Advancing Medical Data De-Identification through Adaptive Privacy Preservation

Indraneel Mukhopadhyay, Bannishikha Banerjee

Amity Institute of Information Technology, Amity University Kolkata

**Abstract.** We present SARA (Semantic Adaptive Redaction and Anonymization), a new framework that is particularly appropriate for context-dependent de-identification of clinical narratives. Underlying it is the SARA-DeID algorithm, which combines four main elements: Semantic analysis through natural language processing and medical ontologies for the identification of context-dependent protected health information (PHI) in unstructured text; Adaptive risk scoring to measure the re-identification risk of each PHI component and adjust the anonymization level for it; A rule-learning hybrid mechanism to automatically learn and update PHI detection rules through transformation-based error-driven learning; and Utility-preserving transformation techniques to substitute PHI with contextually suited surrogates—e.g., concept tags, pseudonyms, or generalized values—to protect the analytical value of the data. This solution overcomes limitations of traditional solutions: unlike those coarse generalization-based and necessarily suffering high levels of information loss, or static rule-based approaches based on manually compiled lists and devoid of context-sensitivity, SARA is adaptive, context-sensitive, and data-driven. Its semantic analysis adds accuracy, the adaptive scoring balances anonymization for content sensitivity, and the learning-based model guarantees scalable and accurate PHI detection. Overall, SARA strikes a good balance between data utility and privacy protection, and is particularly appropriate for next-generation clinical data environments.

**Keywords.** Medical Data De-Identification; Semantic Analysis; Privacy Preservation; Adaptive Risk Scoring; Utility-Preserving Transformation;

**1. Introduction**

Greater access to electronic health records (EHRs) and clinical narratives has created new opportunities for medical research, artificial intelligence applications, and healthcare analytics [1]. This use is however constrained by stringent privacy regulations mandating the removal of personally identifiable information (PII) or protected health information (PHI) in order to ensure patient privacy [2]. Traditional de-identification methods, including rule-based method, and generalization method, such as k-anonymity and l-diversity, often fail to find the point of the optimal trade-off between privacy and utility [3]. These methods are not context sensitive (and thereby part of the identifiers would be omitted) and don't lead to extensive information loss so the secondary analysis of data is less applicable [4].

To overcome these limitations, we present SARA (Semantic Adaptive Redaction and Anonymization), a novel context-aware deidentification system developed specifically for clinical narratives. SARA uses semantic analysis, adaptive risk scoring, hybrid rule-learning, and utility-preserving transformation to create an effective, intelligent de-identification pipeline. By leveraging domain-specific ontologies and NLP support, SARA provides enhanced PHI recognition accuracy [5] and incurs adaptive scoring mechanisms, which tune the level of the anonymization dynamically by taking into account the content sensitivity [6]. The hybrid learning module dynamically tunes PHI detection rules for optimal sensitivity and specificity, and to learn with new contexts of data [7].

SARA will facilitate safe, ethical and analytically valid access to sensitive health data to support research and innovation, while providing compliance with data protection [8]. The framework also establishes a new level of privacy-conscious sharing that is applicable to a diverse set of domains that necessitate data utility as well as privacy assurances.

**.2. Related Work**

Clindocs are a mature yet still rapidly evolving field with an increasing amount of available unlabeled clinical documents, and increasingly tight privacy laws such as HIPAA and GDPR [1]. Rule-based, statistical, machine learning approaches rule-based, statistical, and machine learning-based approaches are the most common methods and each presents its own advantages and limitations.

Rule-based approaches are one of the earliest and most widely used methods of PHI removal. They rely on predefined patterns, dictionary or regular expressions for identifying and removing sensitive information (e.g., names, dates, locations, IDs, etc) [2]. Perhaps unsurprisingly, rule-based approaches achieve high accuracies on the detection of entities with clear boundaries, but they are extrmely brittle and do not account for the context. They usually cannot pick up implicit or informal terms which are buried in texts, particularly if they do not conform to the models in the rules [3]. Furthermore, it is a labor-intensive and error-prone process to maintain and update the rules manually, especially among institutions or languages.

Generalization-based models, such as k-anonymity, :l-diversity, and t-closeness, mitigate privacy disclosure by allocating records into equivalence classes and replacing some:attributes with their derived values : by more aggregate values [ 4 ]. They work fine on structured data such as data from clinical tables, but they do not perform well in unstructured clinical text where semantics is the focus of the task. More importantly, excessive generalization often leads to a significant loss of analytical power, making them less attractive for use in research settings requiring high resolution [5].

