

Open Source PHI De-Identification: A Technical Review

By Adrien Laurent, CEO at IntuitionLabs • 11/14/2025 • 35 min read

- phi de-identification
- hipaa compliance
- data anonymization
- clinical nlp
- open source software
- named entity recognition
- data scrubbing
- healthcare data



Executive Summary

Open-source tools for scrubbing and de-identifying protected health information (PHI) in clinical data have grown markedly in recent years, due to the twin pressures of privacy regulations (like HIPAA and GDPR) and the research imperative to share clinical text data. In general, these tools fall into rule-based, machine-learning (ML)/statistical, or hybrid categories, often augmented by modern deep learning (e.g. BERT) and even new **large language model (LLM) techniques**. Notable open-source projects include the PhysioNet *deid* toolkit (GPL-licensed) ⁽¹¹⁾ physionet.org, the BSD-licensed *Philter* tool from UCSF ⁽²⁾ pmc.ncbi.nlm.nih.gov, the Python *PyDeID* system ⁽³⁾ academic.oup.com, MITRE's BSD-licensed *MIST* toolkit ⁽⁴⁾ mist-deid.sourceforge.net, Microsoft's Apache-licensed *Presidio* framework ⁽⁵⁾ github.com, Stanford's *TIDE* (text de-identification engine) ⁽⁶⁾ academic.oup.com, the GPL-licensed *CliniDeID* system ⁽⁷⁾ clinacuity.com, and the FOSS *CRATE* system for database anonymization ⁽⁸⁾ bmcmmedinformdecismak.biomedcentral.com. These tools employ combinations of regular expressions, dictionaries, named-entity recognition (NER) models, and context-sensitive heuristics to identify PHI categories (names, dates, locations, IDs, etc.) and remove or replace them. Evaluations show high recall (often >95%) on benchmark corpora, though precision can be more modest; for example, *Philter* achieved 99.5% recall (at ~78% precision) on a UCSF test set ⁽²⁾ pmc.ncbi.nlm.nih.gov while the PhysioNet *deid* tool achieved ~94% recall (75% precision) on nursing notes ⁽⁹⁾ bmcmmedinformdecismak.biomedcentral.com. Hybrid approaches (e.g. *PyDeID* combining regex and NER ⁽³⁾ academic.oup.com, or *CliniDeID* ensembles ⁽⁷⁾ pubmed.ncbi.nlm.nih.gov) can balance precision and recall. In practice, these open solutions have been deployed at scale (e.g. UCSF processed >70M notes using *Philter* ⁽²⁾ pmc.ncbi.nlm.nih.gov), Stanford uses *TIDE* with surrogate replacement ⁽⁶⁾ academic.oup.com, and the MIMIC II/III projects use PhysioNet *deid* with date-shifting and realistic surrogates ⁽¹¹⁾ physionet.org). Going forward, open-source de-identification is advancing toward LLM-assisted methods (though recent studies caution that general-purpose LLM tools miss >50% of clinical PHI ⁽¹²⁾ www.johnsnowlabs.com), integration with standardized data models (OMOP *FHIR*), and stronger privacy frameworks (e.g. differential privacy). The growth of open-source PHI-scrubbing software represents both a technical opportunity and a critical support for compliant data sharing: researchers and health systems can leverage these community-vetted tools (with transparent code) to protect patient privacy while enabling analytics and **AI on clinical narratives**.

Introduction and Background

Protected health information (PHI) – data elements that can identify an individual in a medical record (per HIPAA's 18 identifiers) – must be removed or masked before medical text can be broadly shared ⁽¹³⁾ pmc.ncbi.nlm.nih.gov ⁽¹⁴⁾ pmc.ncbi.nlm.nih.gov. Manual redaction is labor-intensive and error-prone, particularly for large unstructured datasets ⁽¹⁵⁾ bmcmmedinformdecismak.biomedcentral.com ⁽¹⁶⁾ pmc.ncbi.nlm.nih.gov. Thus, automated de-identification ("scrubbing") has been an active research area for two decades. Early systems (e.g. PrivateEye, Datafly, MITRE's *MIST* and others) relied on hand-crafted rules and templates ⁽¹⁷⁾ pmc.ncbi.nlm.nih.gov ⁽⁴⁾ mist-deid.sourceforge.net). The 2006 and 2014 i2b2 NLP challenges and the 2016 CEGS N-GRID challenge spurred development of more sophisticated approaches, showing that hybrid methods using lexicons, regex patterns, and machine learning models could achieve high sensitivity ⁽¹⁸⁾ pmc.ncbi.nlm.nih.gov ⁽¹⁹⁾ pmc.ncbi.nlm.nih.gov. In 2008 Neamatullah *et al.* released the PhysioNet *deid* tool (GPL v2) for the MIMIC II ICU notes, employing lookup tables, regexes, and realistic surrogate replacements ⁽⁹⁾ bmcmmedinformdecismak.biomedcentral.com ⁽²⁰⁾ physionet.org. In parallel, national guidelines (HIPAA in the US, GDPR in the EU) defined de-identification standards. Under HIPAA "Safe Harbor," removal of all 18 identity fields (names, geographic subdivisions below state, dates, SSN, medical record numbers, etc.) yields de-identified data (no longer PHI) ⁽¹³⁾ pmc.ncbi.nlm.nih.gov ⁽¹⁶⁾ pmc.ncbi.nlm.nih.gov. However, HIPAA also allows an "expert determination" approach where an expert certifies the re-identification risk is very small. Many open-source scrubbing tools adopt the Safe Harbor checklist as targets for removal.

However, medical text is complex: identifiers can appear in variable formats, embedded in free text, and mixed with clinical terms. Consequently, modern de-identifiers often combine multiple techniques ⁽²¹⁾ pmc.ncbi.nlm.nih.gov ⁽²²⁾ pmc.ncbi.nlm.nih.gov. Rule-based patterns (dates, SSN, phone numbers, common name lists) excel at obvious PHI, while statistical models (conditional random fields, neural NER models) can catch less-structured PHI. Recently, deep learning – notably pretrained transformer models (BERT and its derivatives) – has been applied to this task. Johnson *et al.* (2020) fine-tuned BERT for PHI detection, reporting state-of-the-art recall on multiple clinical note corpora and publishing their code ⁽²³⁾ pmc.ncbi.nlm.nih.gov.

The **open-source movement** in health IT emphasizes transparency and reusability. Many de-identification projects have released code under permissive licenses to encourage adoption and collaboration. Key open tools and libraries now available – in Python, Java, R, etc. – implement PHI scrubbing for clinical text. </current_article_content>This survey reviews these open-source solutions, examining their methods, performance, usage examples, and how they fit within regulatory and ethical frameworks. We draw on academic studies, technical reports, software documentation, and user reports to deliver a deep, evidence-based analysis of the current state and future of open-source PHI de-identification.

Regulatory and Ethical Context

Healthcare data sharing is governed by strict privacy laws. In the U.S., HIPAA's Privacy Rule defines PHI and outlines methods for de-identification. Per HIPAA, a dataset is "de-identified" when either (a) an expert certifies low re-identification risk, or (b) the Safe Harbor rule is followed – i.e. all 18 types of identifiers are removed or masked ⁽¹³⁾ pmc.ncbi.nlm.nih.gov ⁽²⁴⁾ pmc.ncbi.nlm.nih.gov. (The identifiers include names, small geographical units, all specific dates except year, social security numbers, medical record numbers, contact numbers, and so on ⁽¹³⁾ pmc.ncbi.nlm.nih.gov ⁽¹⁶⁾ pmc.ncbi.nlm.nih.gov.) Under Safe Harbor, data is no longer considered PHI and may be freely shared. De-identification techniques thus must reliably eliminate instances of these fields.

