Learning structures of the French clinical language: development and validation of word embedding models using 21 million clinical reports from electronic health records

Basile Dura  
Innovation and Data, IT Department  
Assistance Publique – Hôpitaux de Paris  
basile.dura-ext@aphp.fr

Charline Jean  
Innovation and Data, IT Department  
Assistance Publique – Hôpitaux de Paris  
charline.jean-ext@aphp.fr

Xavier Tannier, PhD  
Sorbonne Université, Inserm, Université Sorbonne Paris Nord  
Laboratoire d’Informatique Médicale et d’Ingénierie des Connaissances pour la e-Santé  
xavier.tannier@sorbonne-universite.fr

Alice Calliger  
Innovation and Data, IT Department  
Assistance Publique – Hôpitaux de Paris  
alice.calliger-ext@aphp.fr

Romain Bey, PhD  
Innovation and Data, IT Department  
Assistance Publique – Hôpitaux de Paris  
romain.bey@aphp.fr

Antoine Neuraz, MD, PhD  
Sorbonne Université, Inserm, Centre de Recherche des Cordeliers  
Biomedical Informatics Department, Hôpital Necker-Enfants Malades  
Assistance Publique – Hôpitaux de Paris  
Team HeKA, INRIA  
antoine.neuraz@aphp.fr

Rémi Flicoteaux, MD, PhD  
Medical Information Department  
Assistance Publique – Hôpitaux de Paris  
remi.flicoteaux@aphp.fr

Abstract

Background

Clinical studies using real-world data may benefit from exploiting clinical reports, a particularly rich albeit unstructured medium. To that end, natural language processing can extract relevant information. Methods based on transfer learning using pre-trained language models have achieved state-of-the-art results in most NLP applications; however, publicly available models lack exposure to speciality-languages, especially in the medical field.

Objective

We aimed to evaluate the impact of adapting a language model to French clinical reports on downstream medical NLP tasks.

Preprint. Work in progress.
Methods

We leveraged a corpus of 21M clinical reports collected from August 2017 to July 2021 at the Greater Paris University Hospitals (APHP) to produce two CamemBERT architectures on speciality language: one retrained from scratch and the other using CamemBERT as its initialisation. We used two French annotated medical datasets to compare our language models to the original CamemBERT network, evaluating the statistical significance of improvement with the Wilcoxon test.

Results

Our models pretrained on clinical reports increased the average F1-score on APMed (an APHP-specific task) by 3 percentage points to 91%, a statistically significant improvement. They also achieved performance comparable to the original CamemBERT on QUAERO. These results hold true for the fine-tuned and from-scratch versions alike, starting from very few pre-training samples.

Conclusions

We confirm previous literature showing that adapting generalist pre-train language models such as CamenBERT on speciality corpora improves their performance for downstream clinical NLP tasks. Our results suggest that retraining from scratch does not induce a statistically significant performance gain compared to fine-tuning.

Keywords

Natural language processing, electronic health records, clinical data, word embeddings

Abbreviations

| Abbreviation | Meaning |
|--------------|---------|
| APHP         | Assistance Publique – Hôpitaux de Paris (Greater Paris University Hospitals) |
| RWD          | Real-World Data |
| EHR          | Electronic Health Record |
| ML           | Machine Learning |
| NLP          | Natural Language Processing |
| BERT         | Bidirectional Encoder Representations from Transformers |
| CDW          | Clinical Data Warehouse |
| EDS          | Entrepôt des Données de Santé, APHP’s CDW |
| NER          | Named Entity Recognition |
1 Introduction

Medical studies using real-world data (RWD) may benefit from exploiting clinical reports, a rich albeit unstructured part of electronic health records (EHR) collected during care episodes. These reports may contain relevant information that is scarce in structured EHR; by some estimates, up to 80% of entities found in clinical reports are absent from other media [1].

In this context and given the scale of data to analyse, natural language processing (NLP) methods are needed to extract meaningful medical information from this unstructured medium, and help address challenges such as automatic detection of adverse drug reaction, clinical trial eligibility or identification of temporal associations [2].

Initially bound to purely rule-based methods, the NLP field has been shifting towards machine learning (ML) algorithms that can detect patterns automatically. The most recent techniques rely on a first processing stage to represent free-text data into machine-readable input using models known as word embeddings, whose goal is to provide a vector representation that conveys as much semantic and syntactic information as possible.