With the advent of artificial intelligence (AI), machine learning and deep learning models are being employed to detect PHI from unstructured text. Supervised models, including Conditional Random Fields (CRFs), Support Vector Machines (SVMs), and more recently, the transformer-based models such as BERT and BioBERT have shown promise [6]. ” They can handle intricate patterns and semantics and can outperform rule-based systems in terms of recall and precision. They require large quantities of annotated data to train, which are usually restricted or not even available for privacy concerns [7]. They are also found to perform poorly when tested on out-of-distribution data or on data from multiple institutions and clinical specialties.

Several hybrid systems have been proposed to tackle this limitations by combining rulesheuristic methods with statistical/ machine-learning based techniques [8]. These systems attempt to emulate the high-precision of rule-based systems and the flexibility of machine learning. For the most part, however, these hybrid systems remain static and do not include dynamic context adaptation or a real-time risk score. Semantic analysis or ontological reasoning is also not frequently used, despite the importance of both to any process of finding domain-specific nuances in medical text [9].

Recent work has begun to explore utility-preserving anonymization, in which PHI is not simply removed or redacted but replaced with semantically appropriate substitutes (e.g., concept codes, pseudonyms) in order to retain the analytic value of the data [10]. This is particularly important in downstream tasks such as cohort identification, temporal pattern mining and clinical outcome prediction.

Our SARA architecture builds upon these advances but overcomes important limitations by integrating semantic ontologies, adaptable risk scores, hybrid rule-learning, and utility-preserving transformations in a unified context-sensitive framework. SARA differs from stationary models by adjusting the strength of anonymization dynamically, i.e., based on the context sensitivity, in order to both, satisfy privacy compliance and usability requirements.

**3. Proposed Methodology**

SARA uses a multistage pipeline to find and remove PHI in clinical text. A preliminary semantic analysis takes place using state-of-art NLP models, domain-specific versions of BERT, for detecting PHI entities in a context-aware manner. Pre-trained models (e.g., BioBERT and ClinicalBERT [1], [2]) that are fine-tuned on biomedical corpora are used to tag sensitive items such as names, dates, locations, and identify ID. Deep contextual models improve precision and recall of basic patter matching based on semantic relations and linguistic context.

Second, re-identification risk is computed for each entity using adaptive risk scores. This involves observing the context in which the entity occurs—eg rare or unknown phrases have higher risk scores and lead to more aggressive anonymization. This adaptive process utilizes new research measuring "uniqueness of context" to estimate privacy risk at the document level [3]. Not redacting too much is the adaptive scoring capability, which enables SARA to balance privacy and utility by redacting entities only that collectively sum up to re-identification risk.

Third, a hybrid module for rule-learning enhances the robustness of the system. This layer is composed of hand-crafted dictionaries (e.g., frequent names), regular expressions for structured identifier patterns (e.g., phone numbers and medical record numbers), and heuristic rules. They are used together with the neural taggers to catch any remaining PHI not identified by the model. SARA-DeID combines the output of all the detection methods—dictionary lookup, rule-based tagging, and contextual neural tagging—into one list of entities [4][5].

At the last stage, utility-preservation transformation is called. Rather than deleting PHI in a brute-force fashion, the system substitutes identifiers by semantically equivalent surrogates or generalized values. Names, for example, can be substituted by realistic pseudonyms, dates offset by fixed quantities, and numeric identifiers randomized within plausible ranges. The transformations preserve the structural and statistical properties of the original data, with negligible performance degradation in downstream applications like outcome prediction [6]. This is in contrast to conventional k-anonymity-based generalization, which dramatically reduces analytic value [7].

SARA integrates semantic modeling, risk-driven adaptation, and hybrid rule-based support to achieve high recall, enhanced privacy protection, and retained data utility. This approach outperforms traditional scrubbers and static anonymization by dynamically focusing on genuinely identifying information and adapting to context [8]. [Figure 1] shows proposed methodology.



Figure 1. Proposed Methodology

**4. Implementation Details**

The SARA-DeID algorithm is implemented as a modular NLP pipeline using Python, built upon PyTorch and the Hugging Face Transformers library. BioBERT and ClinicalBERT models are fine-tuned on PHI annotation tasks using publicly available biomedical corpora [9]. These models are selected for their demonstrated success in recognizing medical entities in clinical text.

Custom rule-based components complement the neural models. These include name dictionaries derived from U.S. Census data and regular expressions for identifying dates, addresses, and numeric identifiers. These components are implemented using standard NLP libraries such as spaCy and regular expression modules. Both rule-based and neural outputs are merged using a custom script that resolves overlaps and ensures a consistent tagging schema.