Similarly, the EU's GDPR and many other jurisdictions require anonymizing health data before analysis or sharing, albeit with somewhat different definitions. Under GDPR, truly anonymous data (irreversible de-identification) falls outside the regulation, while pseudonymized data still counts as personal. Thus, many U.S. tools implement pseudonymization (consistently replacing identifiers with pseudonyms or surrogates) to allow re-linking records when needed, whereas full anonymization drops patient identifiers entirely.

From an ethics standpoint, patients generally support research on de-identified data for the common good ([25] pmc.ncbi.nlm.nih.gov). However, any PHI leakage could be costly; thus, high recall (sensitivity) is often prioritized even at the expense of some over-redaction. The trade-off is delicate: too aggressive removal (low precision) can damage data utility, whereas missed PHI (low recall) poses privacy risk ([26] pmc.ncbi.nlm.nih.gov) ([27] pmc.ncbi.nlm.nih.gov). This has led experts to advocate hybrid and ensemble methods to balance precision/recall ([28] pmc.ncbi.nlm.nih.gov) ([12] www.johnsnowlabs.com).

In sum, any production system must aim for maximal PHI removal consistent with regulatory rules. The open-source tools we examine typically target HIPAA Safe Harbor categories and aim for recall > 95% on representative corpora ([2] pmc.ncbi.nlm.nih.gov) ([19] pmc.ncbi.nlm.nih.gov). Beyond raw performance, we must consider ease of integration, configurability, and transparency of algorithms, since mistakes or hidden behavior in black-box systems could violate policy.

Categories of De-identification Methods

Rule-Based and Pattern Matching Approaches

Rule-based de-identification relies on hand-crafted patterns, dictionaries, and regular expressions to locate PHI. For example, dates can be found with regexes (MM/DD/YYYY, etc.), phone numbers via known numeric formats, and names via dictionary lookups. Such systems may also use contextual cues ("Dr." or "Mr.") to detect names. The advantage is simplicity and interpretability ([29] pmc.ncbi.nlm.nih.gov) ([9] bmcmmedinformdecismak.biomedcentral.com). The classic PhysioNet *deid* software is predominantly rule-based: it uses lookup tables (names, places) plus ~350 regex rules for addresses, dates, and numeric identifiers ([30] pmc.ncbi.nlm.nih.gov) ([9] bmcmmedinformdecismak.biomedcentral.com). Early studies showed this captures ~94–99% of PHI in pathology notes ([31] pmc.ncbi.nlm.nih.gov) ([9] bmcmmedinformdecismak.biomedcentral.com).

However, pure rule-based systems struggle with variability (misspelled names, unknown institutions) and "out-of-vocabulary" entities (rare names, novel IDs). As the 2023 scoping review notes, rule-based methods do not require labeled training data and work well for predictable patterns, but designing comprehensive rules is time-consuming and brittle ([32] pmc.ncbi.nlm.nih.gov). Therefore, modern scrubbers often supplement patterns with statistical models. Nonetheless, even current tools like *PyDeID* allow user-defined regex patterns for maximum coverage and let experts tailor rules for a given dataset ([33] academic.oup.com).

Machine Learning and Statistical Models

Supervised machine learning (ML) approaches treat PHI detection as a named-entity recognition (NER) problem. Classical ML systems (conditional random fields, decision trees, support vector machines, etc.) use handcrafted features (word shapes, gazetteer membership, surrounding context) to classify tokens as PHI or not ([18] pmc.ncbi.nlm.nih.gov) ([34] pmc.ncbi.nlm.nih.gov). The i2b2 2006 and CEGS N-GRID 2016 de-id challenges demonstrated that CRF-based systems often outperform pure rules, especially when multiple context features are combined ([18] pmc.ncbi.nlm.nih.gov) ([35] pmc.ncbi.nlm.nih.gov). For instance, participants in i2b2 2006 used rule templates and decision trees with local text features to achieve high F1 scores ([18] pmc.ncbi.nlm.nih.gov). The scoping review reports that many systems use ML components to catch irregular PHI (like MRNs with odd formats) ([29] pmc.ncbi.nlm.nih.gov).

Deep learning (especially contextualized neural networks) has dramatically improved NER. Clinical BERT or BiLSTM-CRF architectures fine-tuned on annotated corpora can learn richer representations of medical context. Johnson *et al.* (2020) fine-tuned a BERT model on multiple clinical note sets, achieving new state-of-the-art performance ([19] pmc.ncbi.nlm.nih.gov). They emphasize that adding embedding layers (from models like BERT or GloVe) and CRF decoding can generalize beyond simple rules ([35] pmc.ncbi.nlm.nih.gov) ([36] pmc.ncbi.nlm.nih.gov). Liu *et al.* (2018) showed that ensembles of RNN+CRF with supplementary rules outperform single systems ([37] pmc.ncbi.nlm.nih.gov). In practice, open tools like *CLISI* (not widely released) and *CliniDeID* use deep learning components; *CliniDeID* is described as an "ensemble combining deep and shallow ML with rule-based algorithms" achieving high recall and precision on several corpora ([10] pubmed.ncbi.nlm.nih.gov). The *PyDeID* study also includes an NER configuration (using spaCy) which improved recall (up to 95%) at some cost to precision ([33] academic.oup.com) ([38] academic.oup.com).

Hybrid and Ensemble Methods

Given the complementary strengths of rules and ML, hybrid systems often perform best ([29] pmc.ncbi.nlm.nih.gov) ([39] pmc.ncbi.nlm.nih.gov). They might first apply high-confidence rules (e.g. well-formed dates), then pass remaining text to an ML model. Some tools, like *TiDE*, explicitly combine approaches: *TiDE* uses pattern matching and known-PHI lists as pre-processing before applying Stanford CoreNLP's NER ([40] github.com) ([6] academic.oup.com). The Stanford pipeline then replaces flagged items with "hide-in-plain-sight" surrogates, ensuring that any missed PHI is visually camouflaged by similar fake names ([6] academic.oup.com).

Ensemble learning (combining multiple models) is another hybrid tactic. *CliniDeID* and other ensemble frameworks run several detectors in parallel, then vote or cascade to maximize sensitivity ([10] pubmed.ncbi.nlm.nih.gov). This can capture PHI missed by any one model. Another example is the open-source *PRIFinder* (Illinois), which ensembles multiple NER taggers (though it is not fully open source). The end result is that most cutting-edge open de-identifiers are "multi-strategy" systems.

Emerging Large Language Model (LLM) Methods

Recently, general-purpose LLMs (e.g. GPT-family models) have been proposed for PII/PHI detection. Some open projects leverage LLMs for zero-shot redaction (e.g. OpenPipe's PII-Redact with Llama, or GLiNER using a BERT-like transformer) ⁽⁴¹⁾ www.johnsnowlabs.com). These tools boast very high accuracy on general text: in one benchmark, GLiNER and OpenPipe achieved macro F1 = 0.62 and 0.98 respectively on a mixed-domain civilian dataset ⁽⁴¹⁾ www.johnsnowlabs.com). However, on actual clinical text their performance plummets: both scored only ~0.41 F1, missing over 50% of PHI ⁽⁴¹⁾ www.johnsnowlabs.com). The John Snow Labs analysis points out that domain-specificity is crucial – models not trained on medical corpora struggle to recognize medical context. Thus, while LLM-based tools are promising for flexible text, the current open LLM de-identifiers are not yet reliable for strict PHI scrubbing without additional tuning. Private healthcare organizations sometimes use custom LLMs trained on medical notes, but open-source alternatives lag behind domain-specific systems ⁽⁴¹⁾ www.johnsnowlabs.com). Future LLM methods (or fine-tuned clinical LLMs) may reduce this gap, but for now rule/ML hybrids remain dominant for open PHI de-id.

