Natural language processing for rapid response to emergent diseases: calcium channel blockers and hypertension in the COVID-19 pandemic

Antoine Neuraz, Ivan Lerner, William Digan, Nicolas Paris, Rosy Tsopra, Alice Rogier, David Baudoin, Kevin Bretonnel Cohen, Anita Burgun, Nicolas Garcelon, Bastien Rance, AP-HP/Universities/INSERM COVID-19 Research Collaboration; AP-HP COVID CDR Initiative

Submitted to: Journal of Medical Internet Research
on: June 02, 2020

Table of Contents

Natural language processing for rapid response to emergent diseases: calcium channel blockers and hypertension in the COVID-19 pandemic

Antoine Neuraz[1, 2, 3] MD; Ivan Lerner[1, 2] MD; William Digan[2, 4] MSc; Nicolas Paris[5] MSc; Rosy Tsopra[2, 4] MD, PhD; Alice Rogier[2, 4] MSc; David Baudoin[2, 4] MSc; Kevin Bretonnel Cohen[6] PhD; Anita Burgun[1, 2, 4] MD, PhD; Nicolas Garcelon[2, 7] PhD; Bastien Rance[2, 4] PhD; AP-HP/Universities/INSERM COVID-19 Research Collaboration; AP-HP COVID CDR Initiative[8]

[1]Department of Biomedical Informatics Necker-Enfant Malades Hospital Assistance Publique – Hôpitaux de Paris (AP-HP) Paris FR
[2]Centre de Recherche des Cordeliers INSERM UMRS 1138 Team 22 Universite? de Paris Paris FR
[3]LIMSI CNRS Université Paris Saclay Orsay FR
[4]Department of Biomedical Informatics Georges Pompidou European Hospital Assistance Publique – Ho?pitaux de Paris (AP-HP) Paris FR
[5]DSI WIND Assistance Publique – Ho?pitaux de Paris (AP-HP) Paris FR
[6]School of Medicine University of Colorado Denver US
[7]Institut Imagine, INSERM U1163 Universite? Paris Descartes Universite? de Paris Paris FR
[8]AP-HP/Universities/INSERM COVID-19 Research Collaboration: AP-HP COVID CDR Initiative Paris FR

Corresponding Author:
Antoine Neuraz MD
Department of Biomedical Informatics
Necker-Enfant Malades Hospital
Assistance Publique – Hôpitaux de Paris (AP-HP)
Bat Imagine, Bureau 145
149 rue de Sèvres
Paris
FR

Abstract

Background: A novel disease poses special challenges for informatics solutions: biomedical informatics relies for the most part on structured data; structured data require a pre-existing data/knowledge model; but novel diseases do not have pre-existing knowledge models. In an emergent epidemic, language processing could allow rapid conversion of unstructured text to a novel knowledge model. But, although this has often been suggested, there has never before been an opportunity to actually test that claim in real time. The current pandemic presents such an opportunity.

Objective: To evaluate the added value of information from clinical text in the response to emergent diseases.

Methods: We explored the effect of long-term treatment by calcium channel blockers on the outcome of COVID infection in patients with high blood pressure during in-patient hospital stay, using two sources of information: data available strictly from structured Electronic Health Records and data available through structured Electronic Health Records and text mining.

Results: In this multicenter study involving 39 hospitals, text mining increased statistical power sufficiently to change a negative result for an adjusted hazard ratio to a positive one. Compared to the baseline structured data, it increased the number of patients available for inclusion in the study by 2.95 times, the amount of available information on medications by 7.2 times, and additional phenotypic information by 11.9 times.

Conclusions: Calcium channel blockers are associated with decreased in-hospital mortality in patients with COVID-19 infections. This was found because a natural language processing pipeline could be adapted quickly to the domain of the novel disease and still perform well enough to extract useful information. When that information is used to supplement existing structured data, a sample size can be increased enough to see treatment effects that were not previously statistically detectable.

(JMIR Preprints 02/06/2020:20773)
DOI: https://doi.org/10.2196/preprints.20773

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Short Paper

Natural language processing for rapid response to emergent diseases: calcium channel blockers and hypertension in the COVID-19 pandemic

Abstract

Background: A novel disease poses special challenges for informatics solutions: biomedical informatics relies for the most part on structured data; structured data require a pre-existing data/knowledge model; but novel diseases do not have pre-existing knowledge models. In an emergent epidemic, language processing could allow rapid conversion of unstructured text to a novel knowledge model. But, although this has often been suggested, there has never before been an opportunity to actually test that claim in real time. The current pandemic presents such an opportunity.

Objective: To evaluate the added value of information from clinical text in the response to emergent diseases.

