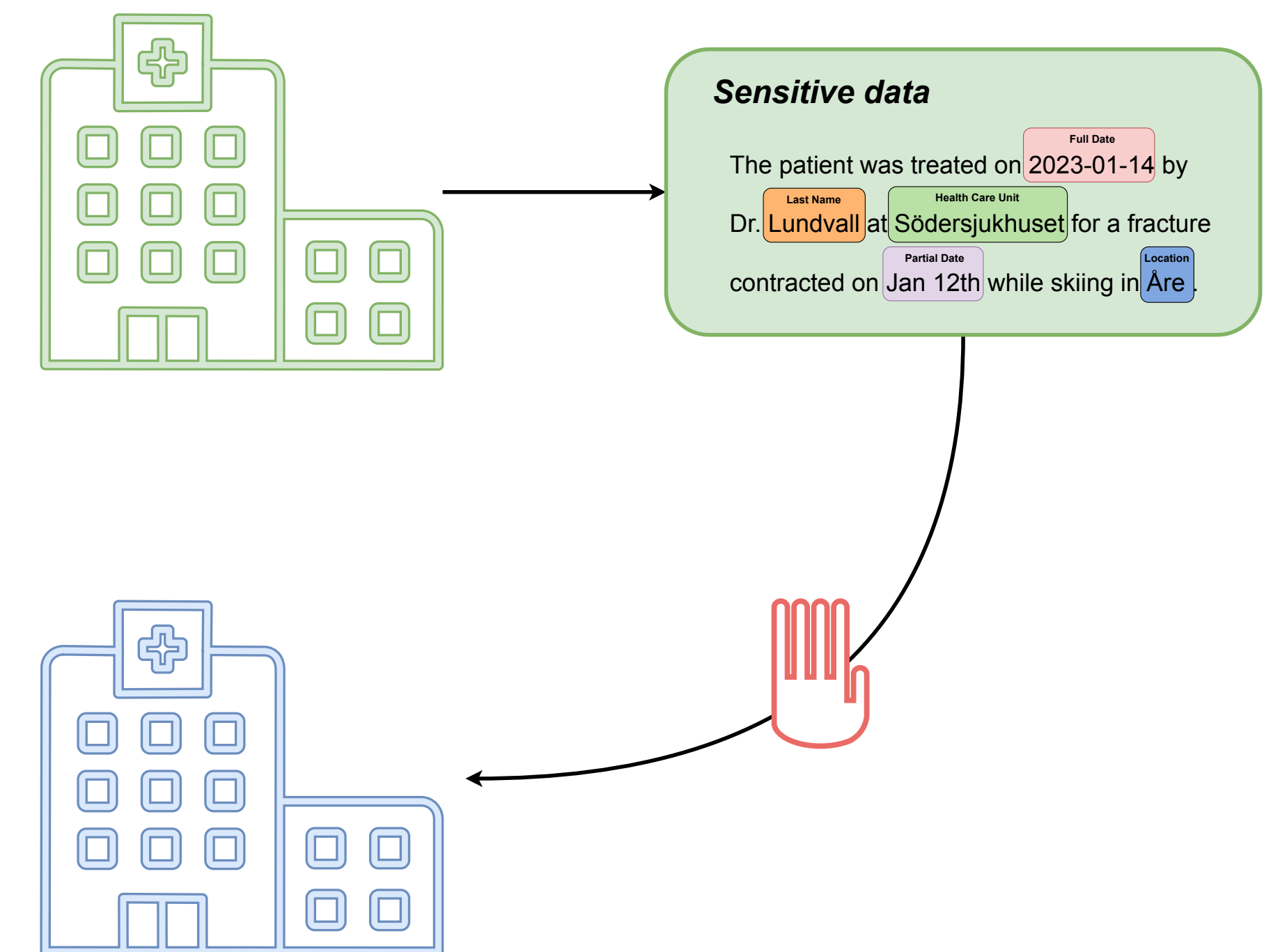


Introduction

NLP is used in sensitive domains — like the clinical domain. These domains suffer from **data scarcity** and **restrictions on sharing**. One strategy is to **automatically de-identify** the data. But datasets used to create de-identifiers are also **sensitive**!

What if we use synthetic data instead?



Experiments

Previous research has focused on:

1. If synthesis works at all
2. What the privacy implications are

We instead systematically study **when** synthetic data are viable — when working with **constrained resources**.

