Evaluating the impact of text duplications on a corpus of more than 600,000 clinical narratives in a French Hospital

William Digan, Maxime Wack, Vincent Looten, Antoine Neuraz, Anita Burgun, Bastien Rance

Evaluating the impact of text duplications on a corpus of more than 600,000 clinical narratives in a French Hospital

William Digan[ab], Maxime Wack[ab], Vincent Looten[ab], Antoine Neuraz[ac], Anita Burgun[abc], Bastien Rance[ab]

[a] INSERM, Centre de Recherche des Cordeliers, UMRS 1138, Université Paris-Descartes, Université Sorbonne Paris Cité, Paris, France,

[b] Department of Medical Informatics, Hôpital Européen Georges Pompidou, AP-HP, Paris, France[c] Department of Medical Informatics, Necker Children Hospital, AP-HP, Paris, France

Abstract

A significant part of medical knowledge is stored as unstructured free text. However, clinical narratives are known to contain duplicated sections due to clinicians' copy/paste parts of a former report into a new one. In this study, we aim at evaluating the duplications found within patient records in more than 650,000 French clinical narratives. We adapted a method to identify efficiently duplicated zones in a reasonable time. We evaluated the potential impact of duplications in two use cases: the presence of (i) treatments and/or (ii) relative dates. We identified an average rate of duplication of 33%. We found that 20% of the document contained drugs mentioned only in duplicated zones and that 1.45% of the document contained mentions of relative dates in duplicated zone, that could potentially lead to erroneous interpretation. We suggest the systematic identification and annotation of duplicated zones in clinical narratives for information extraction and temporal-oriented tasks.

Keywords: Electronic Health Records, Natural Language Processing, Algorithms

Introduction

The use of electronic health records (EHRs) and clinical data warehouses [1] (CDWs) lead to a better collection and preservation of patient information. CDWs store all kinds of data, including laboratory results, diagnostic codes, and clinical narratives (free text medical reports). In fact, CDWs are major tools for translational research. While a large portion of the information are stored in structured ways, and virtually directly reusable, a significant part of medical knowledge is stored as unstructured free text. Some studies have even shown that free text contains up to 80% [2] of overall information. With the availability of information, new ways of exploring data have emerged. For example, high-throughput phenotyping, machine learning or statistical models (including through the use of deep learning). However, free text can be subject to different types of issues (quality, typos…), that could profoundly bias results of analysis and models. One potential problem could come from duplicated sections in clinical reports [3] (created when clinicians copy/paste parts of a former report into a new one).

While duplications common during the care (information can be replicated from one document to another because of the static nature of the family history, previous treatments, and so on…). However, in the case of secondary use of clinical data and more specifically, in the context of data extractions, duplications can have a strong impact on the chronology of the information (an old information can be present in a recent document), and on the sheer presence of the information. In this study, we aim at assessing if duplications have an impact on different types of models.

This study takes place in the context of big data, and simple naïve approaches are not compatible with the volume of data considered (several months of calculation would be needed for simple tasks). A large body of work has been developed around the detection of plagiarism and duplication in clinical narratives. The volume of duplications has been evaluated as high as 80% in Northern American Hospital [4–6]. However, the exploration of duplications in French narratives remains limited, and the potential impact of such duplications is not easy to evaluate.

State of the Art

The identification of duplicated zones or plagiarism has generated a large body of work over the years. However, no open source solutions are available and able to handle the volume of text compatible with our purpose. In medicine, several studies focus on the characterization of copy and paste redundancy.

In their publication of 2013 [6], Cohen *et al.* studied the impact of 'copy and paste' redundancy in a large corpus of text. For that purpose, they developed a character based fingerprint method. This technic is inspired by blast [7] a bioinformatics algorithm which aims to find similar sequence. The authors considered 22,654 notes for 1604 patients. They found that clinical text had a redundancy level of 29%.