Methods: We explored the effect of long-term treatment by calcium channel blockers on the outcome of COVID infection in patients with high blood pressure during in-patient hospital stay, using two sources of information: data available strictly from structured Electronic Health Records and data available through structured Electronic Health Records and text mining.

Results: In this multicenter study involving 39 hospitals, text mining increased statistical power sufficiently to change a negative result for an adjusted hazard ratio to a positive one. Compared to the baseline structured data, it increased the number of patients available for inclusion in the study by 2.95 times, the amount of available information on medications by 7.2 times, and additional phenotypic information by 11.9 times.

Conclusions: In our study, calcium channel blockers are associated with decreased in-hospital mortality in patients with COVID-19 infections. This was found because a natural language processing pipeline could be adapted quickly to the domain of the novel disease and still perform well enough to extract useful information. When that information is used to supplement existing structured data, a sample size can be increased enough to see treatment effects that were not previously statistically detectable.

Keywords: calcium channel blockers; hypertension; natural language processing; electronic health records; COVID-19; public health

Introduction

Outbreaks of novel diseases can pose enormous strains on public health systems. Since the time of Snow's pioneering work on the epidemiology of the London cholera outbreak of 1854[1], it has been clear that information is key to the successful abatement of these enormous public health challenges. Today the health care system has access to quantities of data that would have been unimaginable in Snow's time. The fact that they are in electronic format makes them manipulable and exploitable in rapid fashion. But a novel disease poses special challenges for informatics solutions: biomedical informatics relies for the most part on structured data; structured data require a pre-existing data/knowledge model; and part of being a novel disease is not having a pre-existing knowledge model. This poses a formidable obstacle to leveraging informatics solutions to the kind of public health crisis that the world faces as we write this. One solution to the lack of structured information is natural language processing.

Biomedical text mining, or the use of textual data, in electronic health records, has often been proposed as a method for converting unstructured data to the structured data that is needed in public health informatics. One of its advantages is that text mining can be developed rapidly [2] which could permit the leveraging of electronic health records of patients with a novel disease as quickly as it is entered into health records. But, although this has often been suggested [3] , there has never before been an opportunity to actually test that claim in real time. Thus, the current novel coronavirus, with all of its tragedies, also presents an opportunity to advance the state of public health informatics. This paper tests those possibilities with a case study on the use of calcium channel blockers (CCB) in patients with high blood pressure (HBP) on the risk of death from COVID infection. The association between CCB and the outcome of a COVID infection has already been suggested[4], but has not previously been explored in a large multicenter clinical study.

Materials and Methods

Materials

The data that is used in this study comes from 39 different hospitals in the Paris metropolitan area, named AP-HP. Focusing on this region of the country and on a large number of hospitals allowed a diversity of patient demographics than would be unavailable in most other parts of the country. As of May 4th, the EDS-COVID data set contained 84,966 electronic records of suspected or confirmed COVID patients [see Table 1 for further details on the data set]. The records are comprised of structured fields and free text documents, including clinical notes and narratives. Most textual documents do not follow a specific structure and contain different types of patient information, e.g., patient history, family history, laboratory results, drug history and prescriptions. This makes them an excellent test case for the real abilities of text mining. We used the following pipeline:

- Typical pre-processing steps (i.e., text cleaning, sentence detection) were applied to the full data set (See Supplementary Appendix 1 for a detailed description);
- Extraction of drug names and details of administration (dose, route of administration, frequency, and duration) was done via a deep learning approach based on BERT contextual embeddings [5] (NLP Medication) ;
- Extraction of specific phenotypes associated with COVID *(e.g.*, obesity, smoking status), scores (*e.g.*, sequential organ failure assessment score) and physiological measures (*e.g.*, body mass index), was accomplished via a list of 60 regular expressions (NLP RegExp);
- Extraction of all signs, symptoms and comorbidities included in the Unified Medical Language System [6] (UMLS) was done with the quickUMLS algorithm [7] (NLP UMLS).

The NLP medication extraction model was a BiLSTM-CRF [8] on top of a vector representation of tokens using BERT[5]. We fine-tuned BERT multilingual on a set 10 millions of clinical texts from EHRs. The model was trained on the APMed corpus, a manually annotated corpus of French clinical texts described in [9]. We used the FLAIR[10] implementation with 2 layers of 1,024 units for the LSTMs with a ASGD optimizer, and a reduction of the learning rate on plateau.

The NLP RegExp for the extraction of specific phenotypes were a set of 60 regular expressions developed manually and iteratively by medical informatics experts and medical doctors. We evaluated their precision at the sentence level using a random sample of 100 positive sentences of each regular expression.

