Submitter: Lakshmi Radhakrishnan
Data Scientist
Lakshmi.radhakrishnan@ucsf.edu

UC Location represented: UCSF

Award Category: Sautter Awards, Team Recognition Award (Operational Excellence Award)

Project Name: Philter - Automated Clinical Notes De-Identification Pipeline

Project Members: Lakshmi Radhakrishnan, Gundolf Schenk, Kathleen Muenzen, Boris Oskotsky, Sharat Israni, Atul J. Butte

Summary:

An automated clinical text de-identification pipeline to de-identify clinical note text and reports is invaluable for medical research. At UCSF, we have established such a pipeline that is HIPAA[1] compliant, and scalable to millions of clinical texts. To the best of our knowledge, our de-identification pipeline with its principal algorithm packaged as Philter V1.0 is currently the only professionally certified de-identification software for unstructured data and is now being adopted by many other medical institutions, including the University of California, Irvine and the University of California, Davis.

Project Narrative:

The field of Precision Medicine is quickly generating new approaches to disease treatment that are customized for patients based on their personal genetics, medical history, lifestyle and social determinants. For targeted therapies to advance, there is a growing need for in-depth patient data beyond what is available in structured Electronic Health Record (EHR) data. Clinical notes contain detailed accounts of a patient's medical and family history, lifestyle, disease progression, treatment plans, doctor sentiments and prognoses, which are not typically captured in structured EHR data. Such information is highly valuable for personalized disease treatment research projects. For example, de-identified clinical notes proved extremely valuable during the COVID-19 pandemic, when many investigators were interested in using clinical notes to quickly advance COVID-19 research without having to undergo lengthy Institutional Review Board (IRB) reviews. Unfortunately, clinical notes remain largely unexplored in precision medicine research studies due to the presence of Protected Health Information (PHI) and access restrictions posed by IRBs. In large institutions like UCSF, about 2 million new clinical notes are generated monthly, thus steadily building the institution's current corpus of nearly 120 million notes. Hence, a fast and robust pipeline that can reliably de-identify millions of clinical notes at once is invaluable. Since the publication of the algorithm (Norgeot et al. NPJ Digit. Med. 2020), our focus has been (1) to develop a working clinical text de-identification pipeline that is HIPAA compliant; (2) share routinely updated de-identified clinical notes with researchers to reach new frontiers in medical research; and (3) make the de-identification pipeline robust and scalable to hundreds of millions of clinical notes.

In December 2021, we released and implemented Philter V1.0 at our institution. This is a certified de-identification pipeline in accordance with the HIPAA Privacy Rule. We contracted with ArcherHall, a data forensics company recommended by UCOP, to audit our de-identification pipeline according to two methods that can be used to satisfy the Privacy Rule's de-identification standard: Expert Determination and Safe Harbor. It took us three years of development, rigorous checks for lingering PHI, and multiple iterations of verification by the forensics experts to get the algorithm, the associated pipeline, and the machine-redacted clinical notes as certified de-identified. We also replaced the use of traditional Linux file system with a Mongo database together with a system to track data changes and accelerate the de-identification process on a large and diverse corpus of clinical notes with over 150 note types. Combined with enhanced parallel processing and powerful infrastructure, we were able to bring down the processing for redacting our large corpus of clinical notes from months to a mere few weeks, making our de-identification sustainable for a monthly refresh cycle.

To the best of our knowledge, UCSF's Philter V1.0 pipeline is currently the first and only certified, de-identified redaction pipeline that makes clinical notes available to researchers for non-human subjects' research, without the need for further IRB approval. This project aligns well with the UC Tech's Operational Excellence Award Criteria

