Automated De-Identification, Consistent Obfuscation, and Regulatory Grade Validation of 2 Billion Patient Notes

Veysel Kocaman
John Snow Labs
Lindsay Mico
Providence Health
Mustafa Aytug Kaya
mkaya2@gmu.edu
George Mason University
Nadaa Taiyab
Providence Health
David Talby
John Snow Labs
Tae Surh
Providence Health
Yuqing Guo
Providence Health
Vivek Tomer
Providence Health
Robert Kramer
Providence Health

Article

Keywords:

Posted Date: September 5th, 2025

DOI: https://doi.org/10.21203/rs.3.rs-6867162/v1

Additional Declarations: No competing interests reported.

Automated De-Identification, Consistent Obfuscation, and Regulatory Grade Validation of 2 Billion Patient Notes

Veysel Kocaman[1,+], Lindsay Mico[2,+], Mustafa Aytug Kaya[3,*,+], Nadaa Taiyab[2,+], David Talby[1,+], Tae Surh[2,+], Yuqing Guo[2,+], Vivek Tomer[2,+], and Robert Kramer[2,+]

[1]John Snow Labs, city, postcode, country
[2]Researcher, Healthcare Intelligence, Providence Health, Renton, 98057, USA
[3]George Mason University, CSI, Fairfax, 22030, USA
[*]mkaya2@gmu.edu
[+]all authors contributed equally to this work

ABSTRACT

Rich Large, diverse collections of anonymous patient data—including text, numbers, and images— are essential to advancing a broad range of causes, from clinical decision support and real-world evidence to population health and hospital operations. This study presents a novel system used to automatically de-identify unstructured clinical text from 2 billion patient notes, using consistent obfuscation and tokenization to link them into a unified longitudinal dataset. To the best of our knowledge, this is the first such system to be externally certified for regulatory-grade accuracy on real-world data at this scale.

The system is based on proprietary medical language models and the modified Spark NLP - a distributed computing NLP framework for efficient execution on large clusters of commodity hardware. It satisfies the Expert Determination de-identification criteria under HIPAA (Health Insurance Portability and Accountability Act) Privacy Rules, establishing a baseline requirement of <5% PHI prevalence both in aggregate and per record. It achieves 99% Protected Health Information (PHI) obfuscation, and achieves 100% masking or shifting of target data fields. This level of accuracy surpasses even that of a triple manual review by 3 human annotators.

Obfuscation adds another layer of protection by rendering PHI elements indistinguishable from missed elements. Name changes, date shifting, and tokenizing identifiers are done consistently across documents about the same patient. Equity analysis was performed to ensure the system is not biased across demographic groups for gender, age, ethnicity, and state. Finally, an independent audit including adversarial testing on 790 randomly selected patients was performed, in which a dedicated “red team” working for 3 months was not able to re-identify any of the patients.

Introduction

Electronic Health Records (EHRs) have become ubiquitous in healthcare, with over 96% of acute care hospitals and 86% of office-based physicians in the USA utilizing them[1]. While structured (structured data: information organized in a fixed format, such as tables or fields, making it easy to search and analyze) data dominates billing and claims, a significant portion of clinical data exists as unstructured (unstructured data: information not organized in a fixed format, such as free-text notes or narrative documents) text, crucial for advancing population health, real-world evidence, patient safety, and drug discovery. However, due to its sensitive nature, unstructured clinical text is locked in secure databases, accessible to researchers only after extensive ethics review. This entails months of waiting before a research project can start, as well as limiting the size and scope of data that research has access to. Removing protected health information (PHI) can reduce the need for Institutional Review Boards (IRB) reviews, thus accelerating medical research.

The federally regulated HIPAA Privacy Rule provides two methods for de-identifying protected health information (PHI): the Expert Determination method and the Safe Harbor method. De-identified patient data is defined as health information that has been stripped of all “direct identifiers”—that is, any information that can be used to uniquely identify the patient. According to HIPAA’s Safe Harbor guidelines, there are 18 specific types of direct identifiers (as defined by the US Department of Health and Human Services), although any other data point that could uniquely identify a patient must also be considered.

Given the scale and diversity of data managed in our system as one of the largest U.S. healthcare providers - currently operating 52 hospitals and 1,085 clinics - it became an imperative to develop a methodology to de-identify PHI at scale. Key requirements were handling both structured and unstructured data, linking all data points about the same patient over time, and conforming to the Expert Determination method while retaining important medical and operational information as per

regulators' guidance.

Recent studies indicate that automated de-identification models based on deep learning (AI neural networks) can outperform human annotators (person who labels data for AI training) in identifying PHI, with hybrid methods showing the most promise[2]. Our system leverages artificial intelligence, machine learning, and cloud computing to achieve thorough and legally compliant de-identification of both structured and unstructured EHR data. Built on an automated de-identification software library by John Snow Labs (JSL)[3], this approach has successfully de-identified approximately 2 billion clinical notes and other documents. The medical language models which form the backbone of the system, offer regulatory-grade accuracy for de-identification of unstructured text backed by peer-reviewed academic research[4] and certification through independent expert determination across multiple organizations and countries[5].

Another requirement of the solution is consistent data obfuscation: the ability to replace PHI fields such as the patient's name, address, and other identifiers with medically appropriate random surrogates while meeting six data consistency requirements[3]. Obfuscation also has to be done consistently over time: When the patient's name is replaced by a random name, all dates are shifted by a random number of days, or identifiers are tokenized (replaced with unique codes), the same values must be used across all documents about the same patient over the years. This approach ensures that the de-identified dataset has practical applicability in real-world scenarios - and also reduces re-identification risk while preserving essential medical information.

Scaling capabilities are crucial when dealing with billion-scale datasets in healthcare, particularly for the simultaneous processing of historical and new data. The system had to process the entire historical dataset in a cost-efficient manner, so required the ability to distribute processing on large clusters of cheap commodity hardware. Cloud based processing was required to easily wind down clusters once the historical data processing was donw. Then, efficient processing of approximately 200,000 new notes added daily had to be done in a fast and compute-efficient manner, requiring a compute framework that can facilitate fast incremental processing.

The system was built based on the Apache Spark[6] and Spark NLP[7] software libraries, whose natively distributed architecture allows for parallel processing across multiple nodes, significantly reducing computation time. Spark's ability to scale horizontally optimizes resource utilization, supporting the required scalability while minimizing cost. The high rate of incoming data necessitates a system capable of real-time or near-real-time processing to maintain up-to-date de-identified datasets. Spark's in-memory processing and ability to handle streaming data make it well-suited for this task. Additionally, EHRs contain diverse data types, including structured data and unstructured text of varying lengths. Spark NLP's flexibility in handling different data formats and its extensions for medical language models allow for efficient processing of this heterogeneous data[8].

Equity is another material requirement from such a system. In July 2024, the US Department of Health and Human Services (HHS) issued a final rule regarding section 1557 of the Affordable Care Act (ACA). Section 1557 prohibits discrimination on the basis of race, color, national origin, sex, age, or disability in certain health programs and activities[9]. This requires validating that a de-identification system will not deliver substantially different accuracy for documents about patients from different demographics: for example, be less likely to detect Asian first and last names, hence putting such patients at a higher re-identification risk. While previous studies on bias in de-identification models used small synthetic datasets[10], this is the first study to evaluate equity in medical text de-identification on a large-scale, real-world dataset.

The primary contribution of this study is to develop a certified system to automatically de-identify over 2 billion patient notes. To the best of our knowledge, this is an order of magnitude larger than the scale of the largest certified de-identification system system published to date[11]. The next sections of this paper present the solution methodology, evaluation process, and the results - on accuracy, regulatory thresholds, equity, re-identification risk, and independent audit results.

Results

This section presents the results of the deidentification process applied to the annotated text fields. Table 1 summarizes the text fields by type (small or large text), along with their corresponding table and column names in both Redap and Clarity, and sample sizes. The analysis classified four fields as large text and eleven fields as small text, with a total of 26,419 documents for small text and 2,134 for large text. Small text is defined as having fewer than 2000 characters, while large text contains more than 2000 characters.

The PHI prevalence prior to deidentification varied significantly, ranging from near 0% to 95%, indicating the heterogeneity of information in these fields. For fourteen out of fifteen fields, the deidentification pipeline effectively removed PHI at rates between 89% and 99.9%. Following deidentification, the combined PHI prevalence across all entities was reduced to 5% or less for all fields. Notably, small text fields generally performed better than large text fields, as they contained fewer patient identifiers per entry. Ten out of eleven small text fields achieved a post-deidentification PHI prevalence of 1% or less.