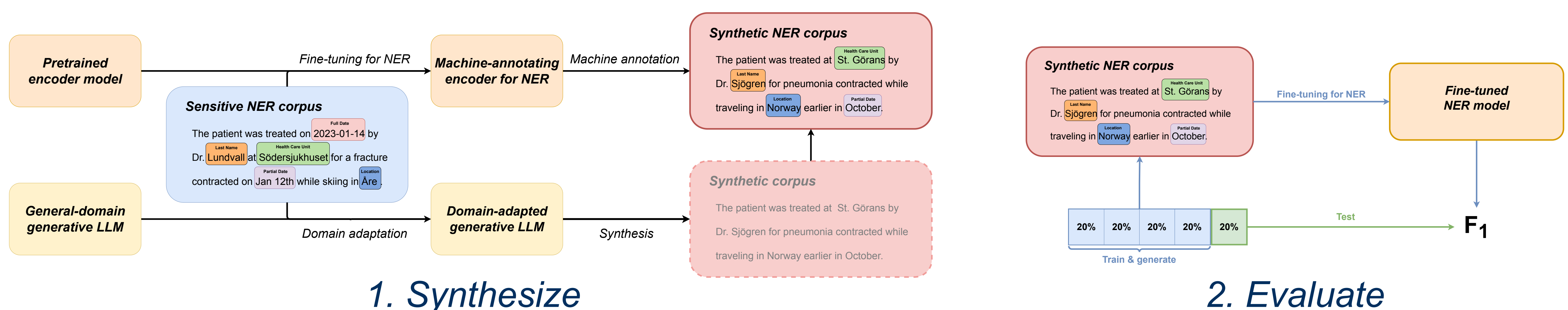
We study the impact of **four constraints**:

1. **Synthesizing varying amounts**
2. **Size of the generative LLM**
3. **Data for domain adaptation**
4. **Data for machine annotation**

Data and models

 MEDDOCAN [1],  FLOR [2], and  roberta-base-bne [3]

 Stockholm EPR PHI Pseudo Corpus (SEPR PHI) [4],  GPT-SW3 [5], and  SweDeClin-BERT [6]



Results

1. Synthetic corpus size

Synthesized amount	SEPR PHI	MEDDOCAN
5%	0.814 ± 0.008	0.938 ± 0.006
100%	0.889 ± 0.009	0.968 ± 0.005
400%	0.896 ± 0.007	0.973 ± 0.003
Gold	0.926 ± 0.005	0.978 ± 0.005

Synthesizing more than the original corpus seems unnecessary.

2. Size of the generative LLM

Model size	SEPR PHI	MEDDOCAN
Small	0.883 ± 0.006	0.973 ± 0.004
Larger	0.896 ± 0.007	0.973 ± 0.003
Gold	0.926 ± 0.005	0.978 ± 0.005

We get far with smaller models! Using the largest models may be overkill.

3. Domain adaptation

% for d.a.	SEPR PHI	MEDDOCAN
0%	0.547 ± 0.178	0.295 ± 0.011
5%	0.873 ± 0.014	0.313 ± 0.032
25%	0.877 ± 0.010	0.970 ± 0.005
50%	0.896 ± 0.007	0.970 ± 0.005
95%	0.896 ± 0.007	0.973 ± 0.003
Gold	0.926 ± 0.005	0.978 ± 0.005

Domain adaption is necessary, but comes with diminishing returns.

4. Machine annotation

% for m.a.	SEPR PHI		MEDDOCAN	
	Gold	Synthetic	Gold	Synthetic
5%	0.707 ± 0.037	0.725 ± 0.039	0.931 ± 0.012	0.942 ± 0.010
25%	0.871 ± 0.010	0.858 ± 0.012	0.967 ± 0.003	0.967 ± 0.004
50%	0.908 ± 0.007	0.889 ± 0.005	0.973 ± 0.004	0.965 ± 0.009
95%	0.926 ± 0.005	0.896 ± 0.007	0.978 ± 0.005	0.973 ± 0.003

Machine annotation is the main factor!

Privacy

% for d.a.	SEPR PHI		MEDDOCAN	
	All 5-grams	Sensitive 5-grams	All 5-grams	Sensitive 5-grams
5%	0.328 ± 0.041	0.233 ± 0.066	0.005 ± 0.000	0.008 ± 0.001
10%	0.216 ± 0.002	0.154 ± 0.016	0.003 ± 0.000	0.006 ± 0.001
25%	0.183 ± 0.015	0.169 ± 0.021	0.003 ± 0.000	0.004 ± 0.000
50%	0.134 ± 0.021	0.141 ± 0.017	0.002 ± 0.000	0.003 ± 0.000
95%	0.122 ± 0.013	0.132 ± 0.010	0.002 ± 0.000	0.003 ± 0.000
0%	0.028 ± 0.002	0.047 ± 0.002	0.001 ± 0.000	0.001 ± 0.000

Privacy implications are unclear — but n-gram overlap is not the best proxy for privacy.

References, paper, and more!