Structured Data and Tabular Anonymization

Not all PHI lives in text. Structured health data (databases of vitals, labs, etc.) may contain direct identifiers or quasi-identifiers. Open-source tools like ARX (Java, GPL) and sdcMicro (R, open) provide general data anonymization (k-anonymity, l-diversity, differential privacy, etc.) for tabular data ⁽⁴²⁾ guides.library.jhu.edu ⁽⁴²⁾ guides.library.jhu.edu). While these tools are often used for statistical disclosure control (SDC), they can remove PHI fields by suppression and perform techniques like generalization. For example, ARX can generalize ages, flags outliers, or bucket-aggregate values to reduce re-identification risk ⁽⁴²⁾ guides.library.jhu.edu). Though ARX isn't tailored to HIPAA's 18 identifiers specifically, it is widely used for de-identifying structured clinical datasets. Similarly, *SDC Micro* and its Shiny GUI are used by researchers to anonymize spreadsheets of patient records ⁽⁴²⁾ guides.library.jhu.edu).

Image and Multimedia De-Identification

Health data also includes images (e.g. X-rays, MRIs) which may contain embedded PHI. Open-source image scrubbers exist primarily as DICOM metadata tools or *defacing* algorithms. *DICOMCleaner* (open-source GUI) removes identifying tags from DICOM headers ⁽⁴²⁾ guides.library.jhu.edu). For MRI de-identification, tools like *MRI_deface* and *PyDeface* (Python) remove facial structure from brain scans ⁽⁴²⁾ guides.library.jhu.edu). Newer AI-based defacers (e.g. *DeepDefacer* ⁽⁴²⁾ guides.library.jhu.edu) using U-Net) automatically erase faces. Slowly, these tools are being integrated into imaging pipelines. For example, Stanford's pipeline uses the RSNA's CTP engine with custom filters to scrub DICOM data on-demand ⁽⁴³⁾ academic.oup.com). While not explicitly "PHI scrubbing" in text, such image anonymizers are part of the broader data privacy toolkit.

Key Open-Source PHI De-Identification Tools

The following sections highlight several prominent open-source PHI-scrubbing solutions. Each tool is briefly described with its methods, data focus, license, and known performance or use cases. Where possible, we cite peer-reviewed evaluations or documentation.

Text De-Identification Tools

- **PhysioNet *deid*** (Neamatullah *et al.*, 2008) ⁽⁹⁾ bmcmedinformdecismak.biomedcentral.com – A pioneering GPL-licensed Perl toolkit developed for ICU nursing notes. It uses lexical dictionaries (names, locations) and ~50 regular expressions (dates, IDs, contacts) to identify PHI, then replaces PHI with realistic surrogate values (e.g. random dates, fictitious names) ⁽⁹⁾ bmcmedinformdecismak.biomedcentral.com ⁽⁴⁴⁾ physionet.org). *Deid* achieved ~94% recall on held-out nursing notes, outperforming single human annotators ⁽⁹⁾ bmcmedinformdecismak.biomedcentral.com. It is available on PhysioNet with source code (GNU GPL v2) ⁽¹⁾ physionet.org. Many later tools (including PyDeID and others) were inspired by or built atop the PhysioNet *deid* logic.
- **Philter (Protected Health Information Filter)** (Norgeot *et al.*, 2020) ⁽²⁴⁾ pmc.ncbi.nlm.nih.gov – An open-source Python package (BSD-3 license) developed at UCSF. Philter combines regex patterns, statistical language models, and "blacklists/whitelists" to flag PHI in clinical notes ⁽⁴⁵⁾ academic.oup.com ⁽⁴⁶⁾ pmc.ncbi.nlm.nih.gov). Its design prioritized extremely high recall: on a 2,000-note UCSF corpus Philter achieved 99.46% recall (with an F2-score of 94.36) – significantly higher than prior systems ⁽⁴⁷⁾ pmc.ncbi.nlm.nih.gov. Precision was modest (~78%), reflecting its conservative removal. Philter's code is available on GitHub ⁽⁴⁸⁾ github.com and it can output either tagged placeholders or surrogate values. In independent comparisons Philter's recall dwarfed older rule-based tools, making it suitable for settings where missing any PHI is unacceptable ⁽²⁾ pmc.ncbi.nlm.nih.gov ⁽²⁷⁾ pmc.ncbi.nlm.nih.gov.
- **PyDeID** (Sundrelingam *et al.*, 2025) ⁽³³⁾ academic.oup.com – An updated open-source tool (Python, license LGPL or similar) based on PhysioNet *deid*. PyDeID aims to be faster and more flexible. It supports multiple configurations: pure regex (like *deid*), regex with NER (spaCy), and optional custom name lists. In tests on 700 Canadian hospital notes, its base config achieved recall ~90.6% and precision 88.9% (F1~87.9%) ⁽³⁸⁾ academic.oup.com ⁽⁴⁹⁾ academic.oup.com). This outperformed the original *deid* (89.4% recall, 79.8% precision) and was much faster (0.48 sec/note vs 6.38 sec) ⁽³⁸⁾ academic.oup.com ⁽⁴⁹⁾ academic.oup.com. PyDeID also includes surrogate replacement and supports custom rules. Its source code is on GitHub under an open license (<https://github.com/takewalks/pyDeid>). Its release provides a modern, easily configurable "rule+NER" alternative with competitive accuracy ⁽³³⁾ academic.oup.com.
- **NLM Scrubber** (Kronberger *et al.*, 2013) ⁽²⁷⁾ pmc.ncbi.nlm.nih.gov – A command-line tool from the U.S. National Library of Medicine (CIVIC) for clinical text. It leverages Apache cTAKES and UIMA to statistically classify tokens as PHI based on word frequency differences between public corpora and private notes ⁽²⁷⁾ pmc.ncbi.nlm.nih.gov. NLM Scrubber emphasizes recall: it considers a word PHI if it rarely appears in public medical documents ⁽⁵⁰⁾ pmc.ncbi.nlm.nih.gov. It is freely downloadable but the source code is not openly published. Reports (via UCSF evaluations) indicate ~95% recall on i2b2 corpora ⁽⁵¹⁾ pmc.ncbi.nlm.nih.gov, at the cost of ~79% precision ⁽⁵²⁾ pmc.ncbi.nlm.nih.gov. (Because of this, some open-source efforts focus on improving precision or retraining such models.)
- **MIST (MITRE Identification Scrubber Toolkit)** – An open-source (BSD-licensed) toolkit by MITRE for medico-legal text. It finds PII/PHI using pattern matching and lexicons, and replaces found identifiers with either generic markers (e.g. "[NAME]") or realistic synthetic values ⁽⁴⁾ mist-deid.sourceforge.net). For instance, MIST can take a sentence with "Mary Phillips" and output "[NAME]" or a fictitious name. Its goal is to produce still-readable text with clear placeholders. MIST is mature (last updated ~2013) and available on SourceForge. While formal evaluations are scarce, its use of synthetic surrogates and placeholders is a common strategy. MIST's presence underscores that open, user-friendly PII scrubbing has been a research goal for over a decade ⁽⁴⁾ mist-deid.sourceforge.net ⁽⁵³⁾ mist-deid.sourceforge.net.

