Utility Preservation of Clinical Text After De-Identification

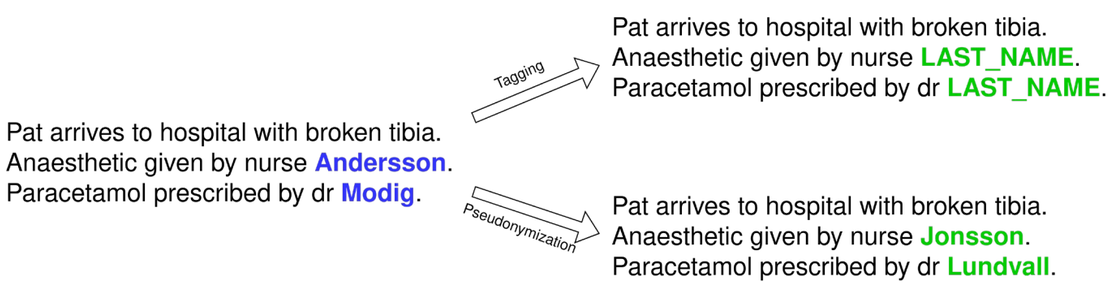
Thomas Vakili & Hercules Dalianis
Department of Computer and System Sciences
Stockholm University, Sweden

Why De-Identification?

The training data for NLP models are sensitive because they often contain information about real humans that may reveal their identity. This is especially important in the clinical domain, where the data being handled are very sensitive. Therefor, there is a need for techniques which **decrease the privacy risks**.

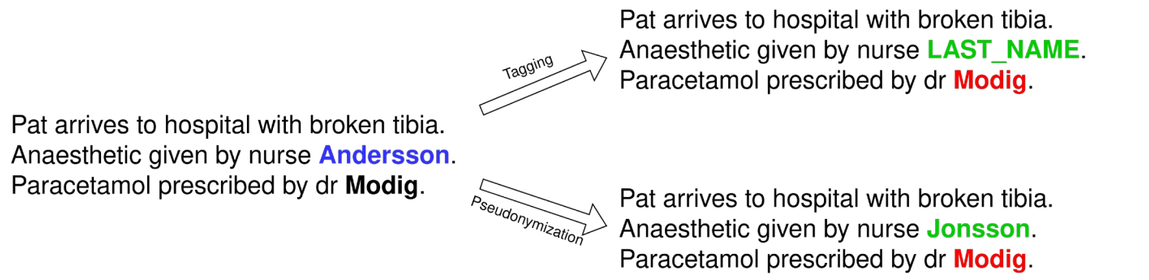
In our work, we look at **automatic de-identification** which is a technique to remove sensitive entities using Named Entity Recognition (NER). We look at two approaches:

1. **Pseudonymization:** which replaces sensitive entities with realistic surrogates
2. **Tagging:** where tag names replace the sensitive entities

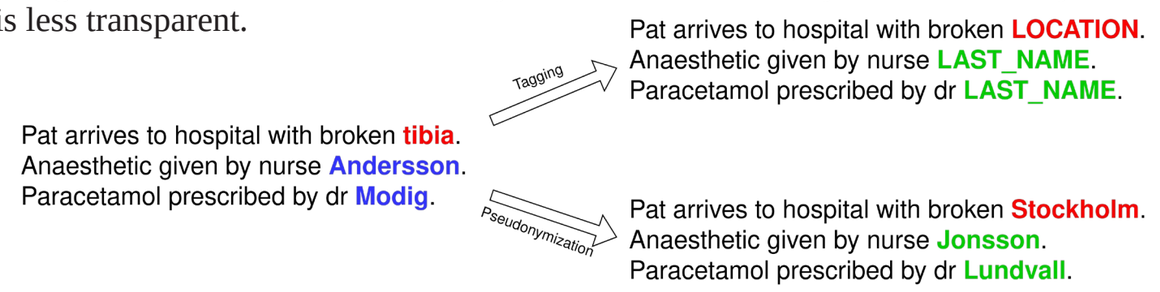


No de-identification system is perfect, and these two approaches have different strengths and weaknesses when considering the **precision-recall tradeoff**.

A de-identifier with **insufficient recall will fail to detect some sensitive entities**. In the example below, we see that pseudonymization allows the missed entity to *hide in plain sight (HIPS)*. Conversely, an attacker reading a tagged version will instantly recognize that "Modig" is a leaked name.



A de-identification system with **insufficient precision will corrupt the data**. In this case, the bone "tibia" is misclassified as a *Location*. Using tags allows a reader to recognize that the model has possibly misclassified the entity, while pseudonymization is less transparent.