In a French preliminary study [3], D'Hondt *et al.* extend the Cohen methods and studied duplications in French clinical notes. The algorithm allowed the use of overlapping fingerprints. They also oriented documents in time. They choose as parameters a fingerprint length of 30 and an overlap of 10 char. Furthermore they identified that in clinical notes, most of the redundancy located on the footer and header section were administrative sections. They worked on the documents from three records and 361 documents. They showed a redundancy level of 33% in clinical notes. Their algorithm allows finding near-duplicated and exact redundancy (at a price of a higher complexity of the algorithm).

A recent study [8] by Gabriel *et al.* was able to scale up to 1.5 million notes in 36.3 hours, regardless of the patient vector. They developed a new method base on windows of three phases. The first step is mini-hashing generation from files. Instead of looking at the character levels they look at the word level and defined a signature. This approach will not be followed in our study because we want to find exact duplication.

Goals

In this study, we aim at evaluating the duplications found within patient records at the European Hospital Georges Pompidou, a French hospital located in Paris. We adopted strategy to enable the treatment of large quantities of text in a reasonable time. Finally, we evaluate the potential impact of

duplications in two use cases: (i) the identification of treatments present in clinical narratives and (ii) the presence of relative dates.

Materials

In this section, we introduce the European Hospital Georges Pompidou and the corpus of text used for the study.

European Hospital Georges Pompidou

The European Hospital Georges Pompidou (HEGP in French) is a 700 beds hospital located in Paris. The HEGP is specialized in oncology, cardiovascular diseases and emergency medicine. The hospital has deployed in 2008 a clinical data warehouse (HEGP CDW) based on i2b2 [9] integrating virtually all the data generated by the hospital information system. Among the data collected, clinical narratives (comprised of clinical reports, letters, imaging reports, and so forth) represent more than 10 million items.

Corpus of Clinical Narratives

Our dataset is a subset of the corpus of the text of the HEGP CDW. We identified all the patients who received chemotherapy since the opening of the hospital 2000 (10,393 patients). We limited the selection of patients to those who had a follow-up of at least a year (i.e. patients with at least two visits distant by 365 days). Because we are interested in duplications within the record of a patient, we selected only patients with at least two distinct documents. Starting from 666,956 documents, we conserved a total of 649,651 documents after a preprocessing step (detail in the method section).

Methods

Definition of Duplicated Zones

In this manuscript, we define a duplication as an identical zone of text found conserved in at least two different documents. We focus on intra-record duplication (i.e. we search for duplication with the record of a patient, and not between patients). The document pairs are oriented in time.

Preprocessing

All documents generated in the hospital comprised administrative information (with the phone number of the service, the names of the staff, and so forth), the clinical notes themselves and footer information regarding the possible secondary use of data. We preprocessed the documents to remove the administrative zones and the footer information section. We also normalized the documents by converting the entire text to lower cases, and transforming multiple spaces into single ones.

Efficient Detection of Duplications

We aim at developing a method able to manage a substantial number of documents. We leverage the approaches developed by Cohen et al. [6] and D' Hondt *et al.* [3] to develop a mix approach. In a nutshell, we rely on fingerprints build from the text to identify identical zone. A fingerprint is a segment of *N* consecutive letters. Fingerprints are not overlapping, if the first fingerprint is constructed from character 1 to N, the second fingerprint starts at position N+1. Similarly to D'Hondt et al., we also leveraged the notion of overlap: we add series of fingerprints with an offset of value OFFSET (i.e. starting at the OFFSETth character). OFFSETs are very similar to Open Reading Frames in DNA. Figure 1 shows a graphical summary of the approach. We detect duplicated ones by comparing fingerprints between pairs of documents. Contiguous or overlapping pairs of fingerprints (in the source and target documents) are merged together. We evaluated different sizes of fingerprints, and values of offsets to find a good compromise between the number of duplicated zones detected and the computed time needed to perform the calculation.