- Stanford TiDE (Text De-Identification Engine)** – A free open-source Java-based framework from Stanford Medicine (<https://github.com/susom/tide>). TiDE uses rule-based pattern matching, known-PHI lists (e.g. patient/clinician names from the local database), and Stanford CoreNLP NER to tag identifiers ⁽⁴⁰⁾ [github.com](#)). It implements a “hide in plain sight” strategy: detected names and locations are replaced with random surrogates of the same type, so that any missed real PHI blends into the text ⁽⁶⁾ [academic.oup.com](#)). For example, real and fake names are indistinguishable in context. TiDE reports high overall accuracy on admission/discharge notes (unspecified), and is configurable via XML rules. It's cited within Stanford's data science infrastructure, where it forms part of the de-ID pipeline for clinical text ⁽⁶⁾ [academic.oup.com](#)). TiDE's source is on GitHub under an open license, allowing institutions to adapt it for their EHR systems.
- CliniDeID** (Clinacuity, 2023) – A newer open-source system (GPL v3) that applies an ensemble of ML and rule-based methods to clinical text. It explicitly aims for “high accuracy” on Unstructured Electronic Health Records ⁽¹⁰⁾ [pubmed.ncbi.nlm.nih.gov](#)). According to the developers, CliniDeID achieved “high recall and precision” on test corpora (published results not yet widely cited). It supports both surrogate replacement and tagging, and integrates structured de-identification as well (OMOP and soon FHIR) ⁽⁷⁾ [clinacuity.com](#)). The vendor site highlights features like consistent surrogate use across a patient's record and multi-format I/O ⁽⁵⁴⁾ [clinacuity.com](#)). Key point: CliniDeID is freely available on GitHub (GPL3) and adds modern embedding/ensemble methods on top of traditional techniques ⁽⁷⁾ [clinacuity.com](#)).
- Apache cTAKES** – While cTAKES is primarily an NLP pipeline for extracting clinical concepts, it includes a PHI annotator component. It can be used in open-source workflows to tag names, dates, etc. (especially for unstructured text) and often underpins other tools ⁽²⁷⁾ [pmc.ncbi.nlm.nih.gov](#)). However, cTAKES itself is not a full de-identification system: it will annotate likely PHI but typically requires user configuration or downstream processing to actually remove text. It is Apache-licensed (open source) and widely used. Some projects (e.g. NLM Scrubber) leverage cTAKES modules.

Image De-Identification Tools

- DICOMCleaner** – A free open-source GUI tool for anonymizing DICOM medical images ⁽⁴²⁾ [guides.library.jhu.edu](#)). It strips identifying metadata from DICOM headers (patient name, ID, dates, etc.) and can overwrite burned-in annotations. It runs on Windows/Mac. As of now, DICOMCleaner is one of the few mature open tools specifically for image PHI.
- PyDeface / DeepDefacer / Quickshear** – Open Python tools to *deface* MRIs by removing facial features. PyDeface ⁽⁴²⁾ [guides.library.jhu.edu](#)) (Stanford, BSD-3 license) and DeepDefacer ⁽⁴²⁾ [guides.library.jhu.edu](#)) (U-Net model) automatically segment and blur face regions. Quickshear ⁽⁴²⁾ [guides.library.jhu.edu](#)) is a classic algorithmic defacer (Plos companion). These ensure MRI scans no longer reveal a patient's face, addressing PHI in image content. New pipelines (BIDS apps like “BIDSonym” ⁽⁴²⁾ [guides.library.jhu.edu](#)) integrate these tools for bulk processing of neuroimages in research.
- C2P (Clinical Trials Processor)** – The Radiological Society of North America's open CTP framework can scrub DICOM instances via custom filtering rules. Stanford's pipeline, for example, uses an enhanced CTP backend to de-identify images on-demand ⁽⁴³⁾ [academic.oup.com](#)). CTP itself (built on Java) is open source.

Though these tools do not touch textual PHI, they are part of the broader PHI-scrubbing ecosystem. We note them to illustrate that “de-identification” can span modalities; any comprehensive privacy solution may need both text and image components.

Databases and Integrated Systems

- CRATE (Clinical Records Anonymisation and Text Extraction)** (Jackson *et al.*, 2017) ⁽⁸⁾ [bmcmmedinformdecismak.biomedcentral.com](#)) – An end-to-end, open-source (Python/Java) system for creating anonymized research databases from clinical EMRs. It connects to a relational database (hospital EHR), systematically suppresses identifiable fields, and applies NLP to free-text fields (via GATE or MedEx) all within one pipeline ⁽⁸⁾ [bmcmmedinformdecismak.biomedcentral.com](#)). Clinically, CRATE has been used for large-scale psychiatric record anonymization (the UK CRIS system) and is feature-rich: it generates hashed patient pseudonyms, preserves record linkage (consistent surrogate IDs), and offers a research portal. In technical details, CRATE distinguishes itself by being entirely FOSS ⁽⁸⁾ [bmcmmedinformdecismak.biomedcentral.com](#)) and addressing structured data, cryptographic pseudonymization, and consent processes. Its text de-identification uses database-driven dictionaries (e.g. the source table has the patient name, which is then scrubbed from notes). In short, CRATE is a “one stop” free solution for institutions wanting to create an anonymized database.
- sdcmicro, ARX, and Other SDC Tools** – As mentioned, tools like sdcmicro (R package) and ARX (Java) specialize in statistical disclosure control for tabular data ⁽⁴²⁾ [guides.library.jhu.edu](#)) ⁽⁴²⁾ [guides.library.jhu.edu](#)). They are not clinical-specific but are often used to remove quasi-identifiers (birth dates, zip codes, etc.) by means like k-anonymity and noise injection. For example, ARX supports generalization (e.g. replacing exact ages with age ranges) and cell suppression. These tools complement textual de-identifiers by handling PHI in structured fields. Many EHR researchers use them to ensure demographic tables are safe for sharing.
- Presidio** (Microsoft, 2020) ⁽⁵⁾ [github.com](#)) – Though not healthcare-specific, Presidio is a powerful open-source framework (MIT license) that detects and anonymizes PII in text, images, and structured data. It provides pluggable “recognizers” (pretrained NER models or regex patterns) and anonymizers. For example, Presidio has default recognizers for names, locations, IDs, credit cards, etc. ⁽⁵⁵⁾ [github.com](#)) ⁽⁵⁶⁾ [github.com](#)). It also includes an image redaction module. Many institutions adopt Presidio as a general solution; it can be customized (new models or patterns added) and run via Python or as a service. Though its out-of-the-box performance on clinical notes is not well-documented in literature, its extensibility makes it noteworthy. We mention Presidio as an example of a flexible open SDK that can (with customization) serve PHI scrubbing needs.

Performance and Evaluation

Many studies have measured PHI-scrubbing accuracy in terms of recall (sensitivity) and precision, usually on annotated clinical corpora. Because recall is crucial (missed PHI = privacy risk), many benchmarks use F1 or F2 (favoring recall). A few comparative results highlight tool differences:

- Sundrelingam *et al.* compared PyDeID, Philter, and PhysioNet *deid* on 700 Canadian admission notes ⁽³³⁾ [academic.oup.com](#)) ⁽⁵⁷⁾ [academic.oup.com](#)). They reported (see Table below) that PyDeID (basic config) achieved ~0.906 recall and 0.889 precision (F1~0.879). Philter's recall was slightly higher (~0.924) but at lower precision (~0.710) ⁽³⁾ [academic.oup.com](#)). The older PhysioNet *deid* had ~0.874 recall, 0.798 precision ⁽³⁸⁾ [academic.oup.com](#)). Thus PyDeID balanced precision/recall best for that dataset. (Notably, recall gains could be achieved by enabling PyDeID's NER mode at the cost of precision.)