Methods such as GloVe [3], Word2Vec [4] or fastText [5] can learn meaningful static representations for words, but novel embeddings algorithms like ELMo [6] and XLNet [7] have since been proposed to include contextual information in the embeddings. Introduced by Devlin et al, the Bidirectional Encoder Representations from Transformers (BERT) [8] proposes an efficient method outputting rich representations for words based on their context that consistently demonstrates state-of-the-art performance in most NLP applications. In French, FlauBERT [9] and CamemBERT [10] are trained on general-purpose French-language documents crawled from the Internet.

Using transfer-learning, such pre-trained models can serve as a basis for a variety of NLP tasks. In the context of a clinical data warehouse (CDW), an ecosystem of researchers and clinicians may rely on a shared pre-trained language model, and fine-tune it on their specific tasks.

Previous work has shown that using specialty language for training BERT-based models can widely increase performances [11, 12]: specialty languages and clinical reports in particular follow a distinct syntax and vocabulary, such that training a model to learn these specificities can represent an advantage. Moreover, Martin et al [10] have determined that a model trained on a carefully selected subcorpus could achieve comparable results despite using less than 10% of the original training data.

In this work, we leverage the CDW of the Greater Paris University Hospitals (Entrepôt des Données de Santé, EDS) to confirm whether there is significant advantage to using a word embedding model specifically trained on French clinical reports for clinical NLP tasks, and address the following questions:

1. Is there an advantage to retraining from scratch, as opposed to fine-tuning an existing model, given the excess computational toll and environmental footprint?
2. How many training steps and examples are necessary to learn useful knowledge about the specialty language?

2 Methods

This study followed the RECORD reporting guideline [13]; the checklist is available in the appendix A.3.

2.1 Dataset

The EDS contains data collected in the EHR of 39 hospitals from the greater Paris area and relative to 11M patients, including 80M clinical text reports.

The training corpus for this work consists of clinical reports gathered between August 2017 and July 2021. Documents are pseudonymised [14] to preserve privacy, by replacing directly identifying entities with fake entities.

Reports containing less than 20 characters were removed and the corpus was resampled to limit the influence of over-represented report types (e.g. prescriptions, consultation or imaging reports).

We pre-processed selected documents with EDS-NLP [15] and spaCy [16] by removing textual pollution, such as administrative information shared by a large proportion of the reports, which could
skew the distribution seen by the model (see appendix B for details). Although clinical reports may contain other forms of duplicate information [17], we remained conservative and did not push the pre-processing further.

This study was authorised by the EDS institutional review board (IRB 00011591, project CSE-19-20). The EDS is approved by the French national data protection agency (CNIL, decision 1980120).

2.2 Models
We used the architecture of CamemBERT-base for all the experiments and compared two training strategies: fine-tuning or retraining it “from scratch”. In what follows, we focus on two models which we compare to the freely-accessible CamemBERT-base model:

1. EDS-fine-tuned, a version fine-tuned on our clinical documents but using the original weights as the initialisation.

2. EDS-from-scratch, a version trained from the ground up. This approach lets us retrain a domain-specific tokenizer.

Since most reports go over the 512-token limit imposed by the BERT architecture, we decided to segment documents into 128-token-long sequences.

2.3 Training
The dataset was split into training (19.6M documents) and validation subsets (1M documents). We pre-trained seven models:

EDS-from-scratch was initialised with random weights. We followed CamemBERT’s training procedure, and ran the experiment for twelve full epochs, totalling more training steps to compensate for the smaller batch size.

EDS-fine-tuned used CamemBERT-base as its initialisation point, and was trained for one epoch on the full dataset. We also trained five other versions to estimate the impact of the number of training samples, using 100K, 300K, 1M, 3M, 10M and 21M documents. We sampled the documents uniformly from the training dataset described earlier. Every model used in this comparison was trained with the same number of steps, corresponding to one full epoch on 21M documents.

We relied on the transformers library by HuggingFace [18], Pytorch [19] and Pytorch-Lightning [20] for our code base.