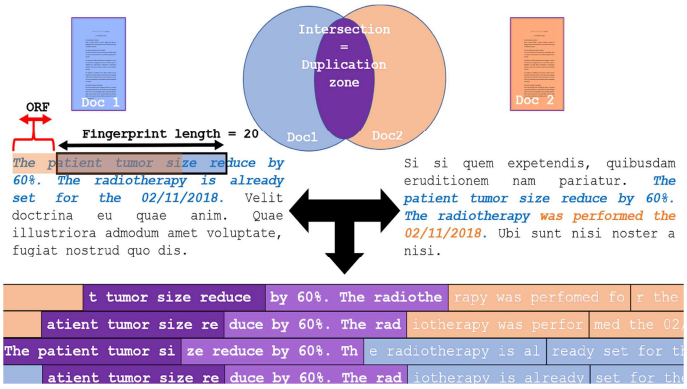

Figure 1: detection of duplicated region between two texts. The method relies on fingerprints.

Evaluating the optimal parameters: We tested combination of parameters for the values of *N* (size of the fingerprints), and OFFSET (value of the offset). We respectively tested the values of 30, 40 and 50 characters from the fingerprints, and 1, 5, 7, 10, 15 and 20 for offsets. In each case, we considered the offset of 1 as our baseline (fingerprints are calculated at each character position). We computed the number of common fingerprints detected, the time needed for the computation.

Computing Duplicated Zones on the Corpus of 650,000+ Documents

Using the optimal values obtained from the previous section (fingerprint length of 30, and offset of 15 characters), we computed duplicated zones on the entire corpus. Duplicated zones are detected among the documents of a single patient. We search for duplicated zones between documents oriented in time: the source document was always older than the target document. Once the pairwise duplication step has been performed, we focus on document levels, and merge all the duplicated zone detected. Figure 1 illustrates the approach.

***Filtering duplicated zones*:** After the merging steps, we identified duplicated zones with a wide variety of length, starting from 30 (the length of a fingerprint). We chose to filter out zones too small, because they likely did not correspond to copy/paste. To select a relevant threshold, we considered the number of duplications found for a given duplicated zone length.

Evaluation the volume of duplicated zones

Finally, we designed three scores to evaluate the volume of duplicated zones in the text;

Global volume of duplication in the corpus: The global duplication score

$$DupScoreGlobal = \frac{\sum_{j=0}^{Totaldocs} DuplicationLength_j}{\sum_{j=0}^{totaldocs} LengthDoc_j}$$

Average duplication score by document: The duplication score per document defined as

$$DupScoreAverageDoc = \frac{\sum_{i=0}^{Totaldocs} \frac{DuplicationLength_i}{LengthDoc_i}}{Totaldocs}$$

Average duplication score per patient: The duplication score per patient is defined as

$$DupScoreAveragePat = \frac{\sum_{k=0}^{TotalPatients} \frac{\sum_{ki=0}^{Totaldocs} DuplicationLength_{ki}}{\sum_{Ki=0}^{totaldocs} LengthDoc\,k_i}}{TotalPatients}$$

In a nutshell, the three score describe different means to measure the amount of duplications. The global duplication socre (*DupScoreGlobal*) evaluate the overall amount of duplication in the corpus (in term of number of characters in duplicated zones). The average duplication score per document evaluates the impact of duplication normalized by document. Finally, the average duplication per patient measures the overall impact of duplication per patients.

Potential Impact of Duplicated Zones on Two Use-Cases

We identified two use cases that could potentially be impacted by the presence of duplicated zones:

***Detecting drugs in duplicated zones*:** We searched for occurrences of medical drugs in our corpus. We used an exact match strategy, based on a list of ingredients and brand names from the Romedi[10] resource. Romedi is a semantic web version of a French public resource of drugs made available by the French National Health Insurance. Molecules such as simple sugars (e.g., glucose), water, inorganic elements (e.g., calcium), and so forth are listed as ingredients in Romedi. However, when mentioned in the clinical narratives, these molecules rarely refer to clinical drugs. Therefore, we eliminate them from the list of drugs identified by Romedi (more precisely, we eliminated the French terms *para*, *olivier*, *alcool*, *sodium*, *potassium*, *calcium*, *glucose*, *magnesium*, *eau*). All drugs were normalized to their corresponding CUI.

We identified the number of drugs present *only* in duplicated zones, and not in the rest of the document. While the presence can be useful for the medical history and for the care of the patient, the presence in portions of text duplicated from former documents could impact machine learning models, or information retrieval processes.