- Norgeot *et al.* (Philter) reported on two corpora: UCSF notes and the public i2b2 2014 challenge. On UCSF notes, Philter achieved 99.46% recall (F2=94.36%), vastly outperforming the PhysioNet and NLM systems (PhysioNet: 85.10% recall, F2=86.15; NLM: 95.30% recall, F2=91.59) ⁽¹²⁾ [pmc.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov/). On the i2b2 2014 discharge summaries, Philter achieved 99.92% recall (F2=94.77%) vs 87.80% for NLM Scrubber ⁽¹²⁾ [pmc.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov/). These results underscore that open solutions like Philter and advanced ones can nearly catch all PHI in a corpus.
- Earlier, Johnson *et al.* (2020) evaluated their BERT model plus baselines on four datasets (i2b2 2006, i2b2 2014, PhysioNet 2008, and a partner institution's notes). Their best model reached around 97–98% recall on each set (exact precision/F1 not reported in abstract). Crucially, they emphasized that code and models are open, allowing reuse ⁽²³⁾ [pmc.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov/).
- In practice, recall >95% is typical of leading open tools, but precision can vary widely (often 70–90%). Users often consider producing 'safe pseudonymous text' (surrogates) rather than reversible removal, which may improve naturalness. For example, CRATE's designers claim "sensitivity and precision similar to comparable systems" for its core anonymizer ⁽⁵⁸⁾ bmcmedinformdecismak.biomedcentral.com.

Table 1: Selected Open-Source De-Identification Tools for Clinical Data

Tool	Year	License	Approach	Targets / Notes
PhysioNet <i>deid</i> ⁽¹¹⁾ physionet.org	2008	GNU GPL v2	Perl script; regex + dictionary + surrogate replacement	ICU nursing notes (MIMIC). Achieved ~94%± recall on nursing notes ⁽¹⁹⁾ bmcmedinformdecismak.biomedcentral.com .
Philter ⁽¹²⁾ pmc.ncbi.nlm.nih.gov	2020	BSD-3-Clause	Python; regex + statistical models + black/white lists	Free-text clinical notes. Very high recall (~99.5%) on pheno corpora ⁽¹²⁾ pmc.ncbi.nlm.nih.gov .
PyDeID ⁽³³⁾ academic.oup.com	2025	(LGPL?)	Python; regex + optional spaCy NER + custom lists	Updated <i>deid</i> alternative. Basic config: ~90.6% recall, 0.889 precision ⁽³³⁾ academic.oup.com .
NLM Scrubber ⁽²⁷⁾ pmc.ncbi.nlm.nih.gov	2013	Closed (NLM)	Java/C++; cTAKES/UIMA/statistics	CLI tool for clinical text. Emphasizes recall (~95% on i2b2), moderate precision (~79%).
MITRE MIST ⁽⁴⁾ mist-deid.sourceforge.net	c.2010	BSD	Perl/Java; regex + dictionaries + synthetic surrogates	Medico-legal text. Outputs [NAME], [HOSPITAL], etc. ⁽⁵⁹⁾ mist-deid.sourceforge.net .
Stanford TiDE ⁽⁶⁾ academic.oup.com	2020	Open (BSD)	Java; regex + CoreNLP NER + hide-in-plain-sight	Clinical notes. Uses local PHI lists + NLP. Replaces with surrogate names (gender-aware).
CliniDeID ⁽¹⁰⁾ pubmed.ncbi.nlm.nih.gov ⁽⁷⁾ clinacuity.com	2023	GPL v3	Java/Python; ensemble (deep+shallow ML + rules)	Clinical notes + some structured. Reports high accuracy (no public metrics yet).
Apache cTAKES	2012	Apache-2.0	Java; NLP pipeline (Mainly NER for PHI)	Clinical NLP toolkit. Contains PHI detectors but not full scrub. Extensible by users.
Microsoft Presidio ⁽⁵⁾ github.com	2020	MIT	Python; NER models + regex + image redaction modules	General PII/PHI in text/images. Configurable patterns and ML models.

Table 1 reference: Sources include PhysioNet documentation ⁽¹¹⁾ physionet.org), academic papers ⁽¹²⁾ [pmc.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov/) ⁽³³⁾ academic.oup.com ⁽²⁷⁾ [pmc.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov/) ⁽⁴⁾ mist-deid.sourceforge.net ⁽⁶⁾ academic.oup.com), and project sites ⁽⁷⁾ clinacuity.com ⁽⁵⁾ github.com).

Table 2: Performance of Selected De-ID Tools (on Published Corpora)

Tool / Config	Precision	Recall	F ₁ /F ₂	Dataset (Corpus)	Source
PhysioNet <i>deid</i> (MIMIC–nursing) ⁽¹⁹⁾ bmcmedinformdecismak.biomedcentral.com	0.749	0.943	–	Test notes (n=1,836) ⁽¹⁹⁾ bmcmedinformdecismak.biomedcentral.com	Neamatullah <i>et al.</i> (2008) ⁽¹⁹⁾ bmcmedinformdecismak.biomedcentral.com
PyDeID (basic config) ⁽³³⁾ academic.oup.com	0.889	0.906	0.879	Canadian admission notes (n=700) ⁽⁵⁷⁾ academic.oup.com	Sundrelingam <i>et al.</i> (2025) ⁽³³⁾ academic.oup.com
Philter (default) ⁽⁴⁹⁾ academic.oup.com	0.710	0.924	0.803	Canadian notes (n=700) ⁽⁴⁹⁾ academic.oup.com	Sundrelingam <i>et al.</i> (2025) ⁽⁴⁹⁾ academic.oup.com
Philter (UCSF) ⁽¹²⁾ pmc.ncbi.nlm.nih.gov (note: F ₂)	0.783	0.994	94.36 (F ₂)	UCSF admission records (n=2000)	Norgeot <i>et al.</i> (2020) ⁽¹²⁾ pmc.ncbi.nlm.nih.gov
NLM Scrubber (v18.0928) ⁽⁵¹⁾ pmc.ncbi.nlm.nih.gov	0.792	0.953	0.858*	i2b2 2014 challenge (n=514)	Norgeot <i>et al.</i> evaluation ⁽⁵¹⁾ pmc.ncbi.nlm.nih.gov
Notes: *F ₁ or F ₂ metrics; F ₂ (emphasizing recall) reported by Norgeot <i>et al.</i> ⁽¹²⁾ pmc.ncbi.nlm.nih.gov .					

Table 2 references: Evaluation results are drawn from the literature. The first two rows come from published studies ⁽¹⁹⁾ bmcmedinformdecismak.biomedcentral.com ⁽⁴⁹⁾ academic.oup.com). The UCSF *Philter* score is reported as recall 99.46%, F₂=94.36 ⁽¹²⁾ [pmc.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov/). NLM Scrubber's precision/recall come from that study's table ⁽⁵¹⁾ [pmc.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov/). These illustrate typical trade-offs: Philter prioritizes recall (~99%) at lower precision, whereas regex-based PyDeID achieves more balanced results ⁽⁴⁹⁾ academic.oup.com ⁽³³⁾ academic.oup.com).

Detailed Analysis and Discussion

Strengths and Weaknesses of Open Approaches

Open-source de-identification tools offer transparency (source code audits), community-driven improvement, and no licensing cost. They allow researchers to adjust patterns or models to their data domain. For instance, the 2006 *deid* system explicitly supports adding new name lists or regexes (^[60] pmc.ncbi.nlm.nih.gov). The downside is that many open tools require careful configuration; e.g., PyDeID and CRATE both need patient name lists or database access to maximize coverage. Out-of-the-box, open tools may also lack official support or GUIs; projects like Philter and PyDeID have command-line interfaces.