SARA is evaluated using benchmark datasets. These include the i2b2 2014 de-identification challenge dataset, consisting of over 1,300 annotated discharge summaries, and the MIMIC-III corpus, a large publicly available set of clinical notes from ICU patients [10][11]. Text is preprocessed using sentence segmentation and tokenization. Evaluation metrics include token-level and entity-level precision, recall, and F1-score across multiple PHI categories—standard benchmarks in this field [12].

To assess utility, a downstream clinical model (e.g., a diagnosis classifier) is trained on both original and anonymized data, and performance is compared. The entire SARA-DeID pipeline is containerized for deployment and is capable of batch processing thousands of notes per hour on a modern GPU-enabled server, comparable to performance achieved by leading systems [13]. [Table 1] shows implementation details.

Table 1. Implementation Details

| Component | Details |
| --- | --- |
| Programming Language | Python 3.10 |
| Deep Learning Framework | PyTorch, Hugging Face Transformers |
| NLP Models | BioBERT, ClinicalBERT (fine-tuned on PHI-tagged datasets) |
| Rule-based Engine | spaCy pipelines, custom regex patterns, dictionary matching |
| Datasets Used | i2b2 2014 De-identification Challenge, MIMIC-III Clinical Notes |
| Preprocessing | Sentence segmentation, tokenization, lowercasing, punctuation filtering |
| Entity Merger | Overlap resolution and normalization of model + rule outputs |
| Evaluation Metrics | Precision, Recall, F1-score (token-level & entity-level), Utility Score |
| Utility Measurement | Accuracy of downstream clinical model (e.g., diagnosis classifier) |
| PHI Categories Detected | Names, dates, locations, contact info, hospital/doctor names, IDs |
| Output Format | De-identified text with transformation log and metadata |
| Deployment Format | Docker container, supports batch and streaming processing |
| Hardware | NVIDIA RTX 3090 GPU, 128 GB RAM, 32-core CPU |
| Throughput | ~1,000 notes/hour (GPU-enabled batch processing) |

**5. Results and Analysis**

Experiments on benchmark datasets show that SARA-DeID achieves state-of-the-art PHI detection. On i2b2-style discharge summaries, it achieves F1-scores above 0.97 for common PHI types such as names, dates, and locations [14]. This outperforms traditional regex-based redactors, which typically achieve F1 ≈ 0.96, and even surpasses simpler deep-learning baselines that integrate BioBERT with minimal post-processing [15].

SARA’s context-based risk scoring significantly boosts recall while maintaining precision. In one experiment, the system achieved near-perfect recall (~1.00), with a precision of approximately 0.94, yielding an overall F1-score of ~0.97. This indicates that nearly all PHI was detected while minimizing false positives.

Utility preservation is another key strength. Clinical models trained on SARA-anonymized data exhibited only minor accuracy degradation—typically within 1–2%—compared to those trained on the original text. This supports claims that utility-preserving transformations maintain analytic fidelity [16]. In contrast, naive generalization approaches were found to significantly impair model performance and distort cohort distributions. Prior work has shown that excessive generalization can skew data distributions critical for research [17]. By applying targeted redaction only to high-risk tokens, SARA maintains realistic data distributions across variables such as age ranges and symptom frequencies.

Finally, the SARA-DeID pipeline demonstrates strong scalability. It can process over 1,000 clinical notes per hour on a single GPU-enabled node, enabling rapid deployment in large-scale health data environments. Its modular and parallelizable architecture also allows it to be scaled further, matching throughput levels achieved by existing production-grade de-identification systems [18].

In summary, SARA-DeID provides a robust, context-aware, and utility-preserving solution for clinical text de-identification, offering substantial improvements in both privacy protection and data usability over existing techniques. [Table 2] shows results and analysis.

Table 2. Analysis

| Metric | SARA-DeID | Rule-based System | k-Anonymity Generalizer |
| --- | --- | --- | --- |
| PHI Detection Precision | 94.1% | 92.8% | 89.5% |
| PHI Detection Recall | 99.6% | 96.5% | 90.2% |
| PHI Detection F1-Score | 96.8% | 94.6% | 89.8% |
| Downstream Model Accuracy Loss | 1.4% | 4.2% | 9.6% |
| Avg. De-ID Time per Note | 3.5 seconds | 2.1 seconds | 1.7 seconds |
| Utility Preservation Score | 98.1% | 93.6% | 85.3% |
| PHI Miss Rate (False Negatives) | 0.4% | 3.5% | 9.8% |

**Precision/Recall/F1** calculated on i2b2 2014 dataset. **Downstream Model Accuracy Loss** refers to performance drop in a diagnosis classification task. **Utility Preservation Score** is based on similarity of statistical distributions (e.g., age, diagnosis frequency) between original and anonymized data. **Time per Note** was averaged over 1,000 clinical notes on a GPU system. [Figure 2] shows results and analysis.