***Relative dates in duplicated zones*:** Our second use case focused on temporality. One major issue when working with text is the identification of the temporality associated with the concepts identified in the text. It is always important to distinguish between events or phenotypes that occurred during or prior the encounter. We searched the duplicated zones for temporality markers using relative dates (i.e. using expressions such as yesterday, two months ago, tomorrow, today, etc.). In such cases, the reference date is assumed to be the date of the creation of the document, but because the expression is located in a duplicated zone, its actual reference date should be identified in the past. We searched the corpus for a series of 8 terms corresponding to relative dates and determined if the terms were located within a duplicated zone.

Implementation of the Pipeline of Detection

We leveraged NextFlow[11] and Docker [12]. Each portion of our pipeline uses a Docker container and Nextflow ensure the parallelization of our processes. The pipeline ran on an Ubuntu 14.04 server, with 15 cores, 64 GB of RAM, and was developed in Python 3.10. Code is accessible on our github repository:
https://github.com/equipe22/duplicatedZoneInClinicalText [13].

Results

Preprocessing

A mean values of 1670 characters were eliminated in general during the preprocessing. Overall, the number of character decreased by 36%. The average length of a text before preprocessing was 4145 and 2474 characters after.

Efficient Detection of Duplications

We compared the execution time and performance with respect to the overlap for different sets of parameters of the detection duplication algorithm (see Table 1).

Table 1 – result of parameters evaluation for 50 patients which have 30 documents in average

fingerprint length	orf size	execution time (second)	% median overlap with the baseline
20	3	653	83
	5	196	77
	7	85	69
	10	34	72
	15	**18**	**67**
	20	10	66
30	3	665	84
	5	274	82
	7	125	80
	10	46	79
	15	**22**	**78**
	20	14	75
40	3	1043	83
	5	395	81
	7	166	80
	10	63	81
	15	**28**	**78**
	20	16	72

Computing Duplicated Zones on the Corpus of 650,000+ documents.

Table 2 –Duplication detection and annotations execution time

	fingerprint generation	merge	drug annotation	time annotation
execution time	3h19	20h19	80 s	23 s

Figure 2– distribution of the patient duplication score

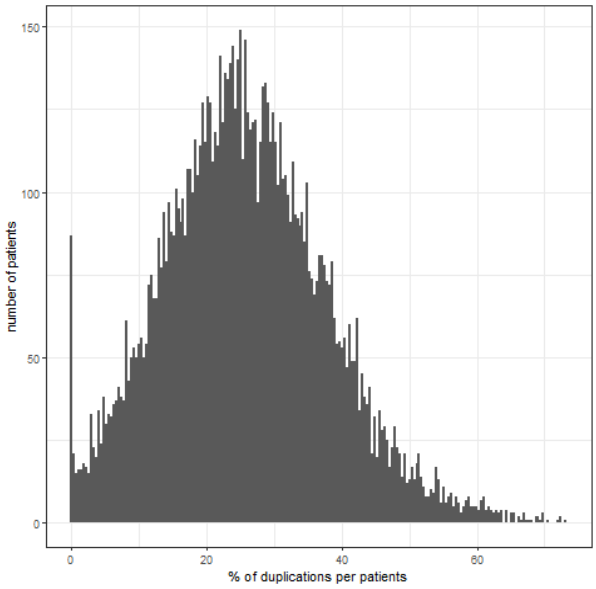

Table 3– summary of duplication score

score	mean	Standard deviation
Global	0.33	0.33
Avg per document	0.25	0.12
Avg per patient	0.28	0.14

Potential Impact of Duplicated Zones on Two Use-Cases

***Detecting drugs in duplicated zones*:** We extracted 2,689,998 brand name and 761,611 ingredients from the corpus. 330,272 documents contain at least one drug mention. Overall, 161,067

documents had a drug detected within a duplicated region. 130,233 documents had at least one drug detected only within the duplicated zone (19.64% of our corpus).

Relative dates in duplicated zones: 45,557 documents contained at least one mention of a relative date. 9,632 documents contained a mention of a relative date within a duplicated region (21% of relative dates, 1.45% of the corpus).

