Multi-modal Dataset for pretraining

This section describes the creation of our dataset. We obtained approval from the Hospital Ethics Committees (Belgian registration number B403201523492) to conduct this study, which involves the retrospective analysis of data from patients treated in the orthopedics department at Cliniques Universitaires Saint Luc in Brussels. A lot of effort was provided in order to protect the privacy of patients, following the GDPR[1]. This involved employing anonymization techniques when possible, and resorting to pseudonymization when complete anonymization was not feasible. Section provides a breakdown of the procedures utilized in handling the images, Section delves into the processing of text and reports, and Section explores how the two modalities are combined.

Images preprocessing

Our initial step involved identifying relevant patients by filtering the PACS (Picture Archiving and Communication System) to maintain patients who underwent imaging studies prescribed by the Orthopaedic surgeons of the hospital, and related to osteoarticular conditions. To ensure data anonymization, privacy related metadata were systematically removed and a new random unique identifier was assigned to each individual patient and to each study.

Upon manual examination of the images, it was observed that certain imaging devices included text reports with privacy-sensitive information (e.g., patient's name in dose reports). These reports were saved as images and mixed with X-rays. To address this issue, we employed the EasyOCR[1] framework to extract text from the images with the objective of identifying problematic images. The choice of the framework was based on an initial comparison with Tesseract OCR[2], our investigation indicated that the EasyOCR framework exhibited superior text identification capabilities when applied to our dataset. Subsequent manual inspection of the extracted texts revealed that images containing private patient information exhibited significantly more text than conventional X-ray images, which typically include simple indications such as laterality or patient position. We retained only those images with a character count below 35, a threshold we verified to be conservative. After this last filtering step, no residual patient privacy data was found in the retained images.

This final step yielded a dataset of 947,062 anonymized X-Ray images, grouped in 252,103 studies from 75,600 patients.

Reports preprocessing

The process of extracting medical reports begins with identifying relevant documents in the Electronic Health Record (EHR) for patients previously identified in the PACS (Picture Archiving and Communication System) in Section . A preliminary filtering step is implemented to select reports, focusing on medical analysis results (included radiology reports), consultation reports, hospitalization reports, and operation protocols.

Given that the reports are stored in PDF format, the Pdfminer Python module was employed to extract text while simultaneously filtering out headers and footers containing administrative information based on the hospital's specific templates.

Despite these precautions, the extracted texts still contain protected health information (PHI), such as the patient's name and date of birth. Manual elimination of this information from the large volume of documents would be impractical. Consequently, the decision was made to create surrogates documents[2] that keep the useful information from the originals but with fictitious PHIs. DEDUCE[3], a rule-based tool designed for identifying PHIs in Dutch medical texts, was adapted to work with French for this purpose[3].

To assess the performance of our modified DEDUCE [3] method used with french documents, 100 reports were randomly selected in the dataset and manually annotated for patient names, person names, locations, institutions, dates, ages, id numbers, phone numbers and url/e-mails. The proposed method was then compared with the annotations, the precision, recall and F1-score were computed for each PHI with results available in Table 1

After identifying PHI in the documents, a systematic replacement was carried out using fictitious but contextually coherent substitute names for individuals, locations, and health institutions. To ensure authenticity in the surrogate data, last names and first names were sourced from the Belgian *Direction générale Statistique* (StatBel)[4] and the French *Institut national de la statistique et des études économiques* (INSEE). For health institution names, lists of nursing homes and hospitals from the Belgian *Institut national d'assurance maladie invalidité* (INAMI) were used. A list of all

[1] `https://github.com/jaidedai/easyocr`

[2] `https://github.com/tesseract-ocr/tesseract`

[3] available at `https://github.com/aenglebert/deduced`

[4] https://statbel.fgov.be/fr/themes/population/noms-et-prenoms

	Count	Precision	Recall	F1-score
Patient names	132.0	0.96	1.00	0.98
Person names	100.0	0.66	0.94	0.78
Locations	52.0	0.98	0.86	0.92
Institutions	23.0	0.76	0.83	0.79
Dates	427.0	0.99	0.98	0.98
Ages	39.0	0.86	0.97	0.91
ID numbers	19.0	0.95	1.00	0.97
Phone numbers	47.0	0.98	0.93	0.96
URL/e-mails	13.0	1.00	1.00	1.00

Table 1: PHI identification metrics

cities in Belgium provided addresses. To further protect privacy, a random shift (between -1000 and +1000 days) was applied to dates, while phone numbers, URLs, and email addresses were simply removed.

The creation of surrogates using the above pseudonymization process is crucial to ensure adherence to privacy regulations, notably GDPR[1]. Furthermore, the incapacity to categorize pseudonymized data as entirely anonymized imposes a limitation on our capacity to publicly share the dataset.

This complete process yielded a dataset of 1,837,427 surrogates of medical documents.

Images and reports pairing

To conduct contrastive vision-language pretraining, pairs of image and text data need to be made. The documents were restricted to radiology reports and aligned with X-ray studies based on their dates (before pseudonymization). In cases where multiple studies and X-ray reports exist for a specific date, we align them in chronological order while disregarding ambiguous instances that necessitate manual examination.

The resulting number of paired studies amounts to 219,675, corresponding to 789,397 individual X-ray images in total. As a study may consist of one or several images, there exists a one-to-many relationship between the reports and the images.

It's worth noting that only pairs of radiographs with associated reports were utilized, which represents only a fraction of the available reports presented in Section . For instance, reports of orthopedic consultations were excluded since they are not directly paired to the images, even though they may contain valuable related information not present in radiologists' reports.

References

[1] Protection Regulation. Regulation (eu) 2016/679 of the european parliament and of the council. *Regulation (eu)*, 679:2016, 2016.

[2] David Carrell, Bradley Malin, John Aberdeen, Samuel Bayer, Cheryl Clark, Ben Wellner, and Lynette Hirschman. Hiding in plain sight: use of realistic surrogates to reduce exposure of protected health information in clinical text. *Journal of the American Medical Informatics Association*, 20(2):342–348, 2013.

[3] Vincent Menger, Floor Scheepers, Lisette Maria van Wijk, and Marco Spruit. Deduce: A pattern matching method for automatic de-identification of dutch medical text. *Telematics and Informatics*, 35(4):727–736, 2018.