Enabling privacy-friendly research with pseudonymized textual reports from disability care

Pim van Oirschot[1,2,*], Lola Vissers[1,2], Joep Tummers[1,2], Maarten Cuypers[1,2]

[1]Radboud university medical center, Department of Primary and Community Care, Nijmegen, The Netherlands,
[2]Academic collaborative Intellectual Disability and Health - Sterker op eigen benen (SOEB), Nijmegen, The Netherlands

*Presenting author, pim.vanoirschot@radboudumc.nl

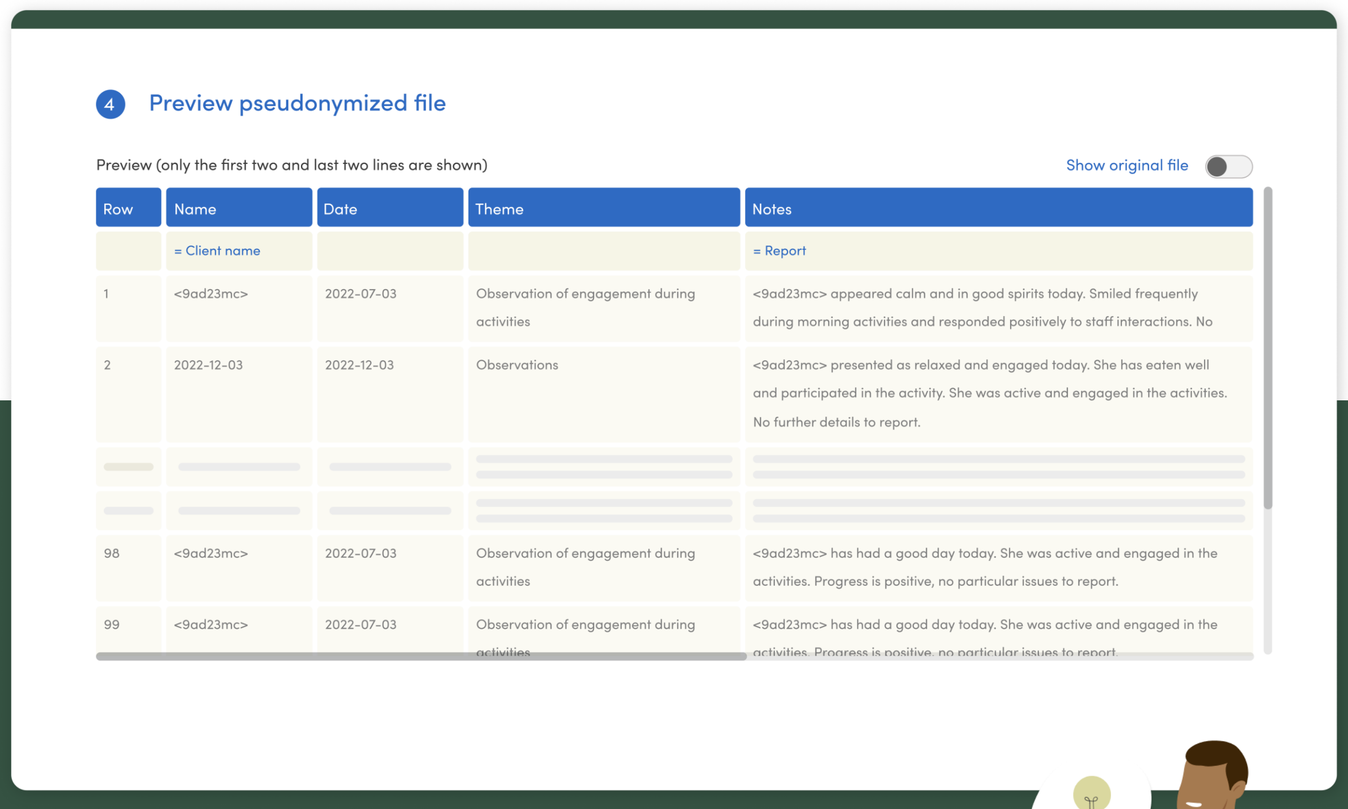
Background

A unique aspect of long-term disability care, compared to other healthcare settings, is the significant time individuals with intellectual disabilities (ID) spend in residential care. This results in a vast amount of (daily) reports on residents' health, well-being and any notable events, which provide valuable insights into the onset and progression of health issues. However, the reports are often unstructured, privacy-sensitive, and voluminous, making large-scale human analysis impractical. Recent advancements in artificial intelligence (AI), particularly large language models, present a potential solution for processing this data, though privacy concerns remain a primary challenge.

Privacytool

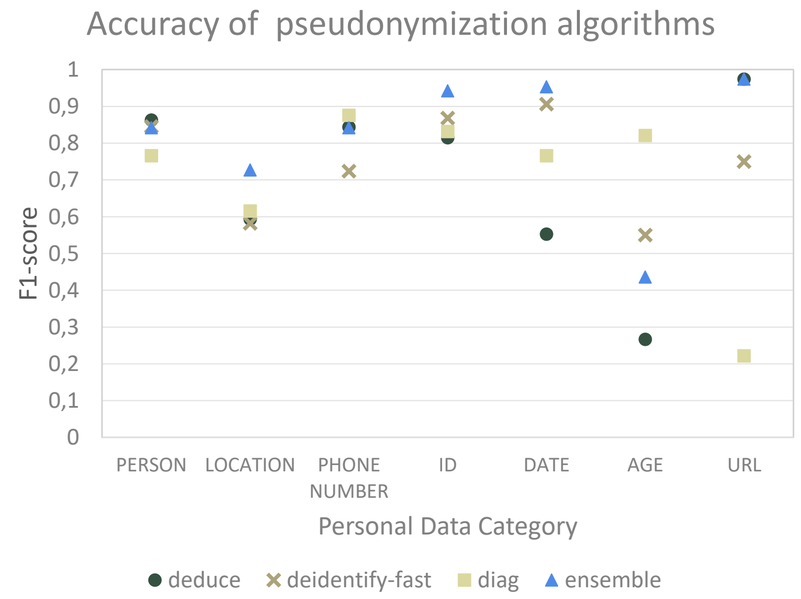
The Privacytool allows safe and fast pseudonymization of textual data, making GDPR-compliant re-use of the data for research and innovation possible, without loss of autonomy and insight in what happens to the data. The user-friendly and easy-to-use application is developed for IT-professionals, application managers and data analysts in healthcare organisations.