The highest post-deidentification PHI prevalence was observed in narrative-full narrative (4.0%) and notes-full note text (3.7%), which can be attributed to the relative length of these large text fields and their functional purpose within Epic (a widely used electronic health record system) to summarize clinical encounters comprehensively. The recall rates (recall measures the percentage of sensitive information the system correctly identifies for removal) across all fields were impressive, ranging

Type	Table	Field	Sample Size
Small Text	SURGICAL_HX	PROC_COMMENTS	2340
	SURGICAL_HX	COMMENTS	2436
	OR_LOG_ALL_PROC	ALL_PROC_AS_ORDERED	2476
	ORDER_PROC	DESCRIPTION	2386
	ORDER_PROC	DIASPLAY_NAME	2365
	SURGICAL_HX	SURGICAL_HX_DATE	2357
	MEDICAL_HX	MEDICAL_HX_DATE	2398
	OR_CASE_ALL_PROC	ALL_PROC_AS_ORDERED	2450
	OR_CASE_PREOPDX	PRE_OP_DX	2427
	HSP_ADMIT_DIAG	ADMIT_DIAG_TEXT	2392
	IP_FLWSHT_MEAS	MEAS_VALUE	2392
Large Text	HNO_NOTE_TEXT	FULL_TEXT	1534
	ORDER_IMPRESSION	IMPRESSION	200
	ORDER_NARRATIVE	FULL_NARRATIVE	200
	ORDER_RES_COMMENT	RESULTS_CMT	200

Table 1. Text Fields Evaluated for Deidentification

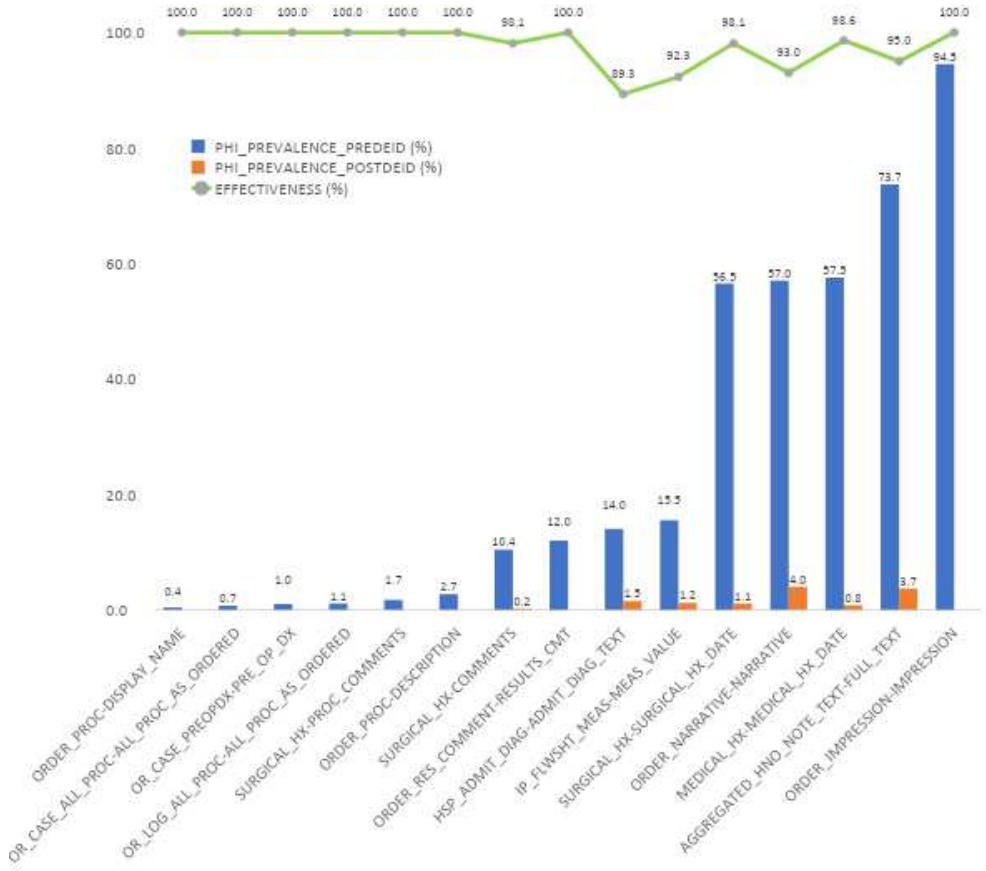

Figure 1. Summary Performance Metrics of Deidentified Text Fields (%)

from 90% to 100% for all PHI entities as seen in Table 2. The data presented in the "Missed Entities" and "Total PHI Entities" columns illustrate the number of PHI entities that were missed and the total number of PHI entities identified across all samples for each field. Please see Figure 1 for more details.

Recall rates for deidentification were notably high, ranging from 90% to 100% across all fields for all PHI entities, as detailed in Table 2. Precision (the percentage of items the system marked as sensitive that actually are sensitive information) for

Field	Recall [95% HDI]	Missed Entities	Total PHI Ents.	N	Precision %
ORDER_PROC-DISPLAY_NAME	100.0% [78.9, 100.0]	0	11	2365	10.4%
OR_CASE_ALL_PROC-ALL_PROC_AS_ORDERED	100.0% [85.5, 100.0]	0	17	2450	33.7%
ORDER_RES_COMMENT-RESULTS_CMT	100.0% [89.7, 100.0]	0	25	200	85.3%
SURGICAL_HX-PROC_COMMENTS	100.0% [93.7, 100.0]	0	42	2340	40.5%
ORDER_PROC-DESCRIPTION	100.0% [95.7, 100.0]	0	64	2386	30.7%
ORDER_IMPRESSION-IMPRESSION	100.0% [99.3, 100.0]	0	409	200	95.3%
HSP_ADMIT_DIAG-ADMIT_DIAG_TEXT	90.9% [88.1, 93.3]	39	429	2463	59.3%
IP_FLWSHT_MEAS-MEAS_VALUE	94.3% [92.2, 96.0]	30	528	2392	77.0%
OR_CASE_PREOPDX-PRE_OP_DX	96.2% [84.2, 99.8]	1	26	2427	11.6%
OR_LOG_ALL_PROC-ALL_PROC_AS_ORDERED	97.0% [87.3, 99.9]	1	33	2476	32.8%
ORDER_NARRATIVE-NARRATIVE	98.3% [97.3, 99.1]	12	719	200	88.0%
SURGICAL_HX-COMMENTS	98.4% [96.7, 99.4]	5	317	2436	92.8%
SURGICAL_HX-SURGICAL_HX_DATE	98.4% [97.8, 99.0]	26	1672	2357	80.2%
MEDICAL_HX-MEDICAL_HX_DATE	99.0% [98.4, 99.4]	18	1718	2398	88.7%
AGGREGATED_HNO_NOTE_TEXT-FULL_TEXT	99.1% [98.9, 99.2]	94	10236	1534	85.9%

Table 2. Recall and Precision with 95% High Density Interval at Field Level

large text fields ranged from the mid-80s to low 90s, which falls within the acceptable range of 80% or more, as considered by experts[12]. There were a total of 16246 PHI entities across the different source fields in the evaluation corpus and from them 16020 where detected and obfuscated resulting in an overall recall of 98.6% for the NLP pipeline. Note that the small text fields with low precision numbers generally also have a very small number of total predicted entities, suggesting that the low precision may in fact be at least partly due to sampling error. Therefore, the total loss of information in the data is very small, even if the precision is low.

Metrics by Entity Type

The effectiveness of the deidentification pipeline in mitigating reidentification risk is primarily assessed through the PHI Prevalence Post-DeID metric. Analysis reveals that the mean and upper bound of the 95% High Density Interval -a statistical range that shows where the most likely values fall- for this metric across all PHI entities and fields falls below 5%, demonstrating the pipeline's robust performance in eliminating personally identifiable information.

Dates, patient names, and ID numbers emerge as the most frequently occurring entities in the text data. However, high-risk identifiers such as email addresses, phone numbers, and street addresses are scarce and virtually eradicated post-deidentification. Similarly, zip codes, city names, and device serial numbers exhibit minimal presence after processing. Although ID numbers are prevalent, they predominantly consist of internal identifiers, which present limited risk without access to secure systems. Notably, external high-risk identifiers like social security numbers were absent from the annotated samples. The pipeline effectively managed the abundant date information, reducing its prevalence to below 4% across all fields. Patient names, detected in 12 fields, were successfully removed by the pipeline. The model demonstrated high accuracy, rarely missing multiple entity types within a single record. Only three such instances were observed among over 30,000 annotated records. Specifically, two records showed concurrent misses of both PATIENT name and DATE entities, while one record had missed detections for both IDNUM and CITY entities . These results indicate that our deidentification pipeline significantly reduces reidentification risk across diverse text fields and entity types, aligning with or surpassing conservative industry standard risk thresholds. The pipeline's performance underscores its potential as a robust tool for safeguarding patient privacy in healthcare data management. For a detailed breakdown of these rare multi-entity misses, refer to Table 3 .

Field	Missed Ents.	>1 Missed
AGGREGATED_HNO_NOTE_TEXT	56	2
HSP_ADMIT_DIAG	38	1
IP_FLWSHT_MEAS	28	0
MEDICAL_HX	18	0
ORDER_NARRATIVE	8	0
OR_CASE_PREOPDX	1	0
OR_LOG_ALL_PROC	1	0
SURGICAL_HX (Comments)	4	0
SURGICAL_HX (Date)	25	0

Table 3. Records with Missed Entities per Field

Independent Audit and Risk Identification

In determining risk identification, we faced challenges due to the lack of clear quantitative benchmarks for reidentification risk in expert determination. To address this, we adopted a conservative approach by establishing baseline requirements that included a threshold of less than 5% PHI prevalence in aggregate and per PHI element per record. This threshold was applied to both direct and indirect identifiers as described above.

An Independent Audit (IA) team conducted a comprehensive audit of our deidentification methodology, focusing on the machine learning model and assessing reidentification risk. The audit process included adversarial testing on 790 randomly selected patient records, a review of the machine learning deidentification process, and an analysis of deidentification results to assess data breach risk. Additionally, external sources such as social media and public records were utilized to attempt patient reidentification. The audit concluded with no successful patient reidentifications.