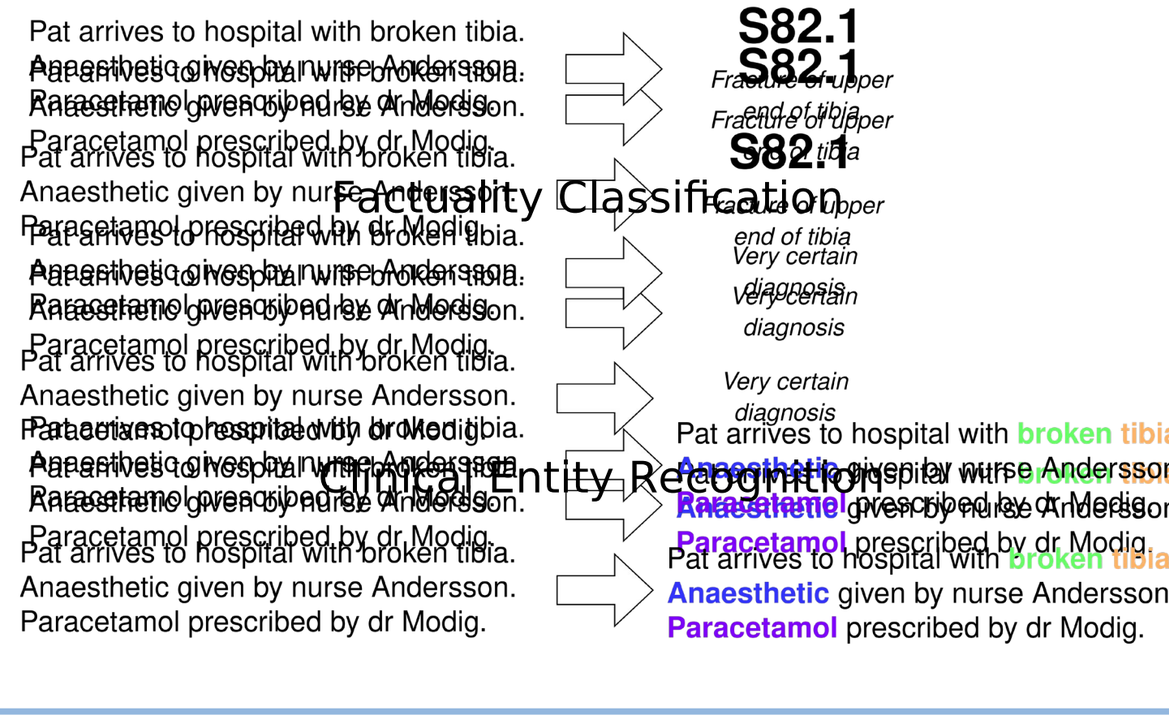
Impact on Machine Learning

We evaluate the impact of data corruption due to insufficient precision by training BERT models on **three clinical datasets**.

1. De-identify each dataset using pseudonymization
2. Train BERT models on real and pseudonymized data
3. Evaluate the performance using held-out test data
4. Perform a Wilcoxon rank-sum test to detect any significant differences

The results from our statistical tests show **no significant deterioration from pseudonymization** in any of the downstream tasks.

Data version	ICD-10	Factuality	Clinical Entity
	Classification	Classification	NER
Unaltered	0.86	0.74	0.85
Pseudonymized	0.86	0.75	0.86



Qualitative Evaluations

We also explored the **percieved quality** of de-identified gastro-intestinal and emergency unit electronic health records from two different hospitals by having experts read large amounts of de-identified texts. This was done twice:

1. With experts in the clinical domain at **Region Stockholm**
2. With a computer scientist at **Lund University**

Participants in the first experiment discussed whether to use pseudonymization or tagging together with a coordinating officer responsible for exporting clinical data. They expressed **a strong preference for tagging** due to the greater transparency of the method.

Apart from the preference for tagging, the qualitative evaluations uncovered several findings. **Abbreviations** were frequently misclassified. For example TBC (tuberculosis) was misclassified as an *Organization*.

Overall, **the participants did not find the texts difficult to read**. This indicates that automatic de-identification also preserves utility for purposes not related to machine learning.

The participants found som instances where the de-identified had failed to tag a sensitive entity. However, the **data leaked was not enough to identify** an individual person.

Conclusions

Pseudonymization does not harm machine learning performance for the three clinical tasks used in this study.

Our evaluators did not find de-identified text difficult to read and found very little sensitive data.

Pseudonymization can make humans uneasy, and our non-technical participants valued the transparency of tagging over the privacy benefits of pseudonymization.

Future Work

This study looked a three clinical tasks: two sequence classification tasks and one token classification task. To strengthen the results, we would like to try pseudonymization on **more datasets**.

The utility for non-machine learning purposes is also very important. Future research could extend our findings by involving a **greater number of participants** in solving a realistic **clinical task** using de-identified electronic health records.

thomas.vakili@dsv.su.se