><td>Smith, Peter L.</td><td>Smith, Peter</td><td>✓</td></tr>
<tr><td>Jennings, Harry</td><td>Jennings</td><td>✗</td></tr>
<tr><td>435 N. 24th St, Suite 40</td><td>435 N. 24th St.</td><td>✓</td></tr>
<tr><td>2351 N. 50th St., Apt 5</td><td>Apt 5</td><td>✗</td></tr>
<tr><td rowspan="2">Zara Habib</td><td>Zara</td><td rowspan="2">✓</td></tr>
<tr><td>Habib</td></tr>
</table>

- Patient Names: Prediction cannot miss more 2 characters to count as a match
- Other Entities: Prediction matches at least 50% of the entity

DATA+AI SUMMIT 2022

The End

ORGANIZED BY databricks