The results demonstrated that 100% of the target structured data had been masked or shifted, resulting in near-total anonymization. The methodology correctly identified an estimated 99% of all PHI elements in text fields, achieving approximately 0% PHI Prevalence Post-DeID for nine out of fifteen text fields. Furthermore, the prevalence was less than 5% per record for any entity type. The approach involved obfuscating rather than merely masking identified PHI elements, rendering them nearly indistinguishable from missed elements while allowing for independent identification of individual PHI elements. This behavior was consistent across key demographics.

The obfuscation technique provides an additional layer of protection compared to traditional methods, making real and synthetic PHI nearly indistinguishable. Although there is a theoretical possibility that future technologies could enhance reidentification capabilities, it is deemed unlikely due to the imperfections inherent in even industry-leading AI/ML tools in accurately identifying all named entities. This limitation presents challenges for automating any subsequent reidentification algorithms at scale. Moreover, we are committed to continually improving our deidentification approach as new technologies emerge.

Pre and Post Deidentification PHI prevalence

Figure 1 provides a comprehensive overview of the Pre-Deidentification (Pre-DeID) and Post-Deidentification (Post-DeID) PHI prevalence across various fields. The Pre-Deidentification (Pre-DeID) PHI prevalence varied significantly, ranging from nearly 0% to 95%, highlighting the heterogeneity of information across different fields. For fourteen out of fifteen fields, the deidentification pipeline effectively removed PHI in 89% to 99.9% of cases. Post-Deidentification (Post-DeID) PHI prevalence across all fields was 5% or less. Small text fields generally performed better than large text fields due to containing fewer patient identifiers per entry. Specifically, ten out of eleven small text fields had a Post-DeID PHI prevalence of 1% or less. The results for recall are equally impressive, ranging from 90%-100% across all fields for all PHI entities in Table 15). The "Missed Entities" and "Total PHI Entities" columns show the number of PHI entities (dates, names, etc.) that were missed in total and the number of total PHI entities found across all the samples for a given field.

The highest Post-DeID PHI prevalence was observed in the "narrative-full_narrative" (4.0%) and "notes-full_note_text" (3.7%) fields, which, while higher relative to other fields, remained low in absolute terms. This higher prevalence reflects the extensive length and comprehensive nature of these large text fields, which are designed to encapsulate the full context of clinical encounters.

The Post-DeID PHI prevalence on an entity basis is a crucial metric for evaluating reidentification risk. After applying the deidentification pipeline, the mean and the upper bound of the 95% High Density Interval (HDI) for Post-DeID PHI prevalence of all entities, including direct and indirect PHI identifiers across all fields, were below 5%, as detailed in Table 4. This is below the threshold generally accepted for direct PHI identifiers to be legally acceptable[12]. As described in table 3 it was extremely

rare for the model to miss multiple entity types in a given record. There were only three records across over 36180 annotated records where the model missed more than one entity type.

Field	Records with Missed Entity	2+ Missed	N
ORDER_NARR	8 (4.00%)	0 (0.00%)	200
AGG_HNO_NOTE	56 (3.65%)	2 (0.13%)	1534
HSP_ADMIT_DIAG	38 (1.54%)	1 (0.04%)	2463
IP_FLWSHIT_MEAS	28 (1.17%)	0 (0.00%)	2392
SURG_HX_DATE	25 (1.06%)	0 (0.00%)	2357
MED_HX_DATE	18 (0.75%)	0 (0.00%)	2398
SURG_HX_COMM	4 (0.16%)	0 (0.00%)	2436
OR_LOG_ALL_PROC	1 (0.04%)	0 (0.00%)	2476
OR_CASE_PREOPDX	1 (0.04%)	0 (0.00%)	2427

Table 4. Proportion of fields with one or more PHI present after de-identification.

Lastly, an independent audit performed by an external consultancy firm was not able to reidentify any personal information on 790 randomly selected data records.

Equity analysis of deidentification performance

The Effectiveness and PHI Prevalence Post-DeID metrics and 95% HDI (utilizing the non-informative prior) for patient notes was calculated for the subgroups from each category: sex, state, and race.

Group	Subgroup	Sample Size	Effectiveness [95% CI]	PHI Prevalence Post Deid [95% CI]
SEX	Male	631	95.6% [93.7, 97.2]	3.2% [2.0, 4.7]
	Female	903	94.6% [92.8, 96.1]	4.0% [2.8, 5.3]
STATE	CA	346	95.9% [93.1, 97.7]	3.2% [1.7, 5.3]
	WA	730	94.5% [92.5, 96.1]	4.4% [3.1, 6.0]
	OR	335	95.0% [91.6, 97.4]	3.0% [1.6, 5.1]
	Low Presence State	123	96.4% [90.9, 99.0]	2.4% [0.6, 6.1]
RACE	Caucasian	1197	94.7% [93.1, 96.0]	3.8% [2.9, 5.0]
	Asian	94	94.8% [88.8, 98.3]	4.3% [1.4, 9.4]
	Other Minority	243	96.8% [93.9, 98.7]	2.5% [1.0, 4.8]

Table 5. Effectiveness and PHI Prevalence Post-Deidentification by Demographic Groups

Effectiveness is a good metric for comparing results across groups because it is independent of the PHI Prevalence Pre-DeID. However, the "bottom-line" metric that best approximates risk of reidentification is PHI Prevalence Post-DeID. The HDI of the subgroups within each category overlap for both metrics when calculated across all subgroups. This suggests that there is no substantial difference between the performance of the model and the bottom-line metrics between subgroups as depicted in Table 5. Note that the HDI for Effectiveness and PHI-Prevalence Post-DeID also overlap when calculated specifically for patient names. Given these results, it is our best professional judgement that the model performs comparably across subgroups and is therefore equitable in terms of ability to identify patient identifiers as seen in figures 2 and 3.

Tables 6 and 6 present a comprehensive analysis of the deidentification pipeline's effectiveness across various entity types and text fields. Tables demonstrate significant reduction in identifiable information after the pipeline's application.

Discussion

This solution and methodology underwent rigorous internal evaluation, followed by an external assessment conducted by an independent auditor. The combined multi-year-long effort demonstrates that this solution produces a dataset with minimal re-identification risk while preserving crucial medical information essential for enhancing patient care and outcomes.

Beyond showing that automated de-identification of unstructured real-world clinical data at regulatory-grade accuracy is viable, this solution is also the first certified system to do at a billion-document scale. The solution addresses the dual mandate of protecting patient privacy and facilitating efforts to accelerate medical research. Unlocking the entire history of unstructured medical notes across a large healthcare system, while integrating the structured and unstructured data modalities consistently

Table	Field	City		Date		Dev		E-mail		Id	
		Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post
AGG_HNO_NOTE_TEXT	FULL_TEXT	2.0	0.1	69.8	1.4	0.3	0.0	0.4	0.0	16.0	0.1
ORDER_IMPRESSION	IMPRESSION	0.0	-	94.5	0.0	0.0	-	0.0	-	26.0	0.0
ORDER_NARRATIVE	NARRATIVE	0.0	-	55.0	0.5	0.0	-	0.0	-	23.0	1.5
ORDER_RES_COMMENT	RESULTS_CMT	0.0	-	7.5	0.0	0.0	-	0.0	-	5.0	0.0
HSP_ADMIT_DIAG	ADMIT_DIAG_TEXT	0.0	-	11.0	1.4	0.0	-	0.0	-	3.0	0.0
IP_FLWSHT_MEAS	MEAS_VALUE	0.2	0.0	6.6	0.1	0.0	-	0.0	-	1.2	0.1
MEDICAL_HX	MEDICAL_HX_DATE	0.0	-	57.3	0.8	0.0	-	0.0	-	0.2	0.0
ORDER_PROC	DESCRIPTION	0.0	-	2.7	0.0	0.0	-	0.0	-	0.0	-
ORDER_PROC	DISPLAY_NAME	0.0	-	0.2	0.0	0.0	-	0.0	-	0.1	0.0
OR_CASE_ALL_PROC	ALL_PROC_AS_ORDERED	0.0	-	0.6	0.0	0.0	-	0.0	-	0.0	-
OR_CASE_PREOPDX	PRE_OP_DX	0.0	-	0.7	0.0	0.0	-	0.0	-	0.2	0.0
OR_LOG_ALL_PROC	ALL_PROC_AS_ORDERED	0.0	-	1.1	0.0	0.1	0.0	0.0	-	0.0	-
SURGICAL_HX	COMMENTS	0.0	-	10.2	0.2	0.2	0.0	0.0	-	0.0	-
SURGICAL_HX	PROC_COMMENTS	0.0	-	1.7	0.0	0.0	-	0.0	-	0.0	-
SURGICAL_HX	SURGICAL_HX_DATE	0.0	-	56.4	1.1	0.0	-	0.0	-	0.1	0.0

Table 6. PHI Entity Prevalence by Field (Pre and Post De-identification) - Part 1

Table	Field	Pat		Phone		Str		Zip	
		Pre	Post	Pre	Post	Pre	Post	Pre	Post
AGG_HNO_NOTE_TEXT	FULL_TEXT	39.7	2.2	4.5	0.0	0.7	0.0	0.6	0.0
ORDER_IMPRESSION	IMPRESSION	0.5	0.0	0.0	-	0.0	-	0.0	-
ORDER_NARRATIVE	NARRATIVE	23.0	2.0	0.0	-	0.0	-	0.0	-
ORDER_RES_COMMENT	RESULTS_CMT	0.0	-	0.0	-	0.0	-	0.0	-
HSP_ADMIT_DIAG	ADMIT_DIAG_TEXT	0.2	0.1	1.4	0.1	0.0	-	0.0	-
IP_FLWSHT_MEAS	MEAS_VALUE	6.4	1.0	2.6	0.0	0.2	0.0	0.1	0.0
MEDICAL_HX	MEDICAL_HX_DATE	0.1	0.0	0.0	-	0.0	-	0.0	-
ORDER_PROC	DESCRIPTION	0.0	-	0.0	-	0.0	-	0.0	-
ORDER_PROC	DISPLAY_NAME	0.1	0.0	0.0	-	0.0	-	0.0	-
OR_CASE_ALL_PROC	ALL_PROC_AS_ORDERED	0.0	-	0.1	0.0	0.0	-	0.0	-
OR_CASE_PREOPDX	PRE_OP_DX	0.1	0.0	0.0	-	0.0	-	0.0	-
OR_LOG_ALL_PROC	ALL_PROC_AS_ORDERED	0.1	0.0	0.0	-	0.0	-	0.0	-
SURGICAL_HX	COMMENTS	0.0	-	0.0	-	0.0	-	0.0	-
SURGICAL_HX	PROC_COMMENTS	0.0	-	0.0	-	0.0	-	0.0	-
SURGICAL_HX	SURGICAL_HX_DATE	0.0	-	0.0	-	0.0	-	0.0	-

Table 7. PHI Entity Prevalence by Field (Pre and Post De-identification) - Part 2

across a longitudinal patient journey, removes IRB reviews as a bottleneck for many research efforts and enables novel research on larger and richer patient datasets.

