

# A Short Review of Ethical Challenges in Clinical NLP

Simon Šuster\*<sup>◇</sup>, Stéphan Tulkens\* and Walter Daelemans\*

\*University of Antwerp and <sup>◇</sup>Antwerp University Hospital



## PRIVACY PROTECTION

Clinical notes contain sensitive personal information, which is protected by legal acts like HIPAA and GDPR. Processing clinical data normally requires prior sanitization (e.g. anonymization) and patients' informed consent. Both are associated with some challenges:

### Sanitization

- can compromise data integrity [1]
- is not foolproof
- allows deductive disclosure

### Informed consent

- is difficult to obtain for thousands of data subjects
- informedness is possible only to some degree (especially with difficult-to-interpret models)
- should not represent a barrier to beneficial research

A default opt-in policy with opt-out possibility would reduce the practical burden of obtaining consent for each research project. Patients could allow research in certain domains, resulting in *tiered consent* [2]. In implementing such a policy, public trust is crucial.

## EASIER-TO-ACCESS DATA

- derived data (trained models, text fragments)
- surrogate data (social media for health)
- data from deceased subjects
- veterinary clinical notes

## Selected references

[1] P Duquenoy, C George, and A Solomonides. Considering something 'ELSE': Ethical, legal and socio-economic factors in medical imaging and medical informatics. *Computer Methods and Programs in Biomedicine*, 92(3):227–237, 2008.

[2] BD Mittelstadt and L Floridi. The ethics of big data: current and foreseeable issues in biomedical contexts. *Science and engineering ethics*, 22(2):303–341, 2016.

[3] WW Chapman, PM Nadkarni, L Hirschman, LW D'Avolio, GK Savova, and Ö Uzuner. Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions. *Journal of the American Medical Informatics Association*, 18(5):540–543, 2011.

## Acknowledgments

Work carried out within the Accumulate IWT SBO project, funded by the government agency for Innovation by Science and Technology (IWT). Credits (icons): bmijnlieff, Sergio Aguado Antillaque

## ETHICS IN CLINICAL NLP

## PROGRESS

Difficult access to data is a major impediment to progress in clinical NLP. Sharing raw or annotated data is rarely possible outside of hospital walls. Collaboration efforts are few as a consequence, and reproducing previous results is often impossible [3].

## BIAS AND SOCIAL IMPACT

- data quality
  - set up nationwide clinical registries
  - work on imputing missing values
- sampling bias
  - build corpora for different specialties, narrative types and geographic areas
- reporting bias
  - increase awareness
  - attempt to model bias
- observational bias
  - acquire background (e.g. demographic) information
- dual use
  - keep oversight