Stanford MEDICINE | Technology & Digital Solutions
STANFORD HEALTH CARE & SCHOOL OF MEDICINE

TDS Connection

THE LATEST FOR TDS PARTNERS | *February 2023*

FEATURED STORY

Groundbreaking TiDE Uses NLP to 'Clean' Patient Data

"Our Research IT team has been at the forefront of innovation for over a decade."
— Dr. Nigam H. Shah, Professor of Medicine and Chief Data Scientist for Stanford Health Care

By Somalee Datta, PhD, Director, Research Technology Engineering Services, TDS

Patient data is a valuable tool for medical research, but extra care must be taken with sensitive information such as names, addresses, phone numbers, medical record numbers, dates, and social security numbers.

That's why my team and I created a powerful and groundbreaking "cleaner" I call TiDE, or Text De-Identification of clinical notes. We designed TiDE to harness advanced technologies to re-use patient data, accurately and with improved privacy — and we were years ahead of the curve.

"Our Research IT team has been at the forefront of innovation for over a decade," said Dr. Nigam H. Shah, Professor of Medicine and Chief Data Scientist for Stanford Health Care. "I benefited directly from their vision to create the STRIDE data warehouse back in 2010, and was thrilled to partner with them to bring STARR-OMOP [STAnford Research Repository-Observational Medical Outcomes Partnership] to our community.

Anna Shvets/Unsplash photo

“STARR-OMOP is one of few research data warehouses to provide de-identified clinical text for broad research use and we started providing that almost 5 years ago!”

At the heart of this story is data. And the kernel of this data is patient’s health data. An academic medical center like Stanford Medicine strives every day to improve human health. We do this by data-driven decision-making.

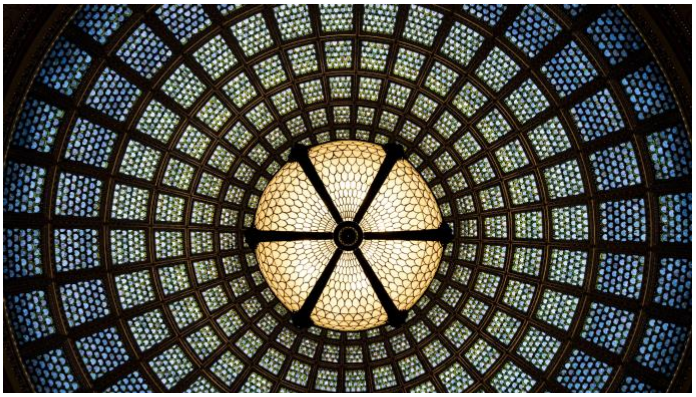
When there is insufficient data, we improve mechanisms to gather data. When there is data, we devise mechanisms to gather the data accurately and in a timely manner and enable data analysis. When the data is deemed sufficient for the purpose of analysis, we put in efforts to make the data FAIR (findable, accessible, interoperable, and reusable). We also make the analysis competitive by giving access to more scientists while following research regulations. Once we have found a winning analysis, we deploy the outcome back in the health system so patients can benefit.

In this circle of data generation, analysis, learning and improving care, there are many regulations that govern its flow, including re-use of the patient data. As technologists, we can help implement these regulations better, so today, we will focus on a common problem, that of anonymizing patient data for approved research.

A subset of the patient data is highly sensitive and is referred to as clinical notes — it’s also the hardest to anonymize.

Say you go to your doctor and you talk a lot, perhaps out of nervousness, or because you’re not sure what information might be relevant. What goes into the clinical notes is substantially more than your diagnosis of hypercholesterolemia or an order of a lipid profile test and a prescription for Lipitor. The clinical notes may contain your genetic predisposition to high cholesterol (maybe your dad suffered from high cholesterol as well), or they may reflect your lifestyle (you are stressed out at work, you don’t find sufficient time to exercise, or a restricted diet depresses you, for example).