The resulting dataset, along with the running system that updates it daily with fresh patient data, enables healthcare providers and researchers to gain valuable insights that can drive improvements in patient care, without compromising individual privacy. It aligns with the current regulations and ethical standards of medical research and practice, ensuring that patient data is utilized responsibly to benefit both individuals and the broader healthcare community.

Methods

We developed a comprehensive de-identification methodology for diverse healthcare data types, both for structured data and unstructured free text. The primary objectives of our deidentification algorithms are to protect patient privacy by utilizing state-of-the-art technology to mask or obfuscate identifiers with high accuracy, minimize information loss by retaining non-identifying information while maintaining referential integrity across tables, preserving temporal relationships, and keeping the original data schema intact. Additionally, our methodology facilitates large-scale de-identification by enabling the processing of historical data and daily updates, with flexibility for future data field inclusion and ongoing maintenance.

Our framework employs a selective approach, choosing the most suitable algorithm based on the specific data type and the

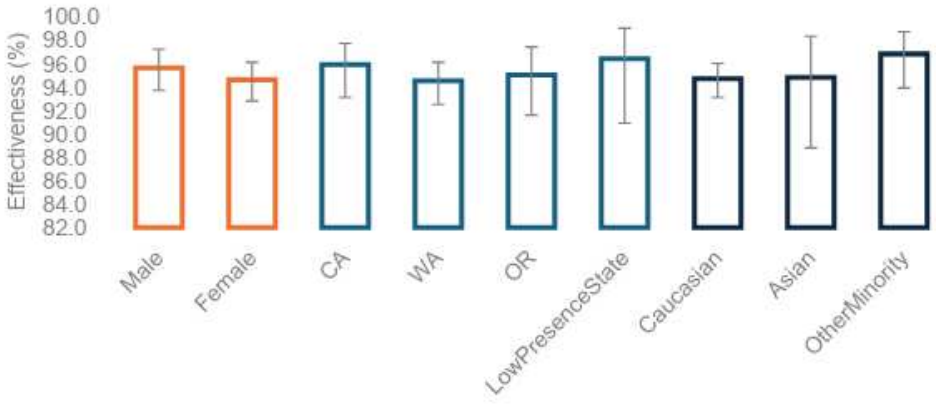

Figure 2. Effectiveness of Deidentification on All PHI Entities by subgroups with 95% High Density Intervals.

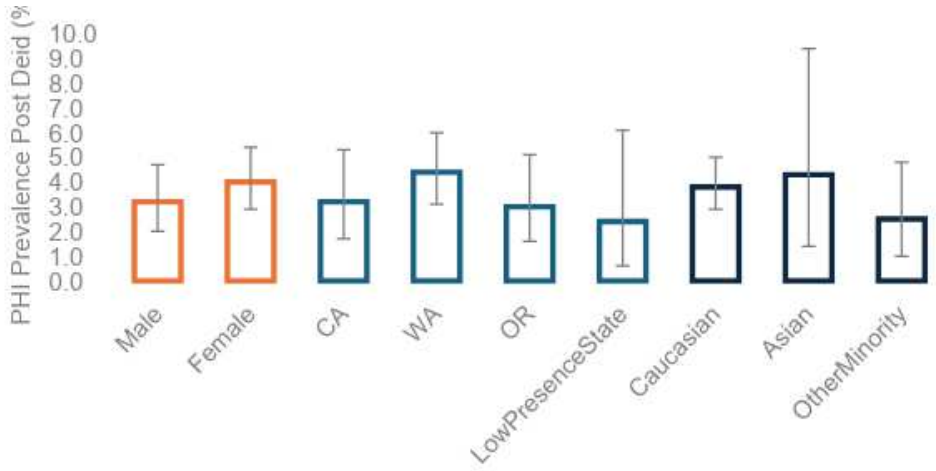

Figure 3. Post-Deidentification PHI Prevalence for all PHI Entities by Subgroup with 95% High Density Intervals.

presence or absence of patient identifiers within the dataset. The process of determining data types and identifying patient identifiers involves a combination of technical approaches, such as analyzing database schemas, complemented by manual review procedures. This comprehensive methodology aims to ensure thorough and accurate de-identification across varied data landscapes. A detailed exposition of the objectives and methodological approach for de-identification follows in subsequent sections.

In order to determine the appropriate de-identification algorithm for each data field, *column categorization* is utilized. This process evaluates both the data type (e.g., dates, numerical, categorical, string) and content to ensure effective PHI protection and use of computational resources. Fields containing explicit patient identifiers, such as addresses or social security numbers, are readily categorized as PHI and subsequently hashed (hashed: converted into a unique, unreadable string of characters using a one-way mathematical function) or masked. However, ambiguous columns require expert judgment to identify potential indirect identifiers. For instance, payor names may inadvertently reveal patient information through employer names or location-specific tribal affiliations. Additionally, text fields present unique challenges based on their length. Small text fields (less than 2,000 characters) often contain duplicated values or preset categories mixed with free text, while large text fields (over 2,000 characters), such as patient notes, are typically more heterogeneous and resource-intensive for annotation and pipeline development, and each warrant employment of distinct techniques. The deidentification process utilizes a comprehensive policy table that specifies algorithms for each field based on content and data type, with default policies applied to unevaluated fields, encompassing 141 Epic-Clarity tables and 926 explicitly defined column policies as depicted in Table 8, thus ensuring thorough protection of patient data across core EMR concepts.

De-Id Policy	Column Count	%
Keep	602	65.0
Entity Hash	173	18.7
Date Shift	100	10.8
Mask	29	3.1
NLP Pipeline	15	1.6
Zip Code Truncation	4	0.4
Date Year Truncation	3	0.3
TOTAL	926	100.0

Table 8. De-Id Policy Distribution

Natural Language Processing (NLP) for PHI Detection

Deep learning models in **NLP**, specifically Named Entity Recognition (NER) models, form the core of our de-identification process. These sophisticated models effectively *detect* and *label* patient identifiers within text data. Once identified, the flagged information undergoes *obfuscation* or *masking* to ensure privacy. The following section provides a detailed explanation of the technology and processes employed for de-identifying text data. De-identification of text data poses unique challenges due to the variety of identifiers and formats. While rules-based methods like text matching. template detection and regular expressions have limitations, deep learning models have significantly improved identifier extraction accuracy.

The de-identification process employs a combination of customized rules and state-of-the-art models from JSL. This decision was influenced by the consistent superior performance of JSL's pre-trained models over other commercial entity extraction solutions, due to their proprietary DL models trained on an in-house curated datasets, a unique healthcare-specific embeddings trained on a large corpus of clinical documents and the use of current model architectures as well as hand-crafted business rules codified into scalable codebase through contextual parsers[3].

Notably, these models have shown the ability to outperform human annotators in specific de-identification scenarios. According to a study on n2b2 clinical text standard de-identification benchmark, it took two independent clinician reviewers to achieve a macro F1 (harmonic mean of precision and recall) score of 95% in identifying protected health information (PHI)[13]. In contrast, John Snow Labs' base models have demonstrated performance surpassing that of two human annotators with an macro F1 score of 96.1%[3]. Building on this foundation, our custom text de-identification pipeline, developed using JSL's software library and tools, achieved a recall of 99% on patient notes, exceeding the performance of three combined independent clinical reviewers.

This superior performance of automated systems in NER tasks for de-identification represents a significant advancement in protecting patient privacy while maintaining data utility for research. The ability of JSL's base models to outperform two human annotators, and our custom pipeline's capacity to surpass three annotators, not only improves efficiency but also enhances the overall accuracy and reliability of the de-identification process. As we continue to refine our approach, these advancements pave the way for more robust and scalable solutions in handling sensitive medical data.

Base NLP Pipeline

John Snow Labs' pre-trained de-identification pipeline consists of five stages: text pre-processing and feature generation, named entity recognition (NER), contextual rules, chunk merging, and obfuscation. This comprehensive approach aims to safeguard sensitive information without sacrificing data integrity, ensuring the possibility of future re-identification if required.

