ScientificScholar®
Knowledge is power
Publisher of Scientific Journals

www.surgicalneurologyint.com

Surgical Neurology International

Editor-in-Chief: Nancy E. Epstein, MD, Professor of Clinical Neurosurgery, School of Medicine, State U. of NY at Stony Brook.

SNI: Computational

Editor
Pieter L. Kubben, MD, PhD
Maastricht University Medical Center, Maastricht, The Netherlands

SNI®
Open Access

Original Article

Evaluating the accuracy of automated and semi-automated anonymization tools for unstructured health records

Layan Abdulilah Alrazihi[1], Sayan Biswas[2], Joshi George[3]

[1]Faculty of Biology, Medicine and Health, University of Manchester, Manchester, [2]Royal Preston Hospital, Lancashire Teaching Hospitals NHS Foundation Trust, Preston, [3]Department of Neurosurgery, Manchester Centre for Clinical Neurosciences, Salford Royal Hospital, Manchester, England, United Kingdom.

E-mail: *Layan Abdulilah Alrazihi - layan_alr@hotmail.com; Sayan Biswas - sayan.biswas@nca.nhs.uk; Joshi George - joshi.george@nca.nhs.uk

***Corresponding author:**
Layan Abdulilah Alrazihi,
Faculty of Biology, Medicine and Health, University of Manchester, Manchester, United Kingdom.

layan_alr@hotmail.com

Received: 06 May 2025
Accepted: 18 June 2025
Published: 01 August 2025

DOI
10.25259/SNI_459_2025

Quick Response Code:

ABSTRACT

Background: Utilization of unstructured clinical text in research is limited by the presence of protected health identifiers (PHI) within the text. To maintain patient privacy, PHI must be de-identified. The use of anonymization tools such as Microsoft Presidio and Philter has been recognized as a potential solution to the challenges of manual de-identification. Therefore, the primary objective of this study is to evaluate the accuracy and feasibility of using Microsoft Presidio and Philter in de-identifying unstructured clinical text.

Methods: A sample of 200 neurosurgical documents, temporally distributed across 10 years, was extracted. The data were processed by Microsoft Presidio and Philter. Each document was manually screened for the ground truth which was used as a reference point to evaluate the accuracy of each tool. Data analysis was conducted using Python.

Results: A median of 8 PHI were manually de-identified per document. Both tools were individually capable of de-identifying a median of 6 PHI per document. Each tool de-identified PHI with an accuracy of 96%. Presidio demonstrated precision of 0.51 and a recall of 0.74, while Philter had precision and recall of 0.35 and 0.79, respectively.

Conclusion: The performance of each tool supports their use in anonymizing unstructured clinical text. Formatting variations between texts limited the performance of both tools. To conclude, further research is required to optimize the tools' output and assess the reliability in de-identifying diverse and previously unseen clinical text, thus allowing the use of unstructured clinical text in medical research.

Keywords: Anonymization, Artificial intelligence, De-identification, Neurosurgery, Protected health identifiers

INTRODUCTION

The increasing use of electronic health records is an international means to improve healthcare quality and efficiency.[2] This transition has made a vast amount of data readily accessible for research purposes.[15] Much of this data is embedded in unstructured clinical text. Free text is variable as it is a natural way to document clinical events. Thus, much of the unstructured text contains protected health identifiers (PHIs). Unstructured clinical text is a valuable resource, particularly for research, service evaluation, and administrative processes that aim to improve patient care. However, confidential patient information restricts the secondary use of electronic health records for research as it does not align with data protection laws in the United Kingdom. Specifically, The Data Protection Act (2018), which sets strict guidelines on data collection and handling to protect patient privacy. A key component of The Data Protection Act (2018) is data minimization, which states that retrieving information from electronic health records should be relevant and limited to only what is necessary for a research purpose.[4]

To enable the safe use of unstructured clinical text for research, patient privacy must be protected through de-identification of PHI.[4,14] Manual de-identification is inefficient and impractical for large datasets. As a result, automated anonymization tools have been developed to address such issues, allowing for faster processing while maintaining patient privacy.[13]

Presidio, developed by Microsoft, is a customizable, automated de-identification tool designed to detect and redact PHI from unstructured text. Researchers have adapted Presidio for clinical unstructured text anonymization, with studies reporting a wide range of F1 scores (0.38–0.85), depending on dataset characteristics and customization levels.[11]