Discussion

Detection of Duplication

We found that fingerprints length did not have an impact on the algorithm speed neither for the generations of fingerprints, nor the identification of duplications. The offset size did have a strong effect on both the execution time and the quality of detection. Compared to the baseline (offset of 1), the lower the offset size is, the better is the quality. However, in the spirit of a scalable approach, the processing time is incompatible with high volume of documents. We selected an offset of 15 for a fingerprint size of 30 for the reminder of our process to preserve a good quality while benefiting from a 200-fold speed improvement of the algorithm.

Filtering: We observed a large number of small-sized duplicated zones of 30 characters (more than 200 million detected duplications). 30 characters are highly unlikely to correspond to a full sentence in French. We decided to use a threshold of 1.5 fingerprints (i.e. 45 characters) to reduce the impact of artefacts that are unlikely to have been generated by a copy/paste process. Using this threshold, we found 29 million detected duplication.

Duplicated zones: Overall, the ratio of duplicated rate of duplications is 33%, in par with findings from the literature [3].20% of document had drugs mentioned only in duplicated zones. 1.45% of the document contained a relative data present in a duplicated zone. While the number is relatively low, the global number of documents is high: several thousands of documents for CDW with 10 million documents. The risk of misinterpretation of relative dates is high; tools such as HeidelTime [14] often used to identify mentions of temporality could provide erroneous normalization of the date since the tool would use the date of the document as a reference (instead of the data of the document source of the duplication).

Technical Significance

The performance of our heuristic allows treating a large amount of text. In this study, we managed a corpus of more than 650,000 documents within less than a day. Our CDW hosts a total of 10 million clinical narratives, some of which are the seldom report in the patient record.

The heuristic approach probably underestimates the volume of duplications. Additional fined grained approaches [3] could be applied to refine our results. We applied our approach to French, but the algorithm could be used for other languages as well.

Significance for secondary use of clinical data

The overall rate of duplicated zone (33%) is reasonable. However, we identified both drugs and relative dates were present in duplicated zones and could have a strong impact on information extractions from the text.

Duplications can have various meanings. The physician can use copy/paste to summarize the past, or to carry medical history from one document to another. Our method does not allow to identify the meaning associated with the copy/paste. However, for any application in which temporality is of importance, relative dates in duplicated zone might present an issue.

Limitation

The HEGP is specialized in oncology and cardiovascular diseases. Our selection of patients did not reflect the variety of the case present in the hospital. However, we did not filter the documents to specific sets of providers. In chronic diseases, with longer follow-ups, it would be possible for the ratio of duplication to be higher.

We used a rule based approach to clean-out the administrative sections of the document. This approach is not transposable, but proved efficient. The structure of the document is highly linked to the Electronic Health Record used, the adoption of standard, etc.

We did not consider inter-patient duplications.Whereas our method could be used similarly to detect duplications among documents from different patients, it was not the purpose of our study. The detection of such zones could be interesting for quality control, or to reduce the work when annotating large corpora of texts for example.

Perspectives

Evolution of the volume of duplication over time: Because of the large variety of profiles, it is complex to provide a good indicator of the evolution of the duplication rate over time. In our corpus, the documents were generated by many providers (medical services). We explored visually this question by representing the duplication rate over time. For comparison purpose, we normalized the time. Figure 3 provides a visualization of the duplication rate over the documents (the 0 in abscissa corresponding to the first document, and 100 to the last). A single point represents the rate of duplication of a single document for a single patient. We can see that there is visually a small trend toward an increase in the rate of duplications in the early part of the distribution, followed up by a plateau.