The text pre-processing stage begins with a document assembler that generates initial annotations from raw text. Subsequently, sentence detection is performed using a deep learning model specifically optimized for clinical and multilingual text, as conventional rule-based methods often prove inadequate for handling non-standard punctuation prevalent in medical notes. Tokenization follows, segmenting sentences into individual tokens (words) and generating features for the Named Entity Recognition (NER) task. Subsequently, tokens are assigned word embedding vectors (numerical representations in space), which were custom-trained using a skip-gram model on a corpus comprising PubMed database abstracts and case studies. These embeddings have a dimensionality of 200 and draw from a substantial vocabulary size of 2.2 million words, ensuring comprehensive coverage of medical terminology. Figure 4 illustrates the whole pipeline, providing a visual representation of the multi-step process that forms the foundation of our text pre-processing approach, optimized for the nuances of clinical text.

The NER model, based on a BLSTM-CNN-Char (character-level AI neural network) architecture[14] detects various PHI elements such as names, organizations, locations, and dates. Two versions of NER models were trained for seven supported languages: *a coarse version* with seven entity types and *a granular version* with thirteen types. These models are central to the de-identification process due to their ability to generalize to unseen data and predict exact entity spans with minimal data loss.

A contextual rule engine complements the NER model, providing flexibility for specific identifier types not covered by the NER model. This engine uses regex (pattern-matching language) matching, prefix and suffix matching, and context analysis to

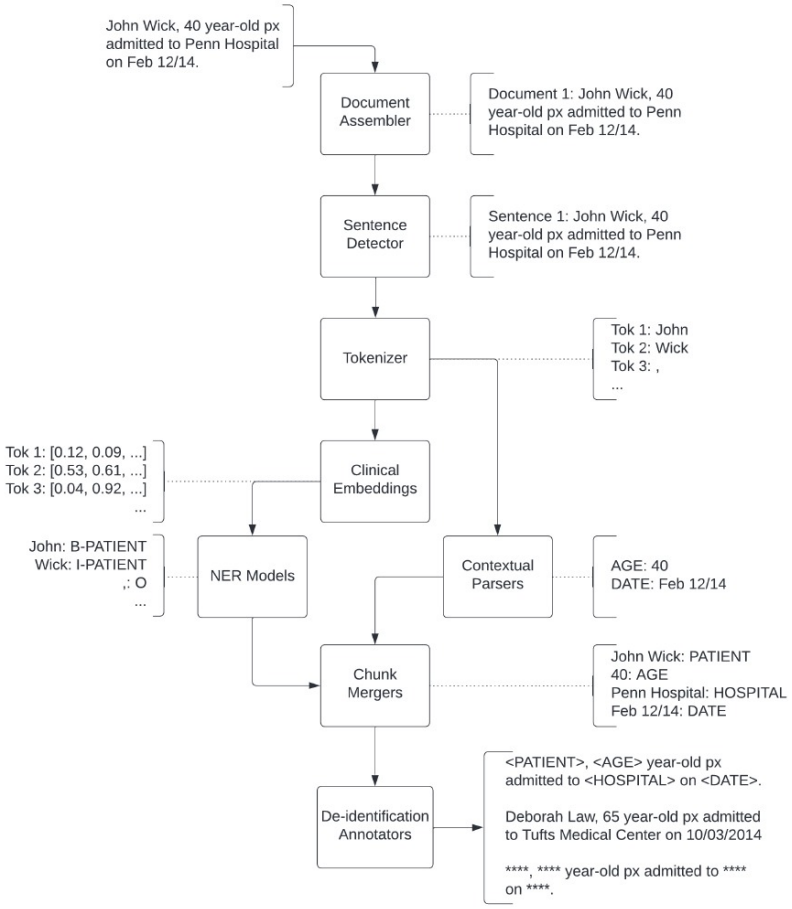

Figure 4. Text Deidentification Pipeline Sequence[3]

reduce false positives and handle complex identifiers like patient IDs or medical record numbers.

The chunk merger resolves conflicting or overlapping detections from both the NER model and rule engine, assigning priorities to each entity type based on the specific use case. This step enhances the overall accuracy of the pipeline.

The final stage, de-identification, generates anonymized text through masking or obfuscation. Masking replaces PHI identifiers with their type or asterisks, while obfuscation substitutes PHI with semantically and linguistically correct surrogate values. To optimize obfuscation, vocabularies are divided into groups based on character length for efficient searching during inference, ensuring the selection of appropriate surrogate values that maintain data integrity.

Augmented NLP pipeline

Our de-identification pipeline enhances John Snow Labs' base pipeline through the sequential application of a series of pre-trained models and regular expressions. Extensive experimentation revealed significant improvements in pipeline performance, particularly in patient name detection, when multiple models were integrated. Notably, older models that exhibited suboptimal performance when used independently proved to be valuable components within a multi-model pipeline. This synergistic effect can be attributed to each model's capacity to identify entities that others might overlook, especially in the context of patient name recognition. In addition to the machine learning models, regular expressions emerged as an effective complementary tool for capturing entities that eluded detection by pre-trained models. For example, while the models demonstrated suboptimal performance in identifying email addresses, URLs, and certain specific date formats, the implementation of simple regular expressions successfully recognized the majority of these entities, thereby substantially enhancing the overall efficacy of the pipeline. Furthermore, the dimensionality of the text field was identified as a critical factor influencing the development of deep learning pipelines. Extensive text fields, such as patient notes and narratives, necessitated the incorporation of numerous models to achieve optimal performance. Conversely, smaller text fields could be de-identified with high accuracy using a more limited number of models. This multi-faceted approach underscores the importance of combining the strengths of machine learning models and rule-based methods in developing robust de-identification systems for clinical text. By addressing the complexities

	Description	Example Text
Original	The original text with identifiable information	Jane is a 48-year-old nurse from Memphis.
NER Detection	Text with Named Entity Recognition (NER) applied	PATIENT is a AGE PROFESSION from CITY.
Masked	Text with sensitive information masked	*** is a *** *** from ***.
Whitelist application	White listed fields are revealed	*** is a 48-year-old *** from ***.
Obfuscated	Text with all identifiable information replaced or altered	Gina is a 45-year-old teacher from Fresno.

Table 9. Different options for representing patient information through various de-identification techniques.

of various text types and entity categories, this strategy offers a comprehensive solution to the challenges of protecting patient privacy while maintaining data utility in healthcare research and applications.

Obfuscation

Obfuscation is the process of replacing the flagged identifiers (PHI entities) with dummy data in accordance with Safe Harbor guidelines. For example, the name “Carey Smith” is replaced with “Ellie Brown” or a phone number like “602-358-7890" is replaced with “780-568-3094". Replacing identifiers with reasonable counterpart reduces the risk of reidentification by making it difficult for an attacker to distinguish between fake identifiers and authentic identifiers that may have been missed.

There are two exceptions in this process: DATE entities in formats recognized by the software are shifted by the days randomly specified for that patient, and ages over 89 years are consolidated into a single category of 90 years or older, as per HIPAA Privacy Rule guidelines[15]. Furthermore, DATE entities in formats that cannot be parsed are masked. For example, 04/05/2019 for a patient with a 30-day date shift is changed to 05/05/2019. However, a 03/04/20 might simply be replaced with <DATE> because it is not clear whether the year is 2020 or 1920.

Obfuscation in automated de-identification presents significant challenges in maintaining data consistency and readability, as highlighted by[16]:

- Name consistency requires mapping all instances of a patient’s name to the same fake name, both within and across documents, potentially aligned with patient IDs.
- Gender consistency involves replacing names with appropriate gender-specific alternatives.
- Age consistency demands replacing ages within suitable ranges based on contextual cues.
- Day shift and date format consistency require careful handling of temporal data, including normalization and consistent shifting of dates.
- Clinical consistency necessitates preserving gender-specific medical information.
- Length consistency aims to maintain the original text structure by replacing entities with fake data of similar length.

These factors are crucial for creating realistic, consistent obfuscated data that preserves the integrity and utility of the original information while protecting patient privacy. Implementing these considerations is essential for effective automated de-identification in real-world healthcare settings, balancing the need for data utility with privacy protection requirements.

To address obfuscation challenges, two key components were developed by John Snow Labs: a *normalization* module and a *faker* module. The normalization module standardizes dates, ages, names, and addresses, ensuring consistent representation of concepts throughout the text and enabling accurate date shifting. The faker module generates random data to replace original concepts, maintaining semantic integrity by considering titles, genders, and addresses. It includes options for masking or total random obfuscation of non-normalizable concepts, particularly dates. The faker module can also maintain text formatting by generating replacements based on Levenshtein distance and allows for custom replacement dictionaries. The system incorporates pre-built faker models for multiple languages, ensuring culturally appropriate obfuscation. These components work in tandem to produce consistent, realistic obfuscated data while preserving the original text’s structure and semantic coherence, thus enhancing the effectiveness of automated de-identification in diverse linguistic and cultural contexts.