[1] The Health Insurance Portability and Accountability Act of 1996

as our team has provided a complete solution for making unstructured clinical note text available for researchers. To date, we have made nearly 120 million certified machine redacted clinical notes available to more than 500 researchers. These notes were collected over the past 30 years and represent data from 2.6 million UCSF patients. Several research projects at UCSF have benefitted from the availability of periodically updated, de-identified clinical note texts. To improve usability, we have set up Apache cTAKES™, which is a high throughput natural language processing (NLP) pipeline for extracting clinically relevant information from clinical text. All 120 million clinical notes have been processed through cTAKES, extracting the incredible amount of 6 billion medical concepts. We have also set up programmatic and non-programmatic tools and environments for researchers to be able to easily work with our de-identified notes corpus. The UCSF Information Commons has some best-of-breed research data exploration and computational analysis tools, like Jupyter, RStudio, Hue leveraging Apache Spark, Facebook Presto, and AI/ML libraries such as ScikitLearn and TensorFlow. We also provide robust non-programmatic tools like EMERSE for our clinical note text, that enables users to text search the de-identified clinical notes through a user-friendly interface. This software was developed at the University of Michigan and implemented here in partnership with CTSI.

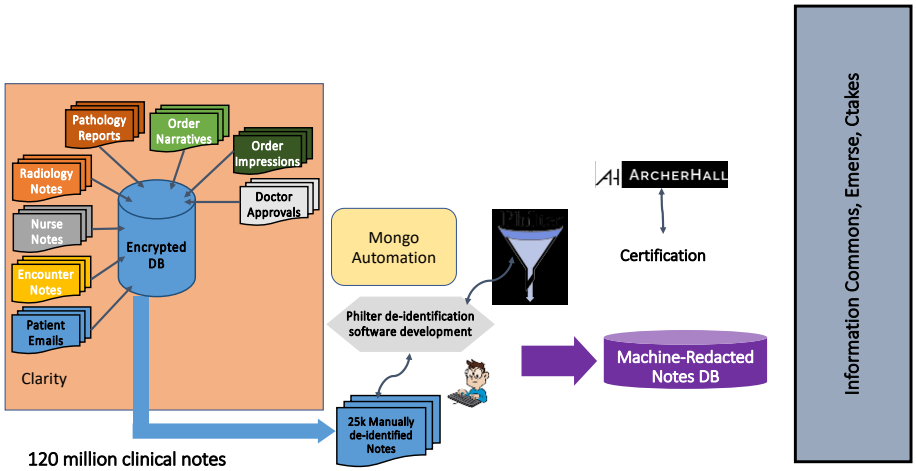

Figure 1 Clinical Notes De-identification Pipeline

The de-identified notes and extracted concepts are actively being used by dozens of researchers at UCSF. Some of the research projects that have used this data resource and successfully published their findings include:

- Rudrapatna V, *et al*. "Accurate Machine Classification of Ulcerative Colitis Mayo Subscores From Electronic Health Record Procedure Reports". The American Journal of Gastroenterology. 2020 Oct; 115(p S420).
- Farrand E, *et al*. "Identifying Patients with Interstitial Lung Disease in Electronic Helath Records: Development and Validation of Machine Learning Algorithms". AMIA. 2022 March.
- Dayan I, *et al*. "Federated learning for predicting clinical outcomes in patients with COVID-19". Nature Medicine. 2021 Sep; 27: 1735–1743.

Other projects that are actively using de-identified clinical notes and extracted concepts include a hip fracture detection study, an evaluation of differential diagnoses and patient similarity in neurodegenerative conditions, and other studies in the areas of social determinants of health, pathology, and oncology.

Our de-identified clinical note texts and reports are now part of our larger de-identified Clinical Data Warehouse and we have started to map the data to the OMOP common data model to be used cross-institutionally. We provide extensive user group sessions and weekly office hours for our research community where researchers are welcome to bring in their questions and concerns, request tutorials and reach out for help. We have seen researchers not only from UCSF but also researchers from other UC and non-UC systems join our office hours and learn from our process. The current process for clinical notes de-identification, and the associated tools and techniques we provide for researchers, is being adapted at other UC systems that do health research, such as UC Irvine and UC Davis, and UCLA. They will no doubt help support multi-institutional studies and create connections at a broader scale.