2.4 Validation
2.4.1 Intrinsic validation
We validated our models using their perplexity measured on a held out validation set, and investigated the influence of the tokenization step. We compared the distribution of tokenized sequence lengths to evaluate whether the tokenizer had learnt some useful information about the clinical vocabulary.

2.4.2 Extrinsic validation
We validated our models on two named entity recognition (NER) tasks, see appendix D for detail:

- APMed [21, 22]: a corpus for extracting drug related information in clinical reports in French.
- QUAERO [23], a compilation of two French corpora annotated to ten types of clinical entities:
  - EMEA includes long texts containing information on marketed drugs from the European Medicines Agency;
  - MEDLINE regroups titles of research articles.

Every task was framed as a token classification problem, using IOB2 notation. We used the same architecture for every test, and trained the models in depth during fine-tuning on the downstream task. We added a classification head consisting of:

- A fully-connected hidden layer with ReLU activation;
- A fully-connected output layer.
Experiments were reproduced ten times with different random initialisations, to obtain a confidence interval around the results. We used seqeval [24] to compute the micro-averaged F1-score, and we evaluated the statistical significance using a Wilcoxon signed-rank test. All tests were 2-sided and p-values were considered statistically significant when lower than 0.05.

3 Results

3.1 Training

Training *EDS-from-scratch* on 21M reports for 12 epochs took 25 days on 8 Tesla V100 GPUs. Each version of EDS-CamemBERT-fine-tuned was trained for 2 days on the same setup. Total carbon emissions were estimated using the MachineLearning Impact calculator [25] at respectively 10 and 110 kgCO2eq for each version of *EDS-fine-tuned* and *EDS-from-scratch*.

3.2 Intrinsic validation

The median number of tokens needed to represent one document was 1724 for CamemBERT’s original tokenizer, and 1500 using our EDS-specific tokenizer.

The models’ loss on unseen data was still decreasing at the end of training (see appendix C).

3.3 Extrinsic validation

3.3.1 Comparison with CamemBERT-base

We compared the transfer-learning capabilities of our models with CamemBERT-base, and recapitulated the results in Table 1.

Table 1: Performance of our models on multiple extrinsic tasks compared to CambemBERT-base. Results are formatted as mean (+/-std).

| Model                   | APMed (F1-score) | EMEA | QUAERO (F1-score) | Total          |
|-------------------------|------------------|------|-------------------|----------------|
| *EDS-fine-tuned*        | .902 (±0.003)*   | .729 (±0.008) * | .597 (±0.007) *  | .655 (±0.007) |
| *EDS-from-scratch*      | .908 (±0.005)*   | .693 (±0.012) * | .601 (±0.01) *   | .642 (±0.007) * |
| CamemBERT-base          | .866 (±0.007)    | .737 (±0.006) * | .584 (±0.004)    | .651 (±0.004) * |
| Best QUAERO [26] model  |                 |      | .749              | .698           |

The results on APMed, the EDS-specific dataset, show a statistically significant improvement when using re-trained language models (maximum \( p = 4 \cdot 10^{-3} \)). However, the difference between *EDS-fine-tuned* and *EDS-from-scratch* is not significant (\( p = 0.1 \)).

On QUAERO, *EDS-fine-tuned* performed better than *EDS-from-scratch* overall (\( p = 4 \cdot 10^{-3} \)).

3.3.2 Impact of the number of training steps and training examples

Figure 1 investigates the impact of the number of training steps performed on *EDS-from-scratch* on its performance on the APMed dataset.

Moreover, Table 2 shows the impact of the number of examples when fine-tuning CamemBERT-base.

Table 2: Performance on APMed with respect to the number of the training examples

| Training examples | APMed (F1-score) |
|-------------------|------------------|
| 100K              | .900 (±0.004)    |
| 300K              | .900 (±0.006)    |
| 1M                | .904 (±0.004)    |
| 3M                | .904 (±0.003)    |
| 10M               | .901 (±0.004)    |
| 21M (EDS-fine-tuned) | .902 (±0.003)  |
4 Discussion

In this work, we investigated the impact of pre-training a BERT-based language model on clinical reports by comparing the performance on two medical down-stream NER tasks.

Our results on the APMed corpus confirm previous literature findings that pre-training on speciality language leads to a statistically significant performance improvement.