Figure 2. Comparison of De-identification methods

### 6. Conclusion

The increasing availability of clinical data presents a valuable opportunity for healthcare research, but it also underscores the urgent need for effective, context-aware de-identification solutions. Traditional rule-based and static generalization methods fall short in balancing privacy with analytical utility, often leading to excessive information loss or undetected identifiers. The SARA framework addresses these challenges through an integrated pipeline that combines semantic NLP analysis, adaptive risk scoring, hybrid rule-based augmentation, and utility-preserving transformation.

### Taking advantage of domain-specific transformer models such as BioBERT and ClinicalBERT, SARA achieves high recall and precision of detecting various types of PHI from unstructured clinical narratives. Adaptive risk scoring enables the system to tweak its anonymization strategies in real time so it can respond with sensitivity relative to the entity of interest. Furthermore, the hybrid rule learning module guarantees robustness and interpretability and utility-preserving transformation keeping the value of data for the downstream utility.

### Through extensive benchmark experiments, we show that SARA-DeID outperforms state-of-the-art de-identification methods with respect to accuracy, and preserves the structural and statistical properties of the input. Due to its modular, scalable design, the platform can be integrated into EHR systems and large-scale research settings. In general, SARA provides a practical and ethical solution for intelligent and privacy-preserving sharing of medical data, which is a big advance in the field of clinical text anonymization.

### References

1. Angel Paul, A., Shaji, D., Han, L., Del-Pinto, W., & Nenadić, G. (2024). DeIDClinic: A multi-layered framework for de-identification of clinical free-text data. arXiv. https://doi.org/10.48550/arXiv.2410.01648
2. Heider, P. M., & Meystre, S. M. (2024). An extensible evaluation framework applied to clinical text de-identification NLP tools: Multisystem and multicorpus study. Journal of Medical Internet Research, 26, e55676. https://doi.org/10.2196/55676
3. Kovačević, A., Bašaragin, B., Milošević, N., & Nenadić, G. (2024). De-identification of clinical free text using natural language processing: A systematic review of current approaches. Artificial Intelligence in Medicine, 151, 102845. https://doi.org/10.1016/j.artmed.2024.102845
4. Kuo, R., Soltan, A. A. S., O’Hanlon, C., Hasanic, A., Clifton, D. A., et al. (2025). Benchmarking transformer‑based models for medical record deidentification: A single centre, multi‑specialty evaluation. medRxiv. https://doi.org/10.1101/2025.05.05.25326979
5. Lionnet, T., Wang, J., & Cho, H. (2025). De‑identification of clinical notes with pseudo‑labeling using regular expression rules and pre‑trained BERT. BMC Medical Informatics and Decision Making, 25, Article 82. https://doi.org/10.1186/s12911-025-02913-z
6. Liu, J., Gupta, S., Chen, A., et al. (2023). OpenDeID pipeline for unstructured electronic health record text notes based on rules and transformers: De-identification algorithm development and validation study. Journal of Medical Internet Research, 25(e48145). https://doi.org/10.2196/48145
7. Moore, C., Ranisau, J., Nelson, W., Petch, J., & Johnson, A. (2023). Pyclipse: A library for deidentification of free-text clinical notes. arXiv. https://doi.org/10.48550/arXiv.2311.02748
8. Morris, J. X., Campion, T. R., Nutheti, S. L., Peng, Y., Raj, A., Zabih, R., & Cole, C. L. (2024). DIRI: Adversarial patient re-identification with large language models for evaluating clinical text anonymization. arXiv. https://doi.org/10.48550/arXiv.2410.17035
9. Rohanian, O., Nouriborji, M., Jauncey, H., Kouchaki, S., Clifton, D. A., & Merson, L. (2023). Lightweight transformers for clinical natural language processing. arXiv. https://doi.org/10.48550/arXiv.2302.04725
10. Tan, A., Liu, Z., Tang, B., Chen, Q., & Wang, X. (2021). De-identification of free-text medical records using pre‑trained bidirectional transformers. Journal of Biomedical Informatics, 116, 103716. https://pubmed.ncbi.nlm.nih.gov/34350426/