To augment the base capabilities, we implemented entity-specific obfuscation policies designed to enhance the authenticity and variability of synthetic data. These policies incorporate both software-generated information and in-house sourced data. This refined approach addresses several challenges, including the indiscriminate tagging of identifiers (such as all zip codes)

Entity Label	Obfuscation Strategy
CITY	Create from a combination of completely fake town and city names by the Faker Python package and real names from Providence patient data. The top 500 most common real city names are added to this list. The 500th most common real city name has over 13,000 patients in Providence's system.
DATE	Dates that can be normalized are shifted (Section 9.2.2); those that cannot are replaced with <DATE>in the text.
DEVICE	Use JSL faker file
DOCTOR	Use JSL faker file
EMAIL	Randomly generate using the Faker Python package
FAX	Merge into the PHONE entity before obfuscating. See PHONE for more details
HEALTHPLAN	Merge into the IDNUM entity before obfuscating. See IDNUM for more details
HOSPITAL	A complete list of Providence's places of service that are categorized as hospitals is pulled to create this list. This list contains approximately 2,000 individual facility names.
IDNUM	Randomly generate ID numbers in the following formats: • 123456789 • A12345 • 12345678
MEDICALRECORD	Merge into the IDNUM entity before obfuscating. See IDNUM for more details.
NAME	Generate with the Faker Python package in a variety of different formats to reflect the diversity of formats in the text data. Formats: • First Last • LAST, FIRST • First • FIRST • Last, First MI • LAST, FIRST MI Reflects Hispanic and non-Hispanic names common in the United States, and small proportions of names from the following locales: • Philippines • India • Thailand • Canada • Ireland • Australia • England
ORGANIZATION	Randomly generate Company names using the Faker Python package.
PATIENT	Merge into the NAME entity before obfuscating. See NAME for more details.
PHONE	Randomly generate various formats using the Faker Python package.
STATE	Use JSL faker file.
STREET	Randomly generate using the Faker Python package
URL	Randomly generate using the Faker Python package.
USERNAME	Randomly generate using the Faker Python package.
ZIP	Randomly generate using the Faker Python package.

Table 10. Obfuscation Strategies for Different Entity Types

and the obfuscation of non-patient identifiers (e.g., hospitals, organizations). By doing so, we improved recall on critical patient-related information. Our implementation integrates custom-generated fake data, produced using the Faker Python package, with JSL faker files. This combination ensures a comprehensive and diverse pool of substitute information, thereby enhancing the effectiveness of the de-identification process. For a detailed description of this methodology, please refer to table 10.

Complementary Methods

In addition to this core technique, our de-identification strategy employs several complementary methods to protect sensitive information, ensuring comprehensive coverage across various data types and formats. These techniques include encryption

Before Deidentification	After Deidentification by Obfuscation
Swedish Medical Center Progress Note Patient Name: Christine L Kraton Age: 74 y.o. DOB: 9/3/1942 Medical Record Number: 20008970125 Date of admission: 4/26/2020 Attending Physician: Albert John Green, MD Assessment: 1. S/p laparoscopic hemicolectomy: chronic sigmoid diverticulitis, surgery on 7/21/20. 2. Antiphospholipid antibody syndrome Plan: - Patient able to resume full dose anticoagulation per surgery.	Providence St Mary Medical Center Progress Note Patient Name: Olivia Bush Age: 74 y.o. DOB: 3/2/1942 Medical Record Number: I58080210 Date of admission: 10/24/2019 Attending Physician: Christian Randall, MD Assessment: 1. S/p laparoscopic hemicolectomy: chronic sigmoid diverticulitis, surgery on 1/18/20. 2. Antiphospholipid antibody syndrome Plan: - Patient able to resume full dose anticoagulation per surgery.

Table 11. Example of the de-identification process of a typical progress note. Highlighted entities were identified as PHI and replaced by obfuscation techniques or date shifting.

for structured data, masking for unevaluated text fields, date shifting for temporal data, ZIP code truncation for geographic information, and preservation of non-identifying data. Together, these methods address different aspects of PHI protection, ensuring comprehensive coverage of data that is not obfuscated through other means:

- **Encryption** plays a crucial role in our structured data de-identification methodology. This approach applies a one-way cryptographic SHA-256 hash function to sensitive data fields, generating a fixed-length output that effectively masks the input string. To mitigate potential risks associated with publicly released encrypted data, we enhance the standard SHA-256 process by appending a random number to the input string before encryption. This additional step significantly reduces the risk of re-identification, even in scenarios where original patient IDs might be compromised. For instance, a patient ID "123456" combined with a random number "48523759" would produce a unique hash, distinct from the hash of the ID alone. Furthermore, our hashing process maintains data integrity by preserving the original data types of hashed fields and ensuring referential consistency across the dataset. This sophisticated encryption method, coupled with careful implementation, provides a robust solution for protecting patient privacy while maintaining data utility in healthcare research and analytics.

- **Masking** is applied to text fields that have not undergone sampling, annotation, and evaluation as an alternative to hashing, primarily to conserve processing time. This technique involves the complete removal of the original text content from these fields. In its place, a string is inserted to indicate that the contents have been masked, effectively obscuring the sensitive information while signaling the occurrence of this privacy-preserving action.

- **Date Shifting** is another crucial technique in healthcare data de-identification, employed to preserve temporal relationships between events while protecting patient privacy. In our methodology, two primary approaches are utilized: *patient-specific random date shifting* and *truncation to the first day of the year*. For patient-specific dates, excluding birth and death dates, a random shift between -364 and 364 days is consistently applied across all dates associated with a given patient. This method maintains the relative timing of events while obscuring the actual dates. For more sensitive information such as birth and death dates, as well as non-patient-specific dates, truncation to the first day of the year (01-01-YYYY) is performed. This approach significantly reduces the risk of re-identification through publicly available datasets while retaining essential temporal information. The implementation of these date shifting techniques, combined with other de-identification methods, allows for the preservation of data utility for research and analytics purposes while adhering to strict privacy protection standards in healthcare data management.

- **Zip Code Truncation** is implemented for all columns containing ZIP codes of patients, their relatives, household members, or employers. Following Safe Harbor guidelines, ZIP codes can be truncated to the first three digits if the geographic unit created by the truncated ZIP has more than 20,000 residents. If the population is 20,000 or fewer, the first three digits are replaced with '000'. Additionally, foreign ZIP or postal codes are also truncated to '000'. This method effectively reduces the risk of re-identification while maintaining compliance with privacy standards.

- **Keep** columns, which are labeled as such, remain unaltered during the de-identification process because they do not contain any patient identifiers.

Custom Scalability Solutions for Large-Scale Batch Inference

In our efforts to scale Apache Spark for processing large datasets, totaling approximately 1.3 billion notes in the historical dataset with an additional 1-3 million new notes added daily, we encountered significant challenges with the default Spark architecture. Despite its efficiency in pre and post-processing tasks, the native Spark configuration exhibited a substantial drop in efficiency when handling batch inference workloads. Even with substantial hardware resources, the largest reasonable vanilla Spark cluster showed decreased performance and hardware utilization as the cluster size increased. To address this, we developed a custom node and workload manager infrastructure. This approach involved assuming the duties of the master node and orchestrating smaller batches across smaller, independent machines. By doing so, we achieved linear scaling to a theoretically unlimited number of worker nodes, ensuring that all machines were utilized at 95-100% capacity throughout the job execution.

Evaluation

The de-identification process for structured and unstructured data presents distinct challenges. Structured data fields allow for complete removal of patient identifiers through techniques like masking, obfuscation, or hashing. However, unstructured text data poses a more complex problem, as it is not feasible to guarantee 100% removal of all patient identifiers, regardless of the method employed - be it stochastic models, programmatic rules, or manual human review. Given this limitation, statistical approaches are necessary to assess the efficacy of text de-identification pipelines. We have developed a comprehensive methodology for evaluating text de-identification, comprising four key phases: sampling, annotation, evaluation, and fine-tuning. This process begins with drawing representative samples from each text field, followed by annotation using advanced deep learning models supplemented by human review. Performance metrics are then calculated to evaluate the pipeline's effectiveness. Finally, the de-identification pipeline undergoes fine-tuning to optimize its performance based on the evaluation results.

Sampling

To rigorously assess the performance of the de-identification model across gender, race, and location, we employed a targeted sampling methodology aimed at identifying potential biases. This approach was crucial because de-identification models might have underperformed or missed specific demographic groups, potentially leading to disproportionate privacy risks. Such biases could have resulted in inadequate protection for certain populations, compromising their privacy and potentially violating ethical and legal standards. In addition, to ensure a clear evaluation, we excluded patients without a listed gender, active address, declared race, or those identifying with more than one race from the sampling.

Next, a stratified random sample was drawn from patient notes and optimized using Neyman allocation[17]. We opted for Neyman allocation, formulated as:

$$n_h = \frac{N_h S_h}{\sum_{i=1}^{H} N_i S_i} n \quad (h = 1, 2, \ldots, H) \tag{1}$$

where N is the population size, H is the number of strata, N_h is the size of stratum h, S_h is the standard deviation of stratum h, n_h is the sample size of stratum h, and n is the total sample size.Neyman allocation was chosen for its ability to minimize the variance of a sample with a fixed size and a given number of strata, and it offers superior efficiency compared to proportional allocation without optimization, as it strategically distributes the sample across strata based on both their size and variability[17,18]. By accounting for the heterogeneity within each stratum, this method enables a more precise estimation of population parameters while maintaining a balanced representation across different strata, thereby enhancing the overall reliability and efficiency of our sampling methodology.

Table 12 illustrates illustrates the allocation of sample sizes for a stratified sampling design based on gender, race/ethnicity, and geographic location. It compares two allocation methods: proportional allocation, which assigns samples based on the size of each stratum, and Neyman allocation, which optimizes sample sizes based on both stratum size and variability (standard deviation). The table shows the population size (Nh), proportion of the total (Prop. Alloc.), variability (Sh), and resulting sample sizes (nh) for each stratum, totaling 1535 samples.

To implement this approach, we first created a patient population by categorizing all eligible patients into 24 unique subgroups based on gender, state, and race. We then applied temporal and content filters to this population, selecting patient notes dated on or after January 1, 2016. From this filtered dataset, we drew a random sample of 100,000 records, allocating an equal number of samples (approximately 4,000) to each of the 24 subgroups. This balanced sampling strategy ensured that each demographic subgroup was adequately represented in our analysis, mitigating potential biases and allowing for more robust comparisons across different patient segments. Please refer to Table 12 for details on the subgroup categorization.