We also evaluated our models outside the EDS context to check for non-regression, namely on the QUAERO corpora, and observed that EDS-fine-tuned fared comparably to CamemBERT-base on this non-clinical dataset. However, BERT-based methods achieved much lower performance than QUAERO’s rule-based laureates on the MEDLINE subcorpus. We posit that the fine-tuned model retains sufficient general language knowledge to keep relatively high performances on a non-clinical task, and that the short length of MEDLINE’s examples hinders contextual methods, although further validation is needed.

What is more, even on the EDS-specific APMed task, we found no significant advantage to using a model pre-trained from scratch compared to fine-tuning a generalist model. Although the new tokenizer is better apt to capture the medical vocabulary (the median token sequence length on EDS reports drops by 15% when using the EDS-specific tokenizer), the overall performance on EDS-specific NER tasks is similar (Table 2). The impact in terms of inference time is to be evaluated in a future work.

Moreover, our study on the impact of the number of training steps and training examples suggests that relatively few samples are required to reach good test performance. Indeed, our model fine-tuned on 100K samples was able to reach performances similar to models trained on the full 21M-report dataset, at a fraction of the computational and environmental toll. This finding opens up a world of possibilities for smaller-scale CDW, which can adapt general-language models to their distribution at a relatively low cost.

5 Conclusion

In this work, we propose EDS-CamemBERT, a language modelling neural network adapted to the context of French-speaking clinical data warehouses. We show that fine-tuning state-of-the-art language models on clinical reports improves performances on downstream speciality tasks. We demonstrate that in this setting, training a model from scratch bears little advantages to fine-tuning a general-language model, despite providing a better tuned tokenizer. Finally, we provide evidence that very few samples are needed to achieve a statistically significant gap in performance.
Acknowledgement

We thank the Greater Paris University Hospitals CDW for its support and the realisation of data management and data curation tasks.

Conflicts of Interest statement

None declared.

Data and code sharing

Access to the Clinical Data Warehouse’s raw data can be granted following the process described on its website: eds.aphp.fr. A prior validation of the access by the local IRB is required. In the case of non-APHP researchers, the signature of a collaboration contract is also mandatory.

The source code used for building the dataset and training the models is freely available on APHP’s Github account, distributed under a 3-Clause BSD licence. It is documented, versioned and citable through Zenodo.

Funding

This study has been supported by grants from the APHP Foundation.

References

[1] Preethi Raghavan, James L. Chen, Eric Fosler-Lussier, and Albert M. Lai. How essential are unstructured clinical narratives and information fusion to clinical trial recruitment? AMIA Summits on translational science proceedings. 2014:218–223, April 2014. ISSN 2153-4063. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4333685/.

[2] Kory Kreimeyer, Matthew Foster, Abhishek Pandey, Nina Arya, Gwendolyn Halford, Sandra F Jones, Richard Forshee, Mark Walderhaug, and Taxiarchis Botsis. Natural language processing systems for capturing and standardizing unstructured clinical information: A systematic review. Journal of biomedical informatics, 73:14–29, September 2017. ISSN 15320464. doi: 10.1016/j.jbi.2017.07.012. URL https://linkinghub.elsevier.com/retrieve/pii/S1532046417301685.

[3] Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global Vectors for Word Representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1162. URL https://www.aclweb.org/anthology/D14-1162.

[4] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. arXiv:1301.3781 [cs], September 2013. URL http://arxiv.org/abs/1301.3781. arXiv: 1301.3781.

[5] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching Word Vectors with Subword Information. arXiv:1607.04606 [cs], June 2017. URL http://arxiv.org/abs/1607.04606. arXiv: 1607.04606.

[6] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. arXiv:1802.05365 [cs], March 2018. URL http://arxiv.org/abs/1802.05365. arXiv: 1802.05365.

[7] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. XLNet: Generalized Autoregressive Pretraining for Language Understanding. arXiv:1906.08237 [cs], January 2020. URL http://arxiv.org/abs/1906.08237. arXiv: 1906.08237.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805 [cs], May 2019. URL http://arxiv.org/abs/1810.04805. arXiv: 1810.04805.
[22] Jordan Jouffroy, Sarah F Feldman, Ivan Lerner, Bastien Rance, Anita Burgun, and Antoine Neuraz. Hybrid Deep Learning for Medication-Related Information Extraction From Clinical Texts in French: MedExt Algorithm Development Study. *JMIR Medical Informatics*, 9(3): e17934, March 2021. ISSN 2291-9694. doi: 10.2196/17934. URL http://medinform.jmir.org/2021/3/e17934/.