Philter, developed by the University of California, San Francisco (UCSF), is another privacy-focused, open-source de-identification tool. It has been evaluated using two datasets from UCSF, achieving a recall of approximately 99% on both datasets, indicating high sensitivity in detecting PHI. However, its precision is widely reported to be low, raising concerns about over-redaction and loss of meaningful clinical data.[8,9,12,13]

Despite advancements in de-identification tools, research on anonymization in neurosurgical clinical documents remains limited, particularly in the United Kingdom. In addition, no comparative analysis of Presidio and Philter has been conducted to date. This study aims to: (1) Evaluate the accuracy, precision, and recall of automated (Presidio) and semi-automated (Philter) anonymization tools on UK-based neurosurgical unstructured clinical text; (2) assess the feasibility of using these tools for de-identifying PHI in unstructured clinical text; and (3) compare the performance of Presidio and Philter in de-identifying unstructured clinical text.

MATERIALS AND METHODS

Study design

This is a pilot single-center comparative retrospective analysis to evaluate the accuracy of anonymization tools.

This project was approved by the Research and Innovation department at our National Health Service (NHS) trust. All documents were blinded and processed in a secure, black-box environment; therefore, patient consent was not required. Columns including names, date of birth, and NHS or hospital number were removed during the extraction process, and only unstructured clinical text was retrieved.

Document types and source

Two hundred documents temporally distributed across a time period of 10 years from 2012 to 2022 were selected. Documents included unstructured clinical text from patients undergoing lumbar spinal surgery at a tertiary neurosurgical unit in the United Kingdom. Document types included admission notes, postoperative notes, clinical notes, and discharge summaries.

Selection of anonymization tools

Two anonymization tools were selected: Microsoft Presidio and Philter by the UCSF.[11,13] Presidio is an automated anonymization tool that uses a series of text and data processing methods to identify the selected PHI and anonymize them.[11] Philter is semi-automated anonymization tool that uses regular expression, blacklist, and whitelist matching as well as part of speech tagging to detect and anonymize the predefined PHI.[13] Regular expression is a coding tool that uses a structured pattern of characters to detect a specific sequence in text, allowing computers to find patterns of numbers or words within unstructured text.

The selection criteria were based on accessibility and customization. Both tools are open-source and allowed customization of protected health identifiers (PHI) categories and regular expression patterns. The customizability of the tools allows for identification of relevant text, enabling their configuration to adapt to neurosurgical unstructured clinical text in the United Kingdom.

Anonymization

Both anonymization tools were initialized and configured to detect a total of 7 PHI selected based on the General Data Protection Regulation.[6] The PHI and detection methods are listed in Table 1.

Statistical analysis

Each original document was screened to manually identify PHI and extract the ground truth. The ground truth was used as comparator to evaluate the output of each anonymization tool. The number of PHI identified in the ground truth was recorded. The number of PHI correctly identified by Presidio and Philter were recorded individually. A correctly de-identified PHI was defined as redacted text regardless of the label used by Presidio. The number of missed PHI was

Table 1: PHI types and detection methods.

Protected health identifiers (PHI)	Anonymization technique		Pattern
	Presidio	Philter	
Person name	Named entity recognition Context words	Blacklist matching Part of speech tagging	Dr. John Smith Dr. Smith
Date or Time	Named entity recognition Context words	Regular expression	Dd/mm/yy Dd/mm/yyyy Dd month yy Dd month yyyy
Location	Named entity recognition Regular expression	Regular expression Blacklist matching	
UK NHS number or Hospital number	Named entity recognition Regular expression	Regular Expression	123 321 1234
Medical license	Regular expression	Regular expression	1234567
Phone number	Named entity recognition Regular expression	Regular expression	+44 7123456789 (+44) 7123456789 07123456789
Email address	Named entity recognition Regular expression	Regular expression	John.smith@hotmail.co.uk

NHS: National health service

calculated based on these two values for each tool. Missed PHI was defined as PHI that was completely missed. Text that was incompletely redacted was evaluated by three researchers who determined whether it should be counted as correctly de-identified or missed based on contextual analysis. In addition, misidentified text – text which was incorrectly labeled as a PHI – was recorded for each tool. This data was used in statistical analysis conducted on python. Both techniques were evaluated through four performance metrics. These include accuracy, recall, precision, and F1 score. Accuracy was defined as the median number of words correctly excluded as non-PHI by each tool providing a true analysis of accuracy and omitting the compounding error rate of anonymization tools

$$Accuracy = \frac{\begin{array}{c} Median\,words\,per\,document \\ -Median\,PHI\,correctly\,identified \end{array}}{Median\,words\,per\,document}$$

