Literature Review & Report
NLP for Healthcare (CL3.411)

Aryaman Bahl, 2022113010

Deadline: 22 August 2025, 11:59 PM

1 Introduction

De-identification protects patient privacy by removing personal identifiers from healthcare data. Under HIPAA, this requires eliminating 18 identifiers such as names, contact details, and medical record numbers. Clinical free text is especially challenging due to irregular grammar, abbreviations, and domain-specific codes. This report reviews rule-based, machine learning, and model-driven methods, summarizes performance of domain-specific architectures, and proposes hybrid workflows for robust de-identification.

2 Basic Approaches

Rule-based methods rely on handcrafted patterns, dictionaries, and contextual rules to locate protected health information (PHI). They are interpretable, efficient, and effective for structured identifiers (e.g., dates, phone numbers, emails).

Table 1: Basic approaches to medical document de-identification.

Technique	Description & Examples
Regex Pattern Matching	Matches fixed patterns such as dates, phone numbers, or emails. Examples: `\b\d{4}[-/.]\d{2}[-/.]\d{2}\b` (dates), `\b[\w.+-]+@[\w.-]+\.[A-Za-z]{2,}\b` (emails).
Dictionaries Gazetteers	Precompiled lists of entities (e.g., patient names, hospitals, cities) used to flag known PHI.
Contextual Rules	Trigger words or section headers that indicate sensitive information. Examples: "SSN:", "Patient Contact Information".
Redaction Pseudonymization	Detected PHI is either masked (e.g., [NAME]) or replaced with consistent pseudonyms. Example: "John Smith" → "Patient123".

Strengths: Transparent, auditable, deterministic. **Limitations:** Brittle to noise, abbreviations, and unseen formats.

3 ML and Hybrid Methods

De-identification can be framed as sequence labeling: tagging each token as PHI or non-PHI.

Classical ML

Pre-deep learning, models such as Conditional Random Fields (CRFs) relied on hand-crafted features (capitalization, affixes, dictionary presence). Effective with small data but limited in generalization.

Deep Learning

Transformer-based models (BERT, ClinicalBERT, BioBERT) learn contextual embeddings directly from data, significantly improving accuracy on PHI recognition. Fine-tuning on clinical datasets yields high recall and precision.

Hybrid Pipelines

Most production systems combine rule-based precision with ML recall. A typical workflow:

1. Regex/dictionaries for unambiguous PHI (dates, phone numbers).
2. Clinical NER models for contextual PHI (names, locations).
3. Post-processing and redaction with placeholders.

4 Model-based Methods and Performance

The best-performing models are neural architectures fine-tuned for clinical named entity recognition.

Key Architectures

- **BiLSTM-CRF:** Early state-of-the-art sequence tagger.
- **Transformers (BERT variants):** ClinicalBERT and BioBERT outperform generic BERT by pre-training on medical corpora.

Performance

On the i2b2/UTHealth 2014 benchmark, BioBERT achieves the highest F1 (**93.0%**). Domain-specific pre-training consistently improves results over general models.

Table 2: Accuracy (MedNLI) and Exact F1 score (i2b2) across various clinical NLP tasks.

Model	MedNLI	i2b2 2010	i2b2 2012	i2b2 2014
BERT	77.6%	83.5	75.9	92.8
BioBERT	80.8%	86.5	78.9	**93.0**
Clinical BERT	80.8%	86.4	78.5	92.6
Discharge Summary BERT	80.6%	86.4	78.4	92.8
Bio+Clinical BERT	**82.7%**	87.2	**78.9**	92.5
Bio+Discharge Summary BERT	**82.7%**	**87.8**	78.9	92.7

5 Proposed Pipelines

Robust Hybrid Workflow

1. Preprocessing and text normalization.
2. Regex + gazetteers for structured PHI.
3. Clinical NER (e.g., BioBERT) for contextual PHI.
4. Merge predictions with consensus logic.
5. Redaction or pseudonymization with consistent labels.

Generate-and-Verify Variant

A large language model proposes candidate PHI spans (high recall). Verification by ClinicalBERT and regex filters hallucinations (high precision).

Lightweight Baseline

In resource-limited settings, rules and gazetteers alone (optionally with CRFs) provide a transparent, interpretable baseline.

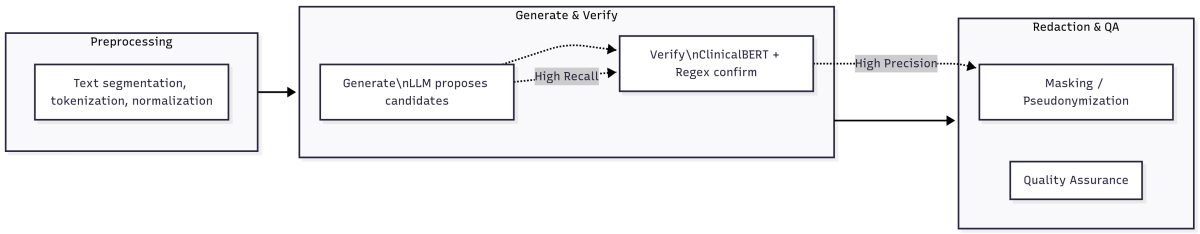

Figure 1: Generate-and-Verify (G&V) pipeline for clinical de-identification. An LLM generates high-recall candidate spans, which are then verified by ClinicalBERT and regex for high precision, followed by redaction and QA.

6 Data Strategy

Since annotated PHI datasets are scarce, synthetic data can be created:

- **Template-based substitution:** Inject synthetic PHI (names, hospitals, addresses) into de-identified text.
- **LLM-driven narrative generation:** Create diverse clinical notes with embedded placeholders.

7 Conclusion

De-identification requires a balance between precision and recall. Rule-based approaches excel for structured PHI, while deep learning models capture context. Hybrid workflows, especially generate-and-verify pipelines, offer a practical path forward. Synthetic data helps overcome annotation bottlenecks, and consistent governance ensures compliance.

References

1. U.S. Department of Health and Human Services. (2003). HIPAA Privacy Rule: Safe Harbor (18 identifiers). 45 CFR §164.514(b)(2). Retrieved from `https://www.ecfr.gov/current/title-45/subtitle-A/subchapter-C/part-164/subpart-E/section-164.514`

2. Neamatullah, I., et al. (2008). Automated de-identification of free-text medical records. *BMC Medical Informatics and Decision Making*, 8, 32. `https://doi.org/10.1186/1472-6947-8-32`

3. Dernoncourt, F., et al. (2017). De-identification of patient notes with recurrent neural networks. *JAMIA*, 24(3), 596–606. `https://doi.org/10.1093/jamia/ocw156`

4. Lee, J., et al. (2020). BioBERT: a pre-trained biomedical language representation model. *Bioinformatics*, 36(4), 1234–1240. `https://doi.org/10.1093/bioinformatics/btz682`

5. Alsentzer, E., et al. (2019). Publicly available clinical BERT embeddings. *arXiv:1904.03323*. `https://arxiv.org/abs/1904.03323`

6. Microsoft Presidio Documentation. Retrieved from `https://microsoft.github.io/presidio/`

7. John Snow Labs. Healthcare de-identification with Spark NLP. Retrieved from `https://nlp.johnsnowlabs.com/`