All the terms extracted by the NLP pipeline, regardless of the method, were automatically annotated according to their modality (negated, hypothetical) and experiencer in the text as described in a previous work [11]. The outputs of the NLP pipeline were normalized to the Observational Medical Outcomes Partnership common data model [12] (OMOP CDM) and were fed back to the database system on a daily basis.

Clinical application: Long-term calcium channel blockers and the outcome of COVID in patients with HBP.

The clinical goal of this case study was to evaluate potential effects of calcium channel blockers (CCB) on in-hospital mortality related to COVID-19 (see [4]). To do this, we used two different sources of data.

1. The first source was two elements of structured data: ICD10 codes, and medication prescriptions from an electronic prescription system).
2. The second source was information on medications and comorbidities extracted by the NLP pipeline from non-structured fields in the electronic health record. The inclusion criterion for patients was COVID-19 disease confirmed by RT-PCR.

We considered a patient as having a long-term treatment by CCB (Supplementary Appendix 2) if there were at least two mentions (in structured data or extracted with NLP respectively) in the last 6 months. We qualified cases as having comorbidities through one occurrence of an ICD10 code (Supplementary Appendix 3) or two NLP mentions, in the last 6 months.
The measured outcome was in-hospital mortality. We used a multivariate Cox proportional hazard model [13], adjusted according to age, gender, and the presence of obesity, diabetes, and cancer. The level of significance was set to 0.05 and all statistical tests were two-sided. We used R statistical software v.3.6.2 with the survival package.

Results

Natural language processing pipeline

As Table 1 shows, natural language processing markedly expanded the quantity of medication and phenotype information available for the analysis. The number of data points for medication increased by 7.2 times ($NLP\, medication / Structured\, medication$), and the number of phenotypes by 15.2 times ¿¿). Among the 84,966 records present in the EDS-COVID cohort (Table 1), 53% of the patients had drug information in their narrative EHR documents whereas only 23% of the patients had medication information available in structured fields in the EHR.

Table 1: Description of the information extracted using the NLP pipelines in the EDS-COVID cohort

Source	Patients, N=84,966	Documents, N=1,524,057	Data
NLP Medication	45,593 (53%)	696,125 (46%)	5,995,945

NLP RegExp	44,498 (52%)	711,900 (46%)	5,449,932
NLP UMLS	44,035 (52%)	833,610 (55%)	19,626,172
Structured medication	19,791 (23%)	-	826,554
ICD10 codes	38,993 (46%)	-	1,643,819

NLP = Natural language processing, UMLS=Unified medical language system, ICD=International classification of diseases version 10, RegExp=Regular Expressions

For specific phenotypes with existing ICD10 codes (Figure 1), information was only available in clinical free–text fields for the majority of patients: 5,133/8,526 (83%) for diabetes, and 2,138/2,871 (74%) for obesity. Some items were absent from the structured data, however, could be recovered using the NLP extraction pipeline, such as COVID specific symptoms like ageusia or anosmia, 2,449 and 2,732 patients respectively.

In terms of quality, the extraction of medication names showed a F1-measure of 93.8% (91.6% after normalization) in all sections. When focusing on the admission and discharge treatments sections, the F1-measure was 96.7% (96% after normalization). The detailed results are shown in Supplementary Appendix 8. Regarding of the phenotypes extracted by regular expressions, we show here the results of those used in the case study: hypertension showed a precision of 99%, and obesity, diabetes and cancer, of 94%, 80% and 91%, respectively.

Case Study

A total of 3,965 patients were included using the NLP pipeline whereas 1,343 only could be included if the study were limited to the use of structured data, increasing the number of patients added for the case study increased by 2.95 times. (Supplementary Appendix 5) A detailed description of the population of COVID positive patients with history of HBP can be found in (Supplementary Appendix 4). In terms of temporal depth of CCB treatment information, Figure 2 shows a higher volume of information coming from text compared to structured data.

When using only structured data, we observed an adjusted hazard ratio (aHR) of 0.83 (95%CI 0.67-1.05) for a treatment by CCB, not statistically significant. When including NLP data, the aHR becomes 0.82 (95%IC 0.71-0.94) and represents a statistically significant reduction of the risk of death. Similar results can be observed that support an increased risk of mortality with the presence of diabetes and cancer as comorbidities. (Table 2)

Table 2: Multivariate Cox survival model (CI: Confidence Interval, HR : Hazard Ratio)

	Structured data			**NLP**		
	HR[1]	95% CI[1]	p-value	HR[1]	95% CI[1]	p-value
Calcium channel blockers	0.83	0.67, 1.05	0.12	0.82	0.71, 0.94	0.005
Age						