Why? Manually removing personally identifiable information (PII) is time-consumable, error-sensitive and increases work pressure. Existing solutions are complex in use and not designed for (ID-)care settings. Consequently, valuable healthcare data is often unexploited for research and innovation.



Effectiveness

We compared the performance of 4 different pseudonymization algorithms: Deduce (Menger et al. 2017), Deidentify-fast (Trienes et al. 2020), Diag (Builtjes et al. 2026) and an algorithm in which the first two of these algorithms are combined, referred to as the Ensemble algorithm. We analysed the algorithm performances on a dataset of approximately 3.400 clinical notes about persons with ID, made by general practitioners (GPs) of 6 GP practices in the Eastern part of the Netherlands, in which 7 categories of personal data were annotated. The accuracy of the algorithms is expressed as the F1-score, which is the harmonic mean of precision and recall. The algorithm performances are shown in the figure below.



"Pseudonymizing data has never been so fast and easy"

"I have complete confidence in this tool, and insight into what happens with the data."

Carmenda, as part of the academic collaborative 'Sterker op eigen benen' within the Radboudumc, is building innovative, data-driven solutions for better and more equitable care for people with intellectual disabilities. Commissioned by the Ministry of Health, Welfare and Sport, they are realizing a secure data infrastructure for research in long-term care.

Radboudumc
Carmenda
STERKER OP EIGEN BENEN

Inclusive & appropriate care