Recall was defined as the ratio of correctly identified PHI (true positive rate):

$$Recall = \frac{Correctly\,identified}{Correctly\,identified + Missed}$$

Precision was defined as the ratio of correctly identified PHI from all the words identified:

$$Precision = \frac{Correctly\,identified}{Correctly\,identified + Misidentified}$$

F1-score was defined as the harmonic mean of precision and recall:

$$F1\,Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

RESULTS

Document characteristics

Four types of documents were analyzed by each tool. Description and frequency of document types are listed in Table 2. Clinical notes comprised much of our sample accounting for 45% of documents. The total number of PHI identified in the original documents was 1449 [Table 3]. Our dataset primarily consisted of PHI in the category date and time (48.2%) and person name (44.1%). Location accounted for 6.6% of the PHI, while medical license, phone number, and NHS number collectively contributed <1%. There were no occurrences of email address in our dataset.

Anonymization

Manual extraction of the ground truth identified a total of 1449 PHI, with a median of 8 per document. Both Presidio and Philter correctly identified a median of 6 PHI per document demonstrating an accuracy of 96% and an error rate of 4% each [Table 4]. Presidio achieved a recall 0.74, precision of 0.51, and an F1 score of 0.60. In comparison, Philter exhibited a recall of 0.79, precision of 0.35, and F1 score of 0.49 [Table 5]. In addition, Philter misidentified a median of 10 words per document, while Presidio misidentified a median of 5 words per document [Figure 1].

Table 2: Documents sourced from neurosurgical notes.

Document type	Description	Frequency (%)
Admission Clerking	Document at admission. Includes patient presenting complaint, medical history, examination, and investigation findings and management plan.	19 (9)
Operation Note	A summary of the surgical procedure performed. Includes techniques used, instruments used, complications if any and postoperative management.	47 (23)
Clinical Note	A document created during inpatient care. Includes assessments and changes in care plan.	90 (45)
Discharge Summary	Document at discharge. Provides a summary of patients' hospitalization. Includes initial complaint, diagnosis, management plan during admission, and instructions after discharge.	46 (23)

Table 3: Frequency of PHI in the ground truth.

PHI category	Frequency in the ground truth *n* (%)
Person	639 (44.1)
Date or time	699 (48.2)
Location	96 (6.6)
Medical license	9 (0.6)
Phone number	5 (0.4)
NHS number	1 (0.1)
Email address	0 (0)
Total	1449 (100)

NHS: National health service, PHI: Protected health identifiers.

Table 4: Comparison of anonymization performance for Presidio and Philter.

	Ground truth	Presidio	Philter
Total PHI (median)	1449 (8)	-	-
PHI correctly identified (median)	-	1067 (6)	1145 (6)
PHI missed	-	382	304
Accuracy	-	96%	96%

PHI: Protected health identifiers.

Table 5: Comparison of recall, precision and F1 scores for Presidio and Philter.

	Presidio	Philter
Recall	0.74	0.79
Precision	0.51	0.35
F1 Score	0.60	0.49

F1 score: The harmonic mean of precision and recall.

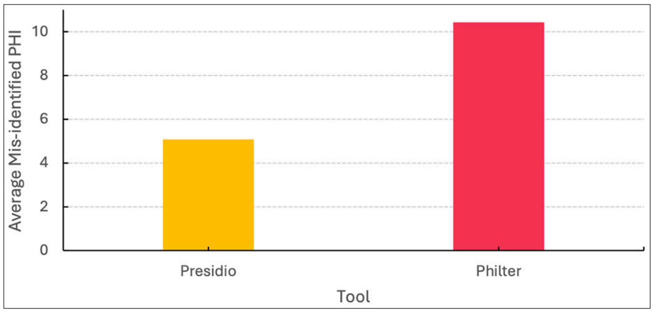

Figure 1: Bar chart comparing the average number of misidentified PHI between Presidio and Philter.