We then applied an NLP pipeline to identify PHI entities within the sampled records. This process allowed us to calculate PHI prevalence, defined as the percentage of records containing at least one PHI entity, for both the full sample and each

Strata	Nh	Prop.	p	Sh	Neyman	nh
F-Cauc-WA	121,785,029	23.80%	0.85	0.3549	22.82%	351
M-Cauc-WA	88,144,946	17.23%	0.86	0.3488	16.23%	249
F-Cauc-OR	47,777,940	9.34%	0.74	0.4412	11.13%	171
F-Cauc-CA	41,631,465	8.14%	0.86	0.3426	7.53%	116
M-Cauc-OR	32,802,216	6.41%	0.74	0.4361	7.55%	116
M-Cauc-CA	30,908,102	6.04%	0.86	0.3443	5.62%	86
F-Other-CA	23,348,944	4.56%	0.83	0.3765	4.64%	71
F-Other-WA	18,190,585	3.55%	0.88	0.3285	3.15%	48
F-Cauc-Low	18,178,775	3.55%	0.78	0.4114	3.95%	61
M-Other-CA	15,234,820	2.98%	0.82	0.3826	3.08%	47
M-Cauc-Low	14,233,714	2.78%	0.79	0.4072	3.06%	47
F-Asian-WA	12,591,889	2.46%	0.89	0.3066	2.04%	31
M-Other-WA	12,214,336	2.39%	0.88	0.3228	2.08%	32
M-Asian-WA	7,296,417	1.43%	0.89	0.3134	1.21%	19
F-Asian-CA	5,788,414	1.13%	0.85	0.3559	1.09%	17
F-Other-OR	5,775,645	1.13%	0.74	0.4363	1.33%	20
M-Other-OR	3,919,761	0.77%	0.73	0.4425	0.92%	14
M-Asian-CA	3,600,504	0.70%	0.86	0.3502	0.67%	10
F-Asian-OR	2,406,397	0.47%	0.73	0.4427	0.56%	9
F-Other-Low	1,942,237	0.38%	0.75	0.4354	0.45%	7
M-Asian-OR	1,405,611	0.27%	0.74	0.4406	0.33%	5
M-Other-Low	1,387,451	0.27%	0.76	0.4289	0.31%	5
F-Asian-Low	701,788	0.14%	0.74	0.4366	0.16%	2
M-Asian-Low	426,915	0.08%	0.76	0.4284	0.10%	1
TOTAL	511,693,901	100%	0.81	0.3881	100.00%	1535

Table 12. Stratified sampling allocation

demographic subgroup. To ensure statistical robustness in our prevalence estimation, we determined the overall sample size using standard statistical parameters. We set a confidence level of 95% (Z=1.96) and a margin of error of 2%. Based on the estimated PHI prevalence of 80% observed in the initial 100,000 records, we used an expected proportion of 0.8. These parameters yielded an overall required sample size of 1,536 records.

To optimize the distribution of this sample across our demographic subgroups, we employed the Neyman allocation schema. This method allowed us to calculate the appropriate sample size for each subgroup, ensuring adequate representation and precision in our subgroup analyses.

Given that small text fields often contain a high percentage of repeated values, random samples were drawn only from distinct values. For these 65 small text fields, columns with low information content relative to the effort required for full-scale annotation, or those clearly composed of patient identifiers, were assigned to the 'Mask' policy group. The review of the remaining small text fields was approached in two steps. During the first round, all rows were manually examined for the presence of any PHI. Those judged by the human reviewer to be free of PHI were considered free of PHI. The remaining 11 small text fields, categorized as containing PHI, were then considered for the full annotation process.

Annotation process

Annotation (labeling data for training or evaluation) is a critical process in identifying patient identifiers within clinical data. A multidisciplinary team of data scientists, clinicians, and healthcare consultants conducts this task, employing a pre-annotation step using deep learning models to extract initial entities. Human annotators then refine these results, guided by a comprehensive annotation guide that evolves throughout the process.

The annotation taxonomy encompasses a wide range of entities, including names, dates, addresses, and contact information. To enhance specificity, suffixes like _PATIENT are added to certain labels. The process defines 'patient' broadly, including relatives and household members, while healthcare providers and facilities are also categorized.

Annotation is performed using Excel spreadsheets in a secure environment for patient notes and small text fields, while John Snow Labs' Annotation Lab software is utilized for larger text fields. Multiple rounds of review and correction ensure accuracy, with improvements continuing throughout the fine-tuning and evaluation phases.

The annotation process underwent strict quality control measures, with each annotation subject to partial or complete review by a secondary annotator. The refinement of annotations was an ongoing process, extending throughout the fine-tuning and evaluation phases. This iterative approach was facilitated by increasingly sophisticated de-identification pipelines, which

Entity Label	Description
AGE	Age of the patient
CITY	City not associated with patient location
CITY_PATIENT	City describing patient location
DATE	Any date, including: 1. Day, month, year 2. Day, year 3. Month, year *Note:* Standalone years, days of week, and seasons are excluded
DEVICE	Device serial number
DOCTOR	Name of healthcare provider
EMAIL	Non-patient email address
EMAIL_PATIENT	Patient email address
FAX	Non-patient fax number
FAX_PATIENT	Patient fax number
HEALTHPLAN	Healthplan ID number (e.g., subscriber number) *Note:* Healthplan names (e.g., 'Aetna') are not included
HOSPITAL	Name of healthcare facility
IDNUM	Any unique ID number that could identify a patient
MEDICALRECORD	Patient medical record number
ORGANIZATION	Name of a non-healthcare organization
PATIENT	Patient name
PHONE	Non-patient phone number
PHONE_PATIENT	Patient phone number
PROFESSION	Patient profession
STREET	Non-patient street address (e.g., 556 N Camino Way)
STREET_PATIENT	Patient street address
URL	Non-patient web address (e.g., patient portal, education materials)
URL_PATIENT	Patient-specific web address (e.g., social media account)
USERNAME	Healthcare provider username
ZIP	Zip code for a non-patient address
ZIP_PATIENT	Zip code for a patient address

Table 13. Annotation Taxonomy

effectively identified overlooked instances and human errors in the initial annotations. The comprehensive annotation effort encompassed 20 distinct text fields, both small and large, totaling 36,180 records. These were meticulously categorized using a taxonomy of 26 unique labels, ensuring a thorough and granular classification of the data. Refer to Table 13 for details.

The final output is a dataset of labeled entities, containing unique identifiers, character positions, entity labels, and text chunks. This dataset serves as a foundation for evaluating the de-identification pipeline, assessing re-identification risks, and refining the process through additional models and rules.

Evaluation

The evaluation of de-identification pipelines involves assessing whether the entities identified by the model match those in the labeled dataset. This process is nuanced due to potential discrepancies between the model's output and the annotated text.

A matching policy was developed to determine whether a model's prediction should be classified as a true positive, false negative, or false positive. For instance, if the model flags "123 Sesame St." as STREET, while the annotated text includes "123 Sesame St., Suite 40," it is considered a true positive as the essential identifying information is removed.

Through experimentation, a match threshold of 50% was established for most entities, meaning if the model identified at least half of the annotated text, it was considered a match. For patient names, a stricter criterion was applied, allowing only the middle initial to be missed.

On the contrary, false positives occur when the model incorrectly flags non-PHI text as an entity, leading to information loss. Additionally, the evaluation considers cases where the model identifies the correct text but applies an incorrect label. However, as the primary focus is on reidentification risk, the matching algorithm does not penalize incorrect labels as long as the identifier is removed.

This evaluation methodology aims to balance the accurate identification of PHI with the preservation of non-sensitive information, ensuring effective deidentification while minimizing unnecessary data loss. Following the annotation and creation

of a labeled dataset, it is possible to evaluate different individual and ensemble models, as well being able to fine-tune them though transfer learning and improve their performance. Deidentification pipelines can be optimized in this way for new fields and use cases.

Evaluation Metrics

We employ both entity-level and record-level metrics to assess the performance of the deidentification process. Entity-level metrics include *recall* and *precision*. Recall calculates the percentage of patient identifiers present in the data that have been correctly identified by the deidentification pipeline whereas precision indicates what percentage of the entities flagged by the pipeline are actual personal identifiers. While a higher precision reduces information loss, our primary goal is to minimize the risk of reidentification, naturally making recall our primary metric and rendering lower precision an acceptable trade-off for maximizing recall.

While entity-level metrics provide valuable insights, they are insufficient on their own to fully evaluate re-identification risk. Record-level metrics play a crucial complementary role in the assessment process, warranting a more comprehensive approach since even a single missed entity within a record can potentially lead to re-identification. To address these concerns, we employ three key record-level metrics: PHI Prevalence Pre-De-identification, PHI Prevalence Post-Deidentification, and Effectiveness. By combining these metrics, we ensure a thorough evaluation of our de-identification process, offering a comprehensive assessment of re-identification risk at both the entity and record levels.