Our Stanford Medicine team has combined the advances in cloud computing (billions of bytes processed blazingly fast), with a technique called “Hiding-in-Plain-Sight (HIPS),” to anonymize clinical notes. We present the details of the method in our 2020 manuscript (Supplementary Section 6). Our method of Text De-Identification not only produces better anonymization for the patient, but it is also better for AI research. We fondly named this method TiDE after a household detergent.



TiDE has a suite of Natural Language Processing (NLP) and other data mining tools that recognize personally identifying patterns. (Pexels image)

Stated simply, TiDE has a suite of Natural Language Processing (NLP) and other data mining tools that recognize personally identifying patterns such as names, medical record numbers, and social security numbers.

The pattern, once identified, is substituted by a similar pattern that is not an identifier — i.e., a realistic and fictitious surrogate. For example, a patient name like Ada Lovelace is recognized as a female name and is replaced by a fictitious female name like Mary Smith. This replacement technique of “like-for-like” is the crux of HIPS. The work of identifying such patterns is referred to as Name Entity Recognition or NER. Let’s see what we gain from these approaches when it comes to patient privacy.

Let’s assume that a hypothetical clinical note contains the following line: “*Patient Ada Lovelace came to see me today from London. She is still sick, months after the birth of her second child, Anne Isabella. I will order some more cardiac tests.*” Let’s further assume that TiDE anonymization replaces the above line with: “*Patient Mary Lovelace came to see me today from Detroit. She is still sick, months after the birth of her second child, Veronica Rose. I will order some more cardiac tests.*”

The NER in TiDE recognized Ada as a name, and more specifically a first name. It further recognized the name to be a female name. TiDE also recognized London to be the name of a city. Notice that for the purpose of this illustration, TiDE couldn’t identify Lovelace as a last name. But in case you’re wondering, TiDE can indeed identify Lovelace as a name. It isn’t an uncommon surname in the United States; Ancestry reports >500,000 records of Lovelace. This is a purely hypothetical example used for illustrative purposes.

The realistic like-for-like substitution strategy — Ada with Mary, Detroit with London, Veronica with Anne, and Rose with Isabella — is what eventually results in HIPS where the real leaked identifier Lovelace (i.e., last name) is now hiding among other fake surrogates.

The “substitution” technique replaces an older and more traditional “redaction” technique used commonly in anonymization where the outcome would have been: “*Patient [FIRSTNAME] Lovelace came to see me today from [CITY]. She is still sick, months after the birth of her second child, [FIRSTNAME] [LASTNAME]. I will order some more cardiac tests.”* With traditional redaction of known PII, all or nearly all leaked PII (here, last name “Lovelace”) can be recognized upon reading. With HIPS, leaked content appears unremarkable. To a non-malign reader, i.e., a researcher bound by legal and ethical obligations of research, not interested in expending significant effort to detect leaked PII, the HIPS solution resolves the residual PII problem.

Note that the redaction technique also changes the sentence pattern for the AI. An AI performs the best when it is learning/training on content similar to what it is being tested on. It doesn’t care whether the name is Ada or Mary, as long as we consistently replace all Ada with Mary. However, if it trains on the pattern “*She is still sick, months after the birth of her second child, [FIRSTNAME] [LASTNAME],”* and subsequently sees a real-world pattern like “*She is still sick, months after the birth of her second child, Anne Isabella*.” AI’s performance on the real-world pattern will be poorer.