45-64	---	---		---	---	
18-44	0.20	0.03, 1.46	0.11	0.35	0.15, 0.80	0.013
65-74	1.50	0.99, 2.27	0.053	1.95	1.54, 2.47	<0.001
75-84	1.68	1.14, 2.48	0.009	2.94	2.35, 3.69	<0.001
85+	2.45	1.66, 3.61	<0.001	3.99	3.16, 5.03	<0.001
Gender						
Female	---	---		---	---	
Male	1.59	1.27, 2.00	<0.001	1.53	1.32, 1.77	<0.001
Obesity	1.07	0.81, 1.42	0.6	1.13	0.90, 1.41	0.3
Diabetes	1.22	0.98, 1.52	0.080	1.25	1.09, 1.45	0.002
Cancer	1.20	0.96, 1.49	0.11	1.34	1.15, 1.56	<0.001

Discussion

The work reported here set out to investigate the potential for biomedical natural language processing to be of use in the context of a rapidly emerging novel disease. To do this, it asks a specific question: does the leveraging of unstructured textual information via natural language processing yield clinically actionable information? To answer that question, we used natural language processing to extract information about hypertension and a medication for treating it from the electronic health records of patients with COVID-19. The results showed that (1) a natural language processing pipeline can be adapted quickly to the domain of a novel disease, (2) it can perform well enough to extract useful information, and (3) when that information is used to supplement the structured data that is already available, the sample size can be increased enough to see treatment effects that were not previously statistically detectable.

Several agencies, notably the European Medicines Agency, have highlighted the benefits of using real-world data for research, and in particular for the generation of complementary evidence and new hypotheses [14]. During the peak of COVID pandemic time available for clinicians to enter EHR data was greatly reduced. Medical informatics became vital to manage the crisis in hospitals and acquire better knowledge of the disease. The NLP pipeline was implemented within two weeks at the beginning of the COVID epidemic in France, building on previous developments in Artificial Intelligence and text mining at AP-HP. More specifically, combining non-specific pre-existing developments (e.g. negation, family history and hypothesis detection) to tailored extractions (i.e. regular expressions) allowed us to obtain fast results of sufficient quality.

About sixty internal research projects exploring EDS COVID data have been submitted for IRB approval within the first eight weeks of COVID epidemic. More than half of them have studied variables such as symptoms (e.g., ageusia), radiological signs (e.g., crazy paving), comorbidities (e.g., obesity), and drug history (e.g., hydroxychloroquine) requiring extraction of information from narrative reports in EHRs.

The case study described in this article shows the possible impact of using information extracted from text in EHR for COVID research. More precisely, the above study would have had different conclusions if information from unstructured fields had been excluded. In our case study, the addition of information from NLP did not change dramatically the HR from the analyses but allowed us to include more patients and therefore narrowed down the confidence intervals and increased the statistical power. Note that the increased statistical power is mainly due to the increase in the number of patients included and the quantity of data available. Further analyses are required to assess the validity of the associations detected here given that some confounding biases may remain and provoke false positive results. Reproducing the analysis on an external population or perform some falsification testing [15] could help improving the validity of these findings.

Acknowledgements

The authors thank the EDS APHP Covid consortium integrating the APHP Health Data Warehouse team as well as all the APHP staff and volunteers who contributed to the implementation of the EDS-COVID database and operating solutions for this database.
The authors would like to acknowledge John Bennett for his thorough editing.
This work was supported by State funding from The French National Research Agency (ANR) under “Investissements d’Avenir” programs (Reference: ANR-10-IAHU-01) and ANR PractikPharma grant (ANR-15-CE23-0028).

Author Contributions

contributions to the conception or design of the work: AN, IL, AB, NG, BR
acquisition, analysis, or interpretation of data: AN, IL, WD, NP, RT, NG, BR
creation of new software used in the work: AN, IL, WD, NP, AR, DB, NG, BR
have drafted the work or substantively revised it: AN, IL, AB, NG, RT, BR, KBC

Competing Interests statement

The authors have no competing interests to declare.

Data availability

Data supporting this study can be made available on request, on condition that a research project is accepted by the scientific and ethics committee of the AP-HP health data warehouse (https://recherche.aphp.fr/eds/recherche/).

Abbreviations

AP-HP: Greater Paris University Hospital, Assistance publique - Hôpitaux de Paris
CCB: calcium channel blocker
CDM: common data model
EHR: electronic health record
HBP: high blood pressure
NLP: natural language processing
OMOP: Observational Medical Outcomes Partnership
UMLS: unified medical language system
JMIR: Journal of Medical Internet Research

References