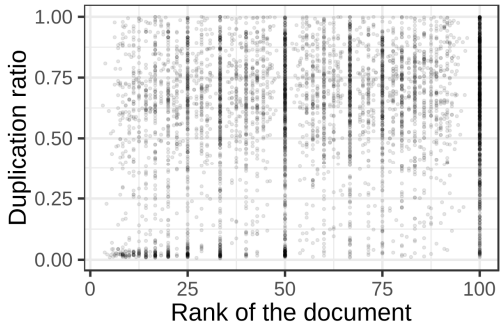

Figure 3 – Duplication representation over time per provider

We explored a visualization of the duplications within a patient records, and their organization over time in Figure 4. We leveraged the circlize [15] visualization to build a graphical summary of the duplicated zone, and their origin for a given patient. The outside circle represents the document of patients. The inner circle represents the provider of the document. Each edge represents a duplicated zones between documents. The date difference between two documents is rendered by the color; darker arcs correspond to the larger number of days than lighter arcs. In our example, Patient 1 and 2 both have 30 documents. The two patients have two distinct pathologies, and therefore two sets of distinct providers. The arcs reflect different hospitalization trajectories.

For Patient 1, the providers reflects an oncology trajectory: digestive surgery (502, 532), imaging (312) and chemotherapy (574). For Patient 2, providers are coherent with urgent care: Internal medicine (812) and emergency medical (108) and reanimation. The systematic identification and annotation of duplicated zones are important for many aspects of data reuse.

While we limited our exploration to drugs and relative dates, other semantic areas would be relevant to explore. For example, procedures and phenotypes. The annotation of duplicated zones could help identify procedures that are not relevant to the current visit.

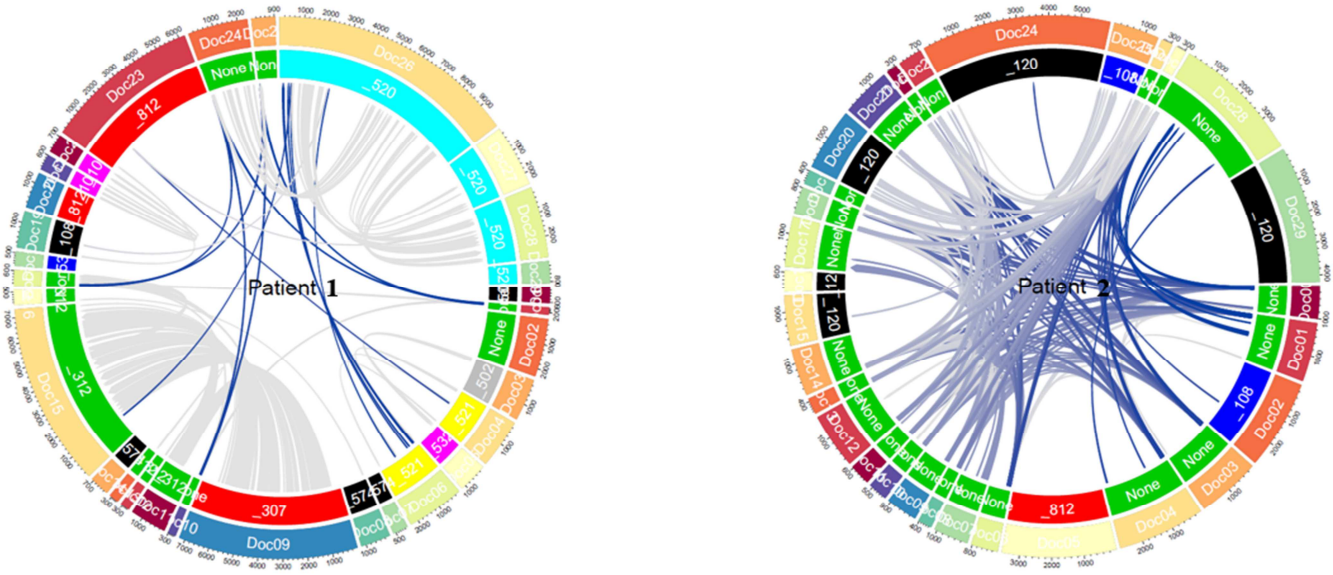

Figure 4– duplication representation over time for two distinct patients. The length of a document reflects its number of characters, an arc between two regions translate the duplication of a portion of text.