[23] Aurélie Névéol, Cyril Grouin, Jeremy Leixa, Sophie Rosset, and Pierre Zweigenbaum. The Quaero French Medical Corpus: A Resource for Medical Entity Recognition and Normalization. page 7.

[24] Hiroki Nakayama. seqeval: A Python framework for sequence labeling evaluation, 2018. URL https://github.com/chakki-works/seqeval. original-date: 2018-02-15T00:02:36Z.

[25] Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. Quantifying the Carbon Emissions of Machine Learning. *arXiv:1910.09700 [cs]*, November 2019. URL http://arxiv.org/abs/1910.09700. arXiv: 1910.09700.

[26] Aurelie Neveol, K Bretonnel Cohen, Cyril Grouin, Thierry Hamon, Thomas Lavergne, Liadh Kelly, Lorraine Goeuriot, Gregoire Rey, Aude Robert, Xavier Tannier, and Pierre Zweigenbaum. Clinical Information Extraction at the CLEF eHealth Evaluation lab 2016. page 15.
A Reproducibility

A.1 Code availability
The codebase used to train and evaluate our models is saved on Zenodo and accessible on APHP’s Github account, under an open-source (3-Clause BSD) licence.

A.2 Data sharing
Trained model weights and the training dataset can be accessed after review by APHP’s institutional review board.

A.3 The RECORD statement checklist
B Training dataset

B.1 Reglementary considerations
This study has been approved by the institutional review board of the EDS (IRB 00011591, decision CSE 19-20). The EDS, authorised by the French National Data Protection Commission (CNIL, decision 1980120), ensures patients’ information through a transparency portal in accordance with the European General Data Protection Regulation. Subjects that objected to the reuse of their data were excluded from this study in accordance with French legislation.

B.2 Flow chart
Figure 2 presents a flowchart describing the cohort definition.

Figure 2: Flow chart detailing the dataset creation

B.3 Sub-sampling
Figures 3 and 4 show the impact of the sampling method on the distribution of document types. Table 4 proposes a description of the meta-categories retained in this study.

B.4 Removal of administrative pollution
We used EDS-NLP [15], developed at APHP, to tag and remove textual pollution. Such items include the systematic mention of patients’ rights regarding the use of their personal data. EDS-NLP builds on spaCy [16] to propose open-source pipeline components specifically designed for French-language CDW.
Figure 3: Distribution of meta-types (similar types are regrouped) before hierarchical sampling

Figure 4: Distribution of meta-types (similar types are regrouped) after hierarchical sampling

B.5 Description
Table 5 presents the age and gender distribution of patients within the cohort.

C Training
Figure 5 presents the validation loss during EDS-from-scratch training.
D Validation tasks

D.1 APMed
The APMed corpus is a named entity recognition task for the detection and normalisation of drug mentions with their posology. It was developed using clinical reports generated at APHP.

D.2 QUAERO
The QUAERO dataset, or QUAERO French Medical Corpus, is a compilation of two corpora annotated for medical entities:

- EMEA includes long texts containing information on marketed drugs from the European Medicines Agency;
- MEDLINE regroups titles of research articles.

