# Using Membership Inference Attacks to Evaluate Privacy-Preserving Language Modeling Fails for Pseudonymizing Data

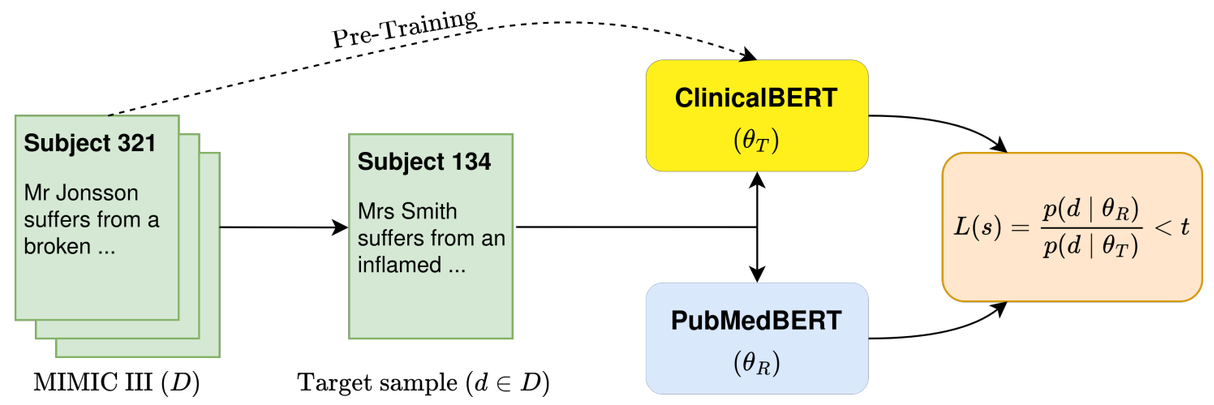
Thomas Vakili & Hercules Dalianis
Department of Computer and System Sciences
Stockholm University, Sweden

## Quantifying the Privacy Risks of Language Models

Language models for clinical tasks can be improved by pre-training using clinical corpora**.** However, language models are susceptible to **privacy attacks** which may reveal **sensitive information** about persons in the data. This is especially problematic when training using inherently sensitive **electronic health records.**

Privacy can be preserved using techniques such as differential privacy and **pseudonymization**. However, there is no consensus on how to compare different techniques. One suggestion is to use **membership inference attacks (MIAs)**.

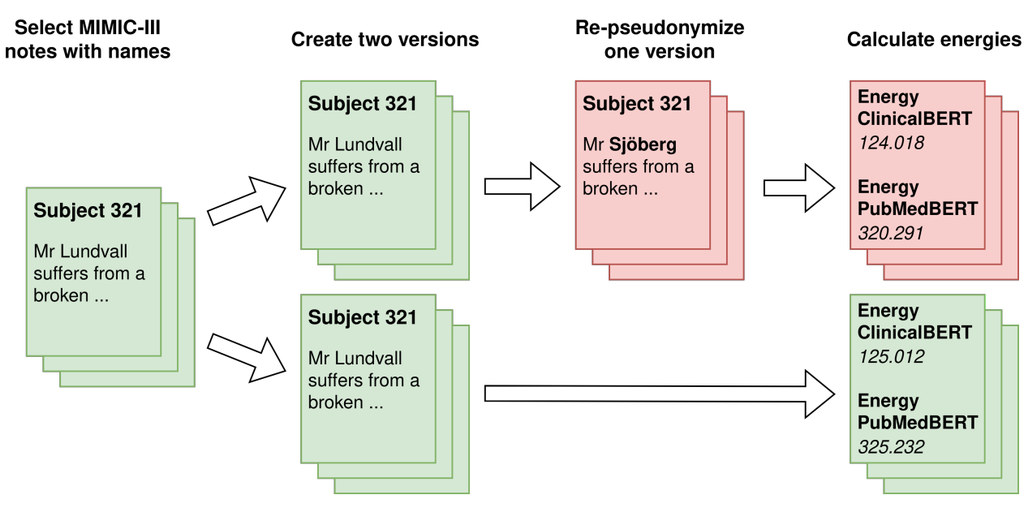
**MIAs** measure **memorization** by mounting attacks that aim to discern if a datapoint was part of a model's training data. Successful attacks are interpreted as a **proxy for privacy risks**.