Conclusions

We developed a method to identify efficiently duplicated zones in clinical narratives. We explored a corpus of more than 650,000 documents belonging to 10376 patients. We identified an average rate of duplication of 33%, in par with value found in other studies. We evaluated the potential impact of duplications in two use-cases, the identification of drugs and the identification of relative dates. We found that 20% of the document contained drugs mentioned only in duplicated zones and that 1.45% of the document contained mentions of relative dates in duplicated zone, that could potentially lead to erroneous interpretation. We suggest the systematic identification and annotation of duplicated zones in clinical narratives for information extraction and temporal-oriented tasks.

Acknowledgements

WD is funded by the ANR PractikPharma. BR is funded in part by the SIRIC CARPEM research program.

Reference

[1] S.N. Murphy, G. Weber, M. Mendis, V. Gainer, H.C. Chueh, S. Churchill, and I. Kohane, Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2), *Journal of the American Medical Informatics Association*. **17** (2010) 124–130. doi:10.1136/jamia.2009.000893.

[2] J.-B. Escudié, B. Rance, G. Malamut, S. Khater, A. Burgun, C. Cellier, and A.-S. Jannot, A novel data-driven workflow combining literature and electronic health records to estimate comorbidities burden for a specific disease: a case study on autoimmune comorbidities in patients with celiac disease, *BMC Med Inform Decis Mak*. **17** (2017). doi:10.1186/s12911-017-0537-y.

[3] E. D'hondt, X. Tannier, and A. Névéol, Redundancy in French Electronic Health Records: A preliminary study, in: Association for Computational Linguistics, 2015: pp. 21–30. doi:10.18653/v1/W15-2603.

[4] J.O. Wrenn, D.M. Stein, S. Bakken, and P.D. Stetson, Quantifying clinical narrative redundancy in an electronic health record, *Journal of the American Medical Informatics Association*. **17** (2010) 49–53. doi:10.1197/jamia.M3390.

[5] R. Zhang, S. Pakhomov, B.T. McInnes, and G.B. Melton, Evaluating measures of redundancy in clinical texts, *AMIA Annu Symp Proc*. **2011** (2011) 1612–1620.

[6] R. Cohen, M. Elhadad, and N. Elhadad, Redundancy in electronic health record corpora: analysis, impact on text mining performance and mitigation strategies, *BMC Bioinformatics*. **14** (2013) 10. doi:10.1186/1471-2105-14-10.

[7] S.F. Altschul, T.L. Madden, A.A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D.J. Lipman, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res*. **25** (1997) 3389–3402. doi:10.1093/nar/25.17.3389.

[8] R.A. Gabriel, T.-T. Kuo, J. McAuley, and C.-N. Hsu, Identifying and characterizing highly similar notes in big clinical note datasets, *Journal of Biomedical Informatics*. **82** (2018) 63–69. doi:10.1016/j.jbi.2018.04.009.

[9] E. Zapletal, N. Rodon, N. Grabar, and P. Degoulet, Methodology of integration of a clinical data warehouse with a clinical information system: the HEGP case, *Stud Health Technol Inform*. **160** (2010) 193–197.

[10] ROMEDI - Référentiel Ouvert du Médicament, (n.d.). http://csfbtp.fr/telechargement.html (accessed September 10, 2018).

[11] Paolo Di Tommaso, Maria Chatzou, Pablo Prieto, Emilio Palumbo, and Cedric Notredame, Nextflow: A tool for deploying reproducible computational pipelines , (2015). doi:10.7490/f1000research.1110183.1.

[12] Docker, Inc, Docker - Build, Ship, and Run Any App, Anywhere, (n.d.). https://www.docker.com/.

[13] equipe22/duplicatedZoneInClinicalText, *GitHub*. (n.d.). https://github.com/equipe22/duplicatedZoneInClinicalText (accessed November 25, 2018).

[14] J. Strötgen, and M. Gertz, Multilingual and cross-domain temporal tagging, *Language Resources and Evaluation*. **47** (2013) 269–298. doi:10.1007/s10579-012-9179-y.

[15] Z. Gu, circlize: Circular Visualization, 2018. https://CRAN.R-project.org/package=circlize (accessed November 4, 2018).

Address for correspondence
Bastien Rance, Email: bastien.rance@aphp.fr
European Hospital Georges Pompidou, 20 rue Leblanc, 75015 Paris France. Tel: +331.56.09.59.85