DISCUSSION

This is the first ever study to our knowledge that has analyzed the efficacy of two different anonymization techniques on UK-based neurosurgical unstructured clinical text. We observed a total number of 1449 PHIs across 200 documents, 48% of which were dates and times. About 44% of the manually identified PHI were Person names including mostly surgeon names. The least occurring PHI was medical license, phone number, NHS number, and email address which reflect that a certain level of anonymity is maintained in unstructured text, particularly because this information is recorded in structured columns on the EPR system.

Both tools showed an accuracy of 96% in identifying PHI. This represents the number of words correctly identified as non-PHI by both tools, demonstrating their effectiveness in distinguishing between PHI and non-PHI words.

Presidio achieved a recall of 0.74. However, its precision was limited to 0.51, indicating that nearly half of the words it classified as PHI were non-PHI. This suggests a higher rate of false positives, which could lead to unnecessary redactions, potentially reducing context and affecting the usability of the de-identified data for research purposes.

Similarly, Philter demonstrated a recall of 0.79, indicating slightly better PHI detection than Presidio. However, its precision was lower at 0.35, meaning that a significant proportion of its identified PHI words were non-PHI. Figure 1 further illustrates this, showing that Philter misidentified nearly twice as many words as Presidio. This reduced precision may be attributed to Philter's reliance on rule-based approaches such as regular expressions, whitelists, and blacklists rather than machine learning techniques.[12,13] As a result, Philter was prone to unnecessary de-identification, particularly with misspelled words, abbreviations, or instances where sentences lacked proper spacing.[9,12] Without recognizing word boundaries effectively, Philter often misinterpreted multiple words as a single entity that did not match its whitelist, leading to excessive redactions. This can be attributed to Philter's lack of deep learning and context awareness.[12]

In contrast, Presidio occasionally demonstrated the ability to recognize words despite minor errors in spelling or spacing, reducing unnecessary de-identification. This suggests that Presidio may offer a more contextually aware approach compared to Philter, which relies strictly on predefined rules.[5,13]

When tested on Australian-based documents, Presidio achieved a recall of 0.81 and an F1 score of 0.85, demonstrating significantly better performance compared to our findings. This difference may be attributed to the preprocessing of documents using optical character recognition (OCR) with Python's Pytesseract library, which likely standardized the text and corrected potential errors before de-identification.[7] Across the literature, Presidio has demonstrated similar performance, with reported precision ranging from 0.76 to 0.88.[1,3,7,10]

Philter, on the other hand, was originally tested during its development at the University of California, where it achieved an impressive recall of approximately 99% across two individual datasets.[13] However, its precision has remained consistently low, with most studies reporting a precision of <0.4.[8,9,12]

Despite these reported performances, neither tool has been evaluated within a neurosurgical UK setting. While fine-tuning these tools has been shown to enhance precision and recall, UK medical documents exhibit significant variations in PHI formatting, which both Presidio and Philter struggle to handle.[8,9] One key limitation is that both tools are primarily trained on US-based datasets, making them less effective for UK-specific PHI patterns. Therefore, UK-based datasets require localized de-identification models tailored to country-specific document structures.

Future de-identification strategies should focus on integrating large language models (LLMs) to preprocess text, correct misspellings, and accurately recognize medical abbreviations and terms.[5] Combining these advanced models with rule-based approaches could create a hybrid de-identification tool with maximized precision and recall, making it more reliable for both clinical, research, and administrative applications.

Limitations

This study has several limitations. First, it is a single-center study, which may limit the generalizability of the findings. In addition, the documents used were not pre-processed or standardized before de-identification, which may have contributed to inconsistencies in performance. Finally, there has been no external validation of these tools, highlighting the need for further multi-center evaluations to assess their effectiveness in diverse clinical environments.

CONCLUSION

Although formatting variations between texts limited the performance of both tools. The use of automated anonymization tools has proven to be a feasible de-identification method. Further research is required to optimize the tools' output and assess the reliability in de-identifying diverse and previously unseen clinical text, thus allowing the use of unstructured clinical text in medical research.

Ethical approval: The Institutional Review Board approval is not required.
Declaration of patient consent: Patient's consent not required as patients identity is not disclosed or compromised.
Financial support and sponsorship: Nil.
Conflicts of interest: There are no conflicts of interest.
Use of artificial intelligence (AI)-assisted technology for manuscript preparation: The authors confirm that there was no use of artificial intelligence (AI)-assisted technology for assisting in the writing or editing of the manuscript and no images were manipulated using AI.