The documents were annotated for clinical entities defined by a subset of the UMLS Semantic Groups: Anatomy, Chemical and Drugs, Devices, Disorders, Geographic Areas, Living Beings, Objects, Phenomena, Physiology, Procedures.
| Item No | STROBE items | Location in manuscript where items are reported | RECORD items | Location in manuscript where items are reported |
|---------|--------------|-------------------------------------------------|--------------|------------------------------------------------|
| **Title and abstract** | | | | |
| 1 | (a) Indicate the study’s design with a commonly used term in the title or the abstract (b) Provide in the abstract an informative and balanced summary of what was done and what was found | a) Abstract - Methods b) Abstract - Methods and Results | RECORD 1.1: The type of data used should be specified in the title or abstract. When possible, the name of the databases used should be included. RECORD 1.2: If applicable, the geographic region and timeframe within which the study took place should be reported in the title or abstract. RECORD 1.3: If linkage between databases was conducted for the study, this should be clearly stated in the title or abstract. | 1.1) Abstract - Methods 1.2) Abstract - Methods 1.3) N/A |
| **Introduction** | | | | |
| 2 | Explain the scientific background and rationale for the investigation being reported | | | |
| **Objectives** | 3 | State specific objectives, including any prespecified hypotheses | | |
| **Methods** | | | | |
| **Study Design** | 4 | Present key elements of study design early in the paper | Sections 2.1, 2.2, 2.3, 2.4 | |
| **Setting** | 5 | Describe the setting, locations, and relevant dates, including periods of recruitment, exposure, follow-up, and data collection | Section 2.1 | |
| **Participants** | 6 | (a) Cohort study - Give the eligibility criteria, and the sources and methods of selection of participants. Describe methods of follow-up - Case-control study - Give the eligibility criteria, and the sources and methods of case ascertainment and control selection. Give the rationale for the choice of cases and controls - Cross-sectional study - Give the eligibility criteria, and the sources and methods of selection of participants (b) Cohort study - For matched studies, give matching criteria and number of exposed and unexposed - Case-control study - For matched studies, give matching criteria and number of controls per case | a) Section 2.1 b) NA | RECORD 6.1: The methods of study population selection (such as codes or algorithms used to identify subjects) should be listed in detail. If this is not possible, an explanation should be provided. RECORD 6.2: Any validation studies of the codes or algorithms used to select the population should be referenced. If validation was conducted for this study and not published elsewhere, detailed methods and results should be provided. RECORD 6.3: If the study involved linkage of databases, consider use of a flow diagram or other graphical display to demonstrate the data linkage process, including the number of individuals with linked data at each stage. | 6.1) Section 2.1 6.2) Section 2.1 6.3) NA |
| **Variables** | 7 | Clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers. Give diagnostic criteria, if applicable. | NA | RECORD 7.1: A complete list of codes and algorithms used to classify exposures, outcomes, confounders, and effect modifiers should be provided. If these cannot be reported, an explanation should be provided. | NA |
| **Data sources/measurement** | 8 | For each variable of interest, give sources of data and details of methods of assessment (measurement). Describe comparability of assessment methods if there is more than one group | Section 2.1 | |
| **Bias** | 9 | Describe any efforts to address potential sources of bias | Section 2.1 | |
| **Study size** | 10 | Explain how the study size was arrived at | Section 2.1 | |
| **Quantitative variables** | 11 | Explain how quantitative variables were handled in the analysis. If applicable, describe which groupings were chosen, and why | NA | |
| **Statistical methods** | 12 | (a) Describe all statistical methods, including those used to control for confounding (b) Describe any methods used to examine subgroups and interactions (c) Explain how missing data were addressed (d) Cohort study - If applicable, explain how loss to follow-up was addressed Case-control study - If applicable, explain how matching of cases and controls was addressed Cross-sectional study - If applicable, describe analytical methods taking account of sampling strategy (e) Describe any sensitivity analyses | a) Sections 2.3, 2.4 b) Section 2.4 c) Section 2.1 d) NA e) Section 2.4 | |