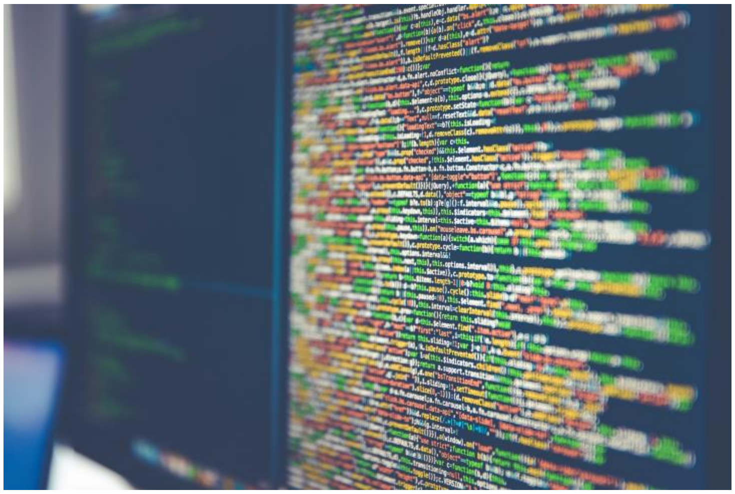
The proof of the AI pudding is in the eating! One of the first publications to use TiDE anonymized clinical text is titled “Discovering monogenic patients with a confirmed molecular diagnosis in millions of clinical notes with MonoMiner,” published in Oct 2022 in *Genetics in Medicine*. Here the authors present a natural language processing (NLP) tool to automatically identify molecularly confirmed monogenic patients from TiDE anonymized free-text clinical notes.

The use of a cloud platform enables performant NERs to operate faster and at a cheaper cost. This is where our engineers can use clever approaches in distributed and/or ephemeral computing to harness large amounts of computing power for a very short duration. TiDE can process approximately 100 million clinical notes in roughly seven hours by deploying 800 standard virtual machines (VMs) in parallel at the total cost of <$450. The total processing time translates to ~0.00025 s/note which is 3 orders of magnitude less than the reported fastest process (0.24 s/note) by Heider *et al*. This significantly reduced cost has allowed Stanford Medicine to publish a monthly updated anonymized observational health research dataset in OMOP (Observational Medical Outcomes Partnership) format.

In the grand scheme of things, the cost is perhaps not the most important aspect. With some additional resources, one can always develop incremental approaches to reduce cost and only update the anonymization of updated notes.

For any algorithm, what is perhaps more important is being able to claim lower error rates in items such as personally identifiable identifiers (PII), such as in the above illustration. In a system like ours, with 4 million active patients, we have more than 100 million clinical notes. To review 1,000,000 clinical notes (1% of total), it would take ~15,000,000 minutes (~15 minutes per note) or 250,000 hours or >$4M (assuming minimum wage).

It is incredibly tedious to review 100 clinical notes, let alone 1,000,000 clinical notes, so we would need a large pool of trained reviewers (~2,500, assuming each individual signs up for 100 hours)! So the feasibility of a human review is ... not that feasible.



Markus Spiske/Unsplash image

We use a mixture of human review and algorithmic quality control to analyze how well we anonymize clinical notes. We start with the top 200 different note types that occur most frequently, pick 1,000 random notes from these, and of these, we pick 100 that have a high amount of PII found or a high number of words found. These 100 go through a manual review for false negatives, i.e., we look for PII like "Anne" left behind. For short, single-line free-form clinical text, like observational measurements data (aka flowsheets), we take 80,000 rows through a word frequency count. We then manually review the lowest frequency 10,000 words for potential PII.

Perhaps an even more pertinent question is whether it is possible to re-identify the patient Ada Lovelace from the TiDE anonymized output, "*Patient Mary Lovelace came to see me today from Detroit. She is still sick, months after the birth of her second child, Veronica Rose. I will order some more cardiac tests.*"

For a malicious entity, it is far more cost effective to attempt breaching IT systems containing identified patient data than to attempt re-identification of HIPS anonymized research data. As an added measure of security, at Stanford, the HIPS anonymized clinical text is treated to the same security requirements as we do for identified patient data. All we have done with TiDE is to increase patient privacy without relaxing legal, ethical, and security measures.

TiDE has been open-source since 2021. We understand that not everyone has access to cloud technologies. We don't assume that access — we also don't demand that you use our distributed programming paradigm.

Instead, the open-source version assumes you might be running the algorithm on your laptop or your local High-Performance Computing (HPC) environment. It is packaged and containerized to run on a set of clinical notes. How you subset your clinical notes is up to you.

In this process of making TIDE accessible, more than anything, we wanted to share that it is indeed possible to improve on traditional anonymization techniques using state-of-the-art technologies without increasing costs, thereby improving patient privacy as well as improving quality of data for AI-driven research.