PHI Prevalence Pre-Deidentification represents the percentage of records in the complete dataset that contain one or more personal identifiers before any de-identification has been applied.

$$\text{PHI Prevalence Pre-DeID} = \frac{\text{Records with PHI Entities}}{\text{Total Records}}$$

PHI Prevalence Post-Deidentification is an "all or nothing" **strict** metric, making it the most conservative measure for evaluating the risk of re-identification[12]. As our primary metric for assessing text de-identification, it indicates the percentage of records that contain one or more personal identifiers after de-identification. This metric not only reveals how many records are potentially 'at risk' due to remaining unobfuscated identifiers but also serves as the principal measure used to evaluate our text de-identification in the context of this analysis.

$$\text{PHI Prevalence Post-DeID} = \frac{\text{Records with Missed PHI Entities}}{\text{Total Records}}$$

Effectiveness measures the model's ability to remove all PHI from a dataset, independent of the initial PHI prevalence. An effectiveness close to one signifies that the model is removing all PHI entities from almost 100% of the records. Effectiveness metric allows for performance comparison of the pipeline across different data fields.

$$\text{Effectiveness} = \frac{\text{Records with No Missed PHI Entities}}{\text{Records with PHI Entities Pre-DeID}}$$

Collectively, the metrics mentioned in this section provide a comprehensive evaluation of the deidentification process, balancing the need for thorough removal of identifiers with the preservation of data utility.

Risk Thresholds and Scope

Direct identifiers are pieces of information that can lead to patient reidentification on their own, such as unique patient names, IDs, email addresses, or phone numbers. Dates and locations are considered indirect identifiers. Experts generally consider a maximum threshold for PHI Prevalence Post Deidentification of <5% for direct identifiers and <50% for indirect identifiers to be legally acceptable. In practice, common risk thresholds for reidentification range from 33% for less sensitive data shared with trusted recipients to 5% for highly sensitive data released to the public domain[12]. In our study, we aimed for a more conservative approach, striving to achieve a PHI Prevalence Post DeID of less than 5% for both direct and indirect identifiers. This threshold is stricter than common practice, as PHI Prevalence Post-DeID is not a direct proxy for reidentification risk due to the obfuscation of PHI, which makes it challenging for an attacker to distinguish between real and fake data.

Although the annotation process included the 26 entity types identified by the NLP pipeline, the evaluation metrics were limited to a subset of entities directly related to patient information. Additionally, certain annotated entities were grouped for ease of reporting and analysis. The final list of 9 entities used for evaluation is listed in Table 14.

To estimate the 95% confidence intervals for post-deidentification prevalences, we employed a Bayesian approach, calculating the Highest Density Interval (HDI) using an uninformative uniform prior. This method allowed us to calculate the posterior beta distribution conservatively, without assuming any specific prior distribution[19].

PHI Entity	Annotated Labeled Entity	Description
CITY	CITY_PATIENT	City describing the location of a patient.
DATE	DATE	Any date.
EMAIL	EMAIL_PATIENT	Patient email address.
IDNUM	IDNUM MEDICALRECORD HEALTHPLAN	Any unique patient related ID number. Patient medical record number. Healthplan ID number.
DEVICE	DEVICE	Device serial number.
PATIENT	PATIENT	Patient name.
PHONE	PHONE_PATIENT FAX_PATIENT	Patient phone. Patient fax.
STREET	STREET_PATIENT	Patient street address
URL	URL_PATIENT	Website, social media site, or other url that directly identifies patient.
ZIP	ZIP_PATIENT	Patient zip code

Table 14. PHI Entities Included in Evaluation Metrics

Statistical Framework

We chose a primarily Bayesian framework to both increase interpretability and decrease the effort of integrating new information by utilizing our prior knowledge when updating our model for each release. By directly modeling the posterior distribution, the probability of a range of values is easily calculated. This approach offers significant advantages over traditional frequentist methods, particularly in the context of estimating PHI Prevalence Post-DeID. The Bayesian method allows for intuitive probability estimation and efficient incorporation of prior information, which is especially valuable when dealing with successive data releases. Our model uses a binomial likelihood to represent the presence or absence of PHI in post-deidentified documents as Bernoulli trials. We employ a beta distribution as the prior, which is conjugate to the binomial, initially set as an uninformative uniform prior for conservatism. The resulting posterior distribution, also a beta distribution, can be easily updated with new information. This feature makes our approach particularly suitable for the ongoing process of assessing deidentification effectiveness across multiple data releases. Furthermore, the Bayesian framework is well-suited to handle the skewed distributions typically encountered at low PHI prevalence levels. To ensure accurate representation of the probability of posterior values, we use the Highest Density Interval (HDI). This is particularly important given the expected skewness of the beta distribution at low PHI prevalence levels. The HDI provides a more nuanced and informative view of the uncertainty surrounding our PHI prevalence estimates compared to traditional confidence intervals. This Bayesian approach ultimately provides a robust, interpretable, and maintainable statistical process for estimating PHI Prevalence Post-DeID. It allows for better-informed decisions in managing the delicate balance between patient privacy and data utility in healthcare research and operations, crucial for assessing and improving the effectiveness of deidentification procedures in healthcare data management.

Ethics Statement

All methods were carried out in accordance with relevant guidelines and regulations. The study protocol was reviewed and approved by the Providence Health Care Institutional Review Board (IRB). Raw, identifiable patient data were accessed under IRB approval. Informed consent was obtained from all subjects and/or their legal guardians, or a waiver of consent was granted by the IRB in accordance with institutional and regulatory requirements.

References

1. Myrick, K. L., Ogburn, D. F. & Ward, B. W. Percentage of office-based physicians using any electronic health record (ehr)/electronic medical record (emr) system and physicians that have a certified ehr/emr system, by us state: National electronic health records survey, 2017. Tech. Rep., National Center for Health Statistics (2019).
2. Negash, B. *et al.* De-identification of free text data containing personal health information: a scoping review of reviews. *Int. J. Popul. Data Sci.* **8** (2023).
3. Kocaman, V., Haq, H. U. & Talby, D. Beyond accuracy: Automated de-identification of large real-world clinical text datasets. *ArXiv* **abs/2312.08495** (2023).
4. Kocaman, V. & Talby, D. Accurate clinical and biomedical named entity recognition at scale. *Softw. Impacts* **13**, 100373, DOI: https://doi.org/10.1016/j.simpa.2022.100373 (2022).
5. Piotrowski, M. Using spark nlp to de-identify doctor notes in the german language. In *NLP Summit* (2022).

6. Zaharia, M. A. *et al.* Apache spark. *Commun. ACM* **59**, 56 – 65 (2016).
7. Kocaman, V. & Talby, D. Spark nlp: Natural language understanding at scale. *ArXiv* **abs/2101.10848** (2021).
8. Tomer, V. Lessons learned de-identifying 700 million patient notes with spark nlp. In *NLP Summit* (2022).
9. U.S. Department of Health & Human Services. Nondiscrimination in health programs and activities (2024). Accessed on October 24, 2024.
10. Xiao, Y., Lim, S. F., Pollard, T. J. & Ghassemi, M. In the name of fairness: Assessing the bias in clinical record de-identification. *Proc. 2023 ACM Conf. on Fairness, Accountability, Transpar.* (2023).
11. Radhakrishnan, L. *et al.* A certified de-identification system for all clinical text documents for information extraction at scale. *JAMIA Open* **6** (2023).
12. El Emam, K. & Arbuckle, L. *Anonymizing Health Data: Case Studies and Methods to Get You Started* (O'Reilly Media, Inc., 2013), 1st edn.
13. Douglass, M., Clifford, G., Reisner, A., Moody, G. & RG, M. Computer-assisted de-identification of free text in the MIMIC II database. In *Computers in Cardiology*, 341–344, DOI: 10.1109/CIC.2004.1442942 (IEEE, 2004).
14. Chiu, J. P. & Nichols, E. Named entity recognition with bidirectional lstm-cnns. *Transactions association for computational linguistics* **4**, 357–370 (2016).
15. U.S. Department of Health & Human Services. Guidance regarding methods for de-identification of protected health information in accordance with the health insurance portability and accountability act (hipaa) privacy rule (2023). Accessed on October 24, 2024.
16. Yogarajan, V., Pfahringer, B. & Mayo, M. A review of automatic end-to-end de-identification: Is high accuracy the only metric? *Appl. Artif. Intell.* **34**, 251 – 269 (2019).
17. Cochran, W. *Sampling Techniques*. Wiley Series in Probability and Statistics (Wiley, 1977).
18. Wright, T. A simple method of exact optimal sample allocation under stratification with any mixed constraint patterns (2014).
19. Hyndman, R. J. Computing and graphing highest density regions. *The Am. Stat.* **50**, 120–126 (1996).

Additional information

Funding Acknowledgement
This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors

Author contributions
Veysel Kocaman contributed significantly to the initial conception and design of the paper, M.Aytug Kaya wrote the majority of the manuscript, and was primarily responsible for articulating the technical details of the study with Veysel Kocaman Lindsay Mico worked in close collaboration with Nadaa Taiyab, Tae Surh, Yuqing Guo, and Vivek Tomer on the conception and design, contributed to writing the technical sections of the manuscript, and provided guidance on the organization of the paper and conducted the experiments, providing critical data for the study. Robert Kramer worked on sampling, equity analysis and all other statistical concepts. David Talby provided thorough proofreading of the manuscript, contributed to the impact analysis, and was responsible for the overall layout of the manuscript. His insights and oversight were crucial in shaping the final form of the paper. All authors read and approved the final manuscript.

Competing interests
The authors declare no competing interests.

Correspondence
All correspondence should be addressed to M.A.K.

Data Availability
The raw data generated and/or analyzed during the current study contain protected health information (PHI) and cannot be shared publicly in accordance with HIPAA regulations. However, de-identified datasets generated during the current study are available from the corresponding author upon reasonable request.