14
| Item No | STROBE items | Location in manuscript where items are reported | RECORD items | Location in manuscript where items are reported |
|--------|--------------|-----------------------------------------------|--------------|-----------------------------------------------|
| 11     | Data access and cleaning methods | RECORD 12.1: Authors should describe the extent to which the investigators had access to the database population used to create the study population. RECORD 12.2: Authors should provide information on the data cleaning methods used in the study. | 12.1) Section 2.1 12.2) Section 2.1 |
| 12     | Linkage | RECORD 12.3: State whether the study included person-level, institutional-level, or other data linkage across two or more databases. The methods of linkage and methods of linkage quality evaluation should be provided. | 12.3) NA |
| 13     | Participants | (a) Report the numbers of individuals at each stage of the study (e.g., numbers potentially eligible, examined for eligibility, confirmed eligible, included in the study, completing follow-up, and analysed) (b) Give reasons for non-participation at each stage. (c) Consider use of a flow diagram | a) Section 3.1 b) Appendix B.2 c) Appendix B.2 | RECORD 13.1: Describe in detail the selection of the persons included in the study (i.e., study population selection) including filtering based on data quality, data availability and linkage. The selection of included persons can be described in the text and/or by means of the study flow diagram. | 13.1) Section 3.1 and appendix B.2 |
| 14     | Descriptive data | (a) Give characteristics of study participants (e.g., demographic, clinical, social) and information on exposures and potential confounders (b) Indicate the number of participants with missing data for each variable of interest (c) Cohort study - summarise follow-up time (e.g., average and total amount) | a) Appendices B.3, B.4, B.5 b) Appendices B.3, B.4 c) NA | 13.1) Section 3.1 and appendix B.2 |
| 15     | Outcome data | Cohort study - Report numbers of outcome events or summary measures over time Case-control study - Report numbers in each exposure category, or summary measures of exposure Cross-sectional study - Report numbers of outcome events or summary measures | NA | 13.1) Section 3.1 and appendix B.2 |
| 16     | Main results | (a) Give unadjusted estimates and, if applicable, confounder-adjusted estimates and their precision (e.g., 95% confidence interval). Make clear which confounders were adjusted for and why they were included (b) Report category boundaries when continuous variables were categorized (c) If relevant, consider translating estimates of relative risk into absolute risk for a meaningful time period | a) Section 3.3 b) NA c) NA | 13.1) Section 3.1 and appendix B.2 |
| 17     | Other analyses | Report other analyses done — e.g., analyses of subgroups and interactions, and sensitivity analyses | Section 3.5 | 13.1) Section 3.1 and appendix B.2 |
Discussion

**Key results**
18 Summarise key results with reference to study objectives
Section 4

**Limitations**
19 Discuss limitations of the study, taking into account sources of potential bias or imprecision. Discuss both direction and magnitude of any potential bias
Section 4

**Interpretation**
20 Give a cautious overall interpretation of results considering objectives, limitations, multiplicity of analyses, results from similar studies, and other relevant evidence
Section 4

**Generalisability**
21 Discuss the generalisability (external validity) of the study results
Section 4

**Other Information**

**Funding**
22 Give the source of funding and the role of the funders for the present study and, if applicable, for the original study on which the present article is based
Funding

**Accessibility of protocol, raw data, and programming code**

**Table 4: Description of the meta-categories**

| Meta-category                  | Description                      |
|-------------------------------|----------------------------------|
| CR Consultation               | Consultation reports            |
| Prescription                  | Prescriptions                    |
| CR Imagerie                   | Imaging reports                  |
| CRH                            | Hospitalisation reports          |
| CR Urgences                   | Emergency room reports           |
| Lettre                        | Lettres                          |
| CR Anapath                    | Anatomopathology reports         |
| CR Opératoire                 | Operation reports                |
| CR Examen technique médical   | Technical examination reports    |
| Observation médicale          | Medical observations             |
| Observation paramédicale      | Paramedical observations         |
| Autre                         | Any documents not included in the other categories |

**Table 5: Distribution of the population within the selected dataset. This is a per-document view, and the age of the patient is calculated relative to the creation date of each individual document.**

| Age distribution (%) | Female  | Male  | Total   |
|----------------------|---------|-------|---------|
| 0-10                 | 761 486 (44%) | 972 365 (56%) | 1 733 851 (8%) |
| 10-20                | 592 199 (48%)  | 642 659 (52%) | 1 234 858 (6%) |
| 20-30                | 939 914 (56%)  | 751 753 (44%) | 1 691 667 (8%) |
| 30-40                | 1 344 079 (59%) | 934 149 (41%) | 2 278 228 (11%) |
| 40-50                | 1 289 314 (54%) | 1 080 854 (46%) | 2 370 168 (11%) |
| 50-60                | 1 523 969 (51%) | 1 478 293 (49%) | 3 002 262 (14%) |
| 60-70                | 1 541 862 (47%) | 1 768 249 (53%) | 3 310 111 (16%) |
| 70-80                | 1 314 876 (46%) | 1 560 407 (54%) | 2 875 283 (14%) |
| 80-90                | 906 750 (53%)  | 807 643 (47%) | 1 714 393 (8%) |
| 90+                  | 356 648 (67%)  | 177 126 (33%) | 533 774 (3%) |
| **Total**            | 10 571 097 (51%) | 10 173 498 (49%) | 20 744 595 |