REFERENCES

1. Asimopoulos D, Siniosoglou I, Argyriou V, Karamitsou T, Fountoukidis E, Goudos SK, *et al*. Benchmarking advanced text anonymisation methods: A comparative study on novel and traditional approaches. In: 13th International conference on modern circuits and systems technologies (MOCAST) authors;2024;1-6.
2. Cresswell KM, McKinstry B, Wolters M, Shah A, Sheikh A. Five key strategic priorities of integrating patient generated health data into United Kingdom electronic health records. J Innov Health Inform 2018;25:254-9.
3. Dalamagkas C, Asimopoulos D, Radoglou-Grammatikis P, Maropoulos N, Lagkas T, Argyriou V, *et al*. AI4COLLAB: An AI-based threat information sharing platform. In: Proceedings of the 2024 IEEE international conference on cyber security and resilience, CSR 2024. Institute of Electrical and Electronics Engineers Inc., 2024. p.783-8
4. Data Protection Act; 2018. Available from: https://www.legislation.gov.uk/ukpga/2018/12/contents/enacted [Last accessed on 2025 Feb 05].
5. Deleger L, Molnar K, Savova G, Xia F, Lingren T, Li Q, *et al*. Large-scale evaluation of automated clinical note de-identification and its impact on information extraction. J Am Med Inform Assoc 2013;20:84-94.
6. I (Legislative acts) regulations regulation (EU) 2016/679 of the European parliament and of the council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data,

and repealing directive 95/46/EC (general data protection regulation) (text with EEA relevance);2016;1-88.

7. Kotevski DP, Smee RI, Field M, Nemes YN, Broadley K, Vajdic CM. Evaluation of an automated presidio anonymisation model for unstructured radiation oncology electronic medical records in an Australian setting. Int J Med Inform 2022;168:104880.
8. Kraljević Z, Shek A, Au-Yeung J, Sheldon EJ, Al-Agil M, Shuaib H, *et al*. Validating Transformers for Redaction of Text from Electronic Health Records in Real-World Healthcare. Proceedings of the IEEE 11th International Conference on Healthcare Informatics (ICHI); 2023. p. 3116-24.
9. Kraljevic Z, Shek A, Yeung JA, Sheldon EJ, Shuaib H, Al-Agil M, *et al*. Validating transformers for redaction of text from electronic health records in real-world healthcare. In: Proceedings - 2023 IEEE 11th international conference on healthcare informatics, ICHI 2023. Institute of Electrical and Electronics Engineers Inc.; 2023. p. 544-9.
10. Lison P, Pilán I, Sánchez D, Batet M, Øvrelid L. Anonymisation models for text data: State of the art, challenges and future directions. United States: Association for Computational Linguistics; 2021.
11. Mendels O, PC, VLN, HS, RT, LL. Microsoft presidio: Context-aware, pluggable, and customizable PII anonymization service for text and images. Microsoft, 2018. [Last accessed on 01 May 2025].
12. Morris JX, Campion TR, Nutheti SL, Peng Y, Raj A, Zabin R, *et al*. DIRI: Adversarial patient reidentification with large language models for evaluating clinical text anonymization. AMIA Jt Summits Transl Sci Proc 2025;2025:355-64.
13. Norgeot B, Muenzen K, Peterson TA, Fan X, Glicksberg BS, Schenk G, *et al*. Protected health information filter (philter): Accurately and securely de-identifying free-text clinical notes. NPJ Digit Med 2020;3:57.
14. Paul A, Shaji D, Han L, Del-Pinto W, Nenadic G. DeIDClinic: A multi-layered framework for de-identification of clinical free-text data. arXiv [Preprint]; 2024. Doi: 10.13140/RG.2.2.20551.92327.
15. Wang Y, Wang L, Rastegar-Mojarad M, Moon S, Shen F, Afzal N, *et al*. Clinical information extraction applications: A literature review. J Biomed Inform 2018;77:34-49.

How to cite this article: Alrazihi LA, Biswas S, George J. Evaluating the accuracy of automated and semi-automated anonymization tools for unstructured health records. Surg Neurol Int. 2025;16:313. doi: 10.25259/SNI_459_2025

Disclaimer

The views and opinions expressed in this article are those of the authors and do not necessarily reflect the official policy or position of the Journal or its management. The information contained in this article should not be considered to be medical advice; patients should consult their own physicians for advice as to their specific medical needs.