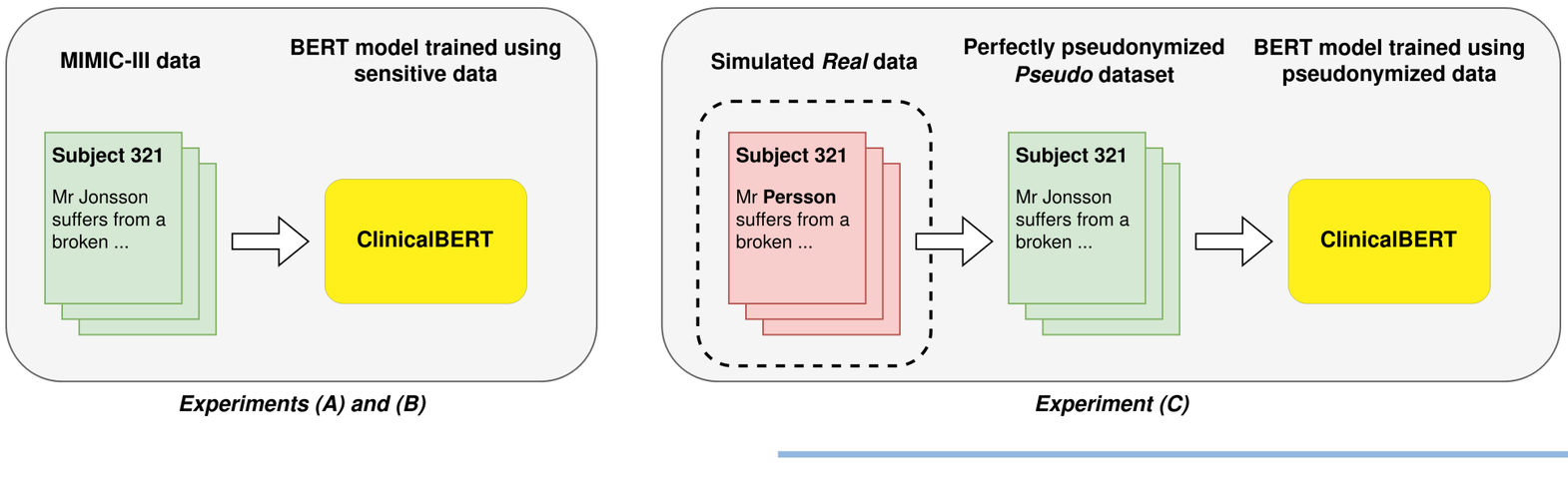



## Token-Level or Datapoint-Level Protection?

Pseudonymization protects information on the **token level**, while MIAs try to detect **datapoint-level membership**. We test how this mismatch affects MIAs for quantifying privacy.

We re-pseudonymize a MIMIC-III subset that was used to train a clinical BERT model by **replacing all names** in 236,114 clinical notes. This allows us to simulate an MIA targeting a model trained using **perfectly pseudonymized data**. If MIAs accurately quantify the privacy benefits of training with pseudonymized data, then **the attack should perform worse**.




## Successful Membership Inference Does Not Imply Privacy Risk

The differences in how well the MIA works are very small. If MIAs are used to quantify privacy, this implies that pseudonymization has negligible privacy benefits.

This is **counter-intuitive** since the privacy measure ensures that **no names are exposed**. The MIA does not capture this.

We conclude that:

- MIAs do not capture the privacy benefits of pseudonymizing training data
- Further research is needed to explore what kinds of memorization are captured by MIAs

| | Attacked data | Accuracy | Precision | Recall | AUC |
|---|---|---|---|---|---|
| (A) | *Pseudo, random sample* | *0.771* | *0.990* | *0.548* | *0.916* |
| (B) | Pseudo, names only | 0.780 | 0.990 | 0.566 | 0.882 |
| (C) | Real, names only | 0.770 | 0.990 | 0.548 | 0.865 |

*thomas.vakili@dsv.su.se*