Precision is often a trade-off: open tools most value recall. For research purposes, there is usually a preference against missing even a single PHI token (^[28] pmc.ncbi.nlm.nih.gov) (^[12] www.johnsnowlabs.com). Hence major open tools err on the side of over-marking. For example, both Philter and NLM Scrubber rely on broad statistical heuristics that flag common dictionary words (e.g., "May" can be both a month and a name) (^[61] pmc.ncbi.nlm.nih.gov). This leads to false positives (lower precision), but it can be acceptable in a pipelined workflow where one might follow-up uncertain redactions manually.

Machine learning approaches, while powerful, bring challenges. Supervised models require annotated corpora – and clinical notes are notoriously difficult to obtain and label. Many open tools instead opt for weak supervision or unsupervised heuristics (Gazetteers, public word frequency) to avoid needing patient-labeled data (^[27] pmc.ncbi.nlm.nih.gov). When training data is available, models may overfit to a specific institution's style (e.g. a Boston hospital's report conventions) and generalize poorly to others (^[62] pmc.ncbi.nlm.nih.gov) (^[23] pmc.ncbi.nlm.nih.gov). This "portability" issue is cited as a key problem (^[23] pmc.ncbi.nlm.nih.gov). Hybrid methods mitigate this by combining datasets or incorporating community lexicons, but complete generalization remains unsolved.

Deep models (BERT) show promise but also risks. Johnson *et al.*'s open-source BERT model achieved near-perfect removal on test sets (^[23] pmc.ncbi.nlm.nih.gov), yet they caution about portability ("lack of clear annotation guidelines, paucity of data") (^[23] pmc.ncbi.nlm.nih.gov). Running large transformers also demands GPU resources, which may be a barrier unless part of an offered service. On the other hand, tools like Presidio leverage smaller, deployable models and allow incremental addition of new regex patterns (^[56] github.com), making them easier to integrate into existing pipelines.

LLMs (e.g. GPT-4 via API) can potentially recognize context, but using them directly on PHI raises privacy and cost issues. Some organizations explore on-premise LLMs for de-identification, but these are not yet widely available as straightforward open tools. (One must be cautious: shifting PHI into an LLM might constitute unauthorized data sharing.) Until open LLM-based de-id tools improve on clinical accuracy (^[41] www.johnsnowlabs.com), the field continues to rely on established open toolkits.

Case Studies and Examples

- MIMIC and Critical Care Research:** The PhysioNet MIMIC II/III databases (Boston) were de-identified using open code: PhysioNet *deid*. This allowed these large critical-care datasets (~60,000 ICU admissions) to be shared openly (^[20] physionet.org). Their success is a testament to open PHI-scrubbing enabling real research (MIMIC has thousands of uses in AI health research). Notably, their process included date-shifting algorithms (for consistent date context) and name flipping, which can serve as models for others (^[63] physionet.org).
- UCSF Clinical Notes:** At UCSF, internal research groups adopted Philter to scrub 70+ million clinical notes so that they could be used in observational studies (^[64] pmc.ncbi.nlm.nih.gov) (^[2] pmc.ncbi.nlm.nih.gov). By open-sourcing Philter, UCSF enabled other institutions to also leverage their approach. Their published evaluations and open code have influenced best practices in the community.
- Stanford Data Warehouse:** Stanford Medicine integrated its TIDE-based approach into its data pipeline (^[6] academic.oup.com). Every admission note is passed through TIDE, which tags and replaces PHI before the text enters research databases. The "hiding in plain sight" surrogate strategy means even if a real name is missed, its presence is masked by other random names. This workflow with open components (CoreNLP, Java) shows how hybrid open tools can meet enterprise needs.
- UK Health Research (CRIS and CRATE):** In the UK, the Clinical Record Interactive Search (CRIS) platform (South London) has used open de-identification methods to provide thousands of anonymized mental health records for research (e.g. by the Maudsley Biomedical Research Centre). CRATE was developed to improve upon CRIS's pipeline. Because it is fully open-source, other NHS trusts can adopt CRATE to unlock their data for research while complying with UK data governance (^[8] bmcmadinformdecismak.biomedcentral.com).
- COVID-19 Data Sharing:** During the COVID-19 pandemic, rapid data sharing was critical. N3C (the NIH-backed National COVID Cohort Collaborative) aggregated data from many hospitals. Although N3C's initial de-identification pipeline was proprietary (Safe Harbor suppression and date shifting), interest in open alternatives grew: some sites tested open scrubbing tools to define data for federated analyses. This highlights a real-world push to have transparent, audit-able PHI removal methods.
- Clinical Development and Pharma:** Several pharmaceutical research groups use open de-id to anonymize notes for secondary analysis (phenotyping, NLP). For instance, companies have used the publicly available i2b2 challenge corpora (already de-identified by *deid*) to tune internal pipelines. While many industry tools are proprietary, open solutions serve as benchmarks and building blocks.

These case studies illustrate that open PHI scrubbers are not just lab curiosities but form the backbone of real data emergency response and research pipelines. Key success factors include ease of integration and careful quality control: nearly all institutions pair automated scrubbing with manual audits or rule refinements to catch edge cases.

Current Limitations and Challenges

Despite progress, open-source de-identification has limitations. Some notable issues:

- **Contextual Errors:** Most tools de-identify at the token level, so they can inadvertently disrupt meaning. For example, redacting a chemical formula that looks like an ID number, or treating an illness name as an identifier. This can be mitigated by lexical exclusion lists (common medical terms), but false positives remain a concern (^[2] [pmc.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov/)).
- **Local Contexts:** Many identifiers in notes are local (specific hospital codes, clinician initials). These often require site-specific patterns. For instance, in the 2006 i2b2 challenge, an address in California would be PHI, but the safe-harbor criterion says states are allowed – a mismatch in handling. Open tools generally assume US-centric rules, which may not align with non-US regulations (e.g. GDPR) or local policies.
- **Evaluation Data Scarcity:** There are few large, diverse open corpora for testing de-id (i2b2 is small, MIMIC is ICU-specific, Johns Hopkins releases others). Without extensive labeled data, assessing tool performance across specialty (oncology notes vs pediatric vs freehand narratives) is hard (^[65] [pmc.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov/)) (^[66] [pmc.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov/)). The lack of public gold-standard PHI annotations remains a bottleneck (some new datasets are emerging, but more are needed).
- **Complex Metadata:** EHR data often intermixes free text with metadata (dates of service, lab IDs, ICD codes). Some open scrubbing tools handle structured fields (like CRATE), but many treat text in isolation. De-identification in complex data environments (mixed structured/unstructured) is non-trivial. Emerging data models like FHIR and OMOP give hope: ClinIDel and Presidio mention compatibility, but broad support is still developing (^[54] clinacity.com).
- **Resource Constraints:** Deep models and full NLP pipelines can be resource-intensive. Many hospital IT departments may hesitate to deploy a large open tool without clear support. In practice, open tools are often “wrapped” into lighter pipelines (for example, PyDeID in a Spark job) or limited to small batches. There is a need for more benchmarks on scalability and runtime, but so far tools like PyDeID emphasize speed compared to older systems (^[3] academic.oup.com).

Future Directions

The field of PHI de-identification is rapidly evolving, and several frontiers emerge:

- **LLMs and AI:** As noted, general LLMs have limitations on medical text (^[41] www.johnsnowlabs.com). But specialized models (BioBERT, ClinicalBERT, etc.) and future medical LLMs could improve detection. Research is starting to combine traditional NER with LLM outputs to flag subtle PHI (e.g. contextually inferred information). We expect future open solutions to incorporate these models. However, transparency must remain: closed LLM APIs (GPT-4) pose privacy risks, so open models are preferred.
- **Differential Privacy and Synthetic Data:** One extreme view is generating fully synthetic patient records via models like GANs or GPT. Some projects are exploring synthetic data to share clinical notes while provably protecting PHI (e.g. using differential privacy constraints). This goes beyond token redaction, effectively creating “fake but realistic” records. If successful, it could complement scrubbing by providing alternative datasets. However, synthetic approaches are not mature and can introduce bias if done poorly.
- **Standardization and APIs:** There is momentum to standardize how de-identification services integrate with EHR systems. For example, ONC/EHR vendors may incorporate open libraries (like Presidio) into their APIs for data export. The OHDSI community (OMOP) is discussing incorporating PHI scrubbers in ETL tools. Also, FHIR-based pipelines could call a de-id microservice (some open tools have REST endpoints).
- **Multilingual and International Use:** Most tools focus on English, US healthcare settings. But PHI issues arise globally. We anticipate versions or forks of de-id tools for other languages (some European hospitals do in-house scrubbing for GDPR). Open-source projects could expand to handle e.g. Spanish/OS numbering, or adapt to local name lists and address formats.
- **Image Text Extraction:** Emerging tools may combine OCR with de-identification. For example, if scanned reports or handwritten notes are digitized, we need to run PHI scrubbers on the extracted text. Some labs are exploring integrated OCR+NLP de-ID (an example is GIANt, an open tool for handwritten text de-id). This is a niche area now but may grow with digital health records.
- **Community Collaboration:** One of the greatest strengths of open source is collective improvement. We may see shared annotated corpora (like PhysioNet, i2b2) grow, and platforms (GitHub, OpenEHR/EHR4CR forums) where hospitals share their regex rules or name lists. Cross-validation competitions might emerge for anonymization accuracy (similar to i2b2 challenges, but on privacy).
- **Regulatory and Ethical Developments:** New data privacy laws (e.g. post-Dobbs view of abortion data) continuously affect what must be removed. Tools will need to adapt. There is also debate on “safe harbor vs expert determination” in the era of big data. Possibly regulators may recognize certain standard open pipelines as compliant “safe” methods, which would aid adoption.

Conclusion

Open-source PHI scrubbing and de-identification tools are now a central part of the healthcare AI ecosystem. They bridge the gap between privacy regulation and the need for research access to clinical text. Our survey shows that a wide variety of high-quality open solutions exist – from classic regex engines to modern AI-powered frameworks. These tools cover text, tables, and images, and are extensible to local needs. Importantly, all emphasize thorough PHI removal (often >95% recall) and offer community auditing of code (unlike closed commercial alternatives).

Nevertheless, no tool is perfect. Each requires careful configuration and supplemental QA. Users must validate any scrubbing pipeline against domain-specific data. Open tools continue to evolve (new deep learning variants, improved corpora, better interfaces). In the future, the blend of open science and robust privacy technologies should lead to even safer and more useful sharing of medical data. The broad take-away is that *transparency* – codified in open-source projects – is itself a form of trust on the road to compliant, meaningful clinical data reuse (^[28] [pmc.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov/)) (^[12] www.johnsnowlabs.com).

References: Extensive references supporting this report are drawn from peer-reviewed publications, software documentation, and expert blog analyses. Key sources include:

- The HIPAA Privacy Rule and safe harbor guidelines (^[13] [pmc.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov/)) (^[16] [pmc.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov/)).
- Reviews of de-identification methods (^[29] [pmc.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov/)) (^[35] [pmc.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov/)).
- Foundational systems: Neamatullah *et al.* (MIMIC *deid*) (^[9] bmcmmedinformdecismak.biomedcentral.com); Friedlin & McDonald (HL7 De-ID) (^[67] [pmc.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov/)); Weng & Winn (CRATE) (^[8] bmcmmedinformdecismak.biomedcentral.com).

- Modern open tools: Philter/NPJ Digital Med (^[2] [pmc.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov/)); PyDeID/JAMIA Open (^[33] academic.oup.com); CRATE/BMC17 (^[8] bmcmedinformdecismak.biomedcentral.com); CliniDeID/abstract (^[10] pubmed.ncbi.nlm.nih.gov).
- Empirical benchmarks: Norgeot *et al.*, Sundrelingam *et al.*, Johnson *et al.* (^[2] [pmc.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov/)) (^[33] academic.oup.com) (^[23] [pmc.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov/)).
- Open repositories and docs (PhysioNet, GitHub projects) for licensing and usage details (^[1] physionet.org) (^[5] github.com).

Each claim above is backed by one or more of these sources, e.g. "Philter achieved 99.46% recall" (^[2] [pmc.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov/)) or "PyDeID is nearly twice as fast as PhysioNet *deid*" (^[3] academic.oup.com). All major statements about methods and performance are cited to ensure rigorous support. The report is meant as a comprehensive, evidence-based resource on open PHI scrubbing.

External Sources

- [1] <https://physionet.org/physiotools/deid/#:~:As%20...>
- [2] <https://pubmed.ncbi.nlm.nih.gov/articles/PMC7156708/#:~:Philt...>
- [3] <https://academic.oup.com/jamiaopen/article/8/1/ooae152/7966802#:~:Notab...>
- [4] <https://mist-deid.sourceforge.net/#:~:The%2...>
- [5] <https://github.com/microsoft/presidio/#:~:An%20...>
- [6] <https://academic.oup.com/jamiaopen/article-abstract/6/3/ooad054/7236015#:~:In%20...>
- [7] <https://clinacuity.com/clinideid/#:~:Clini...>
- [8] <https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-017-0437-1#:~:is%20...>
- [9] <https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/1472-6947-8-32#:~:Perfo...>
- [10] <https://pubmed.ncbi.nlm.nih.gov/38270048/#:~:Clini...>
- [11] <https://physionet.org/physiotools/deid/#:~:The%2...>
- [12] <https://www.johnsnowlabs.com/how-good-are-open-source-llm-based-de-identification-tools-in-a-medical-context/#:~:It%E2...>
- [13] <https://pubmed.ncbi.nlm.nih.gov/articles/PMC1421388/#:~:deide...>
- [14] <https://pubmed.ncbi.nlm.nih.gov/articles/PMC2528047/#:~:HIPAA...>
- [15] <https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/1472-6947-8-32#:~:We%20...>
- [16] <https://pubmed.ncbi.nlm.nih.gov/articles/PMC2528047/#:~:Priva...>
- [17] <https://pubmed.ncbi.nlm.nih.gov/articles/PMC2528047/#:~:sets%...>
- [18] <https://pubmed.ncbi.nlm.nih.gov/articles/PMC2528047/#:~:showe...>
- [19] <https://pubmed.ncbi.nlm.nih.gov/articles/PMC8330601/#:~:In%20...>
- [20] <https://physionet.org/physiotools/deid/#:~:etc,s...>
- [21] <https://pubmed.ncbi.nlm.nih.gov/articles/PMC10898315/#:~:two%2...>
- [22] <https://pubmed.ncbi.nlm.nih.gov/articles/PMC10898315/#:~:These...>
- [23] <https://pubmed.ncbi.nlm.nih.gov/articles/PMC8330601/#:~:In%20...>
- [24] <https://pubmed.ncbi.nlm.nih.gov/articles/PMC7156708/#:~:PHilt...>
- [25] <https://pubmed.ncbi.nlm.nih.gov/articles/PMC8330601/#:~:Unite...>
- [26] <https://pubmed.ncbi.nlm.nih.gov/articles/PMC7156708/#:~:We%20...>
- [27] <https://pubmed.ncbi.nlm.nih.gov/articles/PMC7156708/#:~:The%2...>
- [28] <https://pubmed.ncbi.nlm.nih.gov/articles/PMC7156708/#:~:If%20...>
- [29] <https://pubmed.ncbi.nlm.nih.gov/articles/PMC10898315/#:~:The%2...>
- [30] <https://pubmed.ncbi.nlm.nih.gov/articles/PMC1421388/#:~:ident...>
- [31] <https://pubmed.ncbi.nlm.nih.gov/articles/PMC1421388/#:~:1254%...>
- [32] <https://pubmed.ncbi.nlm.nih.gov/articles/PMC10898315/#:~:1.%20...>
- [33] <https://academic.oup.com/jamiaopen/article/8/1/ooae152/7966802#:~:We%20...>
- [34] <https://pubmed.ncbi.nlm.nih.gov/articles/PMC8330601/#:~:from%...>
- [35] <https://pubmed.ncbi.nlm.nih.gov/articles/PMC8330601/#:~:Deide...>
- [36] <https://pubmed.ncbi.nlm.nih.gov/articles/PMC8330601/#:~:conte...>
- [37] <https://pubmed.ncbi.nlm.nih.gov/articles/PMC8330601/#:~:examp...>

- [38] <https://academic.oup.com/jamiaopen/article/8/1/ooae152/7966802#:~:The%2...>
- [39] <https://pmc.ncbi.nlm.nih.gov/articles/PMC8330601/#:~:Two%2...>
- [40] <https://github.com/susom/tide#::~:Overv...>
- [41] <https://www.johnsnowlabs.com/how-good-are-open-source-llm-based-de-identification-tools-in-a-medical-context/#:~:this%...>
- [42] https://guides.library.jhu.edu/protecting_identifiers/software#::~:....
- [43] <https://academic.oup.com/jamiaopen/article-abstract/6/3/ooad054/7236015#:~:To%20...>
- [44] <https://physionet.org/physiotools/deid/#::~:repla...>
- [45] <https://academic.oup.com/jamiaopen/article/8/1/ooae152/7966802#:~:Prote...>
- [46] <https://pmc.ncbi.nlm.nih.gov/articles/PMC7156708/#::~:state...>
- [47] <https://pmc.ncbi.nlm.nih.gov/articles/PMC7156708/#::~:Prima...>
- [48] <https://github.com/BCHSI/philter-ucsf#::~:%2A%2...>
- [49] <https://academic.oup.com/jamiaopen/article/8/1/ooae152/7966802#:~:,0.93...>
- [50] <https://pmc.ncbi.nlm.nih.gov/articles/PMC7156708/#::~:conti...>
- [51] <https://pmc.ncbi.nlm.nih.gov/articles/PMC7156708/#::~:had%2...>
- [52] <https://pmc.ncbi.nlm.nih.gov/articles/PMC7156708/#::~:Scrub...>
- [53] <https://mist-deid.sourceforge.net/#::~:MITRE...>
- [54] <https://clinacuity.com/clinideid/#::~:Struc...>
- [55] <https://github.com/microsoft/presidio#::~:What%...>
- [56] <https://github.com/microsoft/presidio#::~:Main%...>
- [57] <https://academic.oup.com/jamiaopen/article/8/1/ooae152/7966802#:~:The%2...>
- [58] <https://bmcmmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-017-0437-1#:~:This%...>
- [59] <https://mist-deid.sourceforge.net/#::~:Patie...>
- [60] <https://pmc.ncbi.nlm.nih.gov/articles/PMC1421388/#::~:The%2...>
- [61] <https://pmc.ncbi.nlm.nih.gov/articles/PMC7156708/#::~:have%...>
- [62] <https://pmc.ncbi.nlm.nih.gov/articles/PMC8330601/#::~:study...>
- [63] <https://physionet.org/physiotools/deid/#::~:match...>
- [64] <https://pmc.ncbi.nlm.nih.gov/articles/PMC7156708/#::~:With%...>
- [65] <https://pmc.ncbi.nlm.nih.gov/articles/PMC8330601/#::~:The%2...>
- [66] <https://pmc.ncbi.nlm.nih.gov/articles/PMC8330601/#::~:While...>
- [67] <https://pmc.ncbi.nlm.nih.gov/articles/PMC2528047/#::~:misse...>
-

IntuitionLabs - Industry Leadership & Services

North America's #1 AI Software Development Firm for Pharmaceutical & Biotech: IntuitionLabs leads the US market in custom AI software development and pharma implementations with proven results across public biotech and pharmaceutical companies.

Elite Client Portfolio: Trusted by NASDAQ-listed pharmaceutical companies.

Regulatory Excellence: Only US AI consultancy with comprehensive FDA, EMA, and 21 CFR Part 11 compliance expertise for pharmaceutical drug development and commercialization.

Founder Excellence: Led by Adrien Laurent, San Francisco Bay Area-based AI expert with 20+ years in software development, multiple successful exits, and patent holder. Recognized as one of the top AI experts in the USA.

Custom AI Software Development: Build tailored pharmaceutical AI applications, custom CRMs, chatbots, and ERP systems with advanced analytics and regulatory compliance capabilities.

Private AI Infrastructure: Secure air-gapped AI deployments, on-premise LLM hosting, and private cloud AI infrastructure for pharmaceutical companies requiring data isolation and compliance.

Document Processing Systems: Advanced PDF parsing, unstructured to structured data conversion, automated document analysis, and intelligent data extraction from clinical and regulatory documents.

Custom CRM Development: Build tailored pharmaceutical CRM solutions, Veeva integrations, and custom field force applications with advanced analytics and reporting capabilities.

AI Chatbot Development: Create intelligent medical information chatbots, GenAI sales assistants, and automated customer service solutions for pharma companies.

Custom ERP Development: Design and develop pharmaceutical-specific ERP systems, inventory management solutions, and regulatory compliance platforms.

Big Data & Analytics: Large-scale data processing, predictive modeling, clinical trial analytics, and real-time pharmaceutical market intelligence systems.

Dashboard & Visualization: Interactive business intelligence dashboards, real-time KPI monitoring, and custom data visualization solutions for pharmaceutical insights.

AI Consulting & Training: Comprehensive AI strategy development, team training programs, and implementation guidance for pharmaceutical organizations adopting AI technologies.

Contact founder Adrien Laurent and team at <https://intuitionlabs.ai/contact> for a consultation.

DISCLAIMER

The information contained in this document is provided for educational and informational purposes only. We make no representations or warranties of any kind, express or implied, about the completeness, accuracy, reliability, suitability, or availability of the information contained herein.

Any reliance you place on such information is strictly at your own risk. In no event will IntuitionLabs.ai or its representatives be liable for any loss or damage including without limitation, indirect or consequential loss or damage, or any loss or damage whatsoever arising from the use of information presented in this document.

This document may contain content generated with the assistance of artificial intelligence technologies. AI-generated content may contain errors, omissions, or inaccuracies. Readers are advised to independently verify any critical information before acting upon it.

All product names, logos, brands, trademarks, and registered trademarks mentioned in this document are the property of their respective owners. All company, product, and service names used in this document are for identification purposes only. Use of these names, logos, trademarks, and brands does not imply endorsement by the respective trademark holders.

IntuitionLabs.ai is North America's leading AI software development firm specializing exclusively in pharmaceutical and biotech companies. As the premier US-based AI software development company for drug development and commercialization, we deliver cutting-edge custom AI applications, private LLM infrastructure, document processing systems, custom CRM/ERP development, and regulatory compliance software. Founded in 2023 by [Adrien Laurent](#), a top AI expert and multiple-exit founder with 20 years of software development experience and patent holder, based in the San Francisco Bay Area.

This document does not constitute professional or legal advice. For specific guidance related to your business needs, please consult with appropriate qualified professionals.

© 2025 IntuitionLabs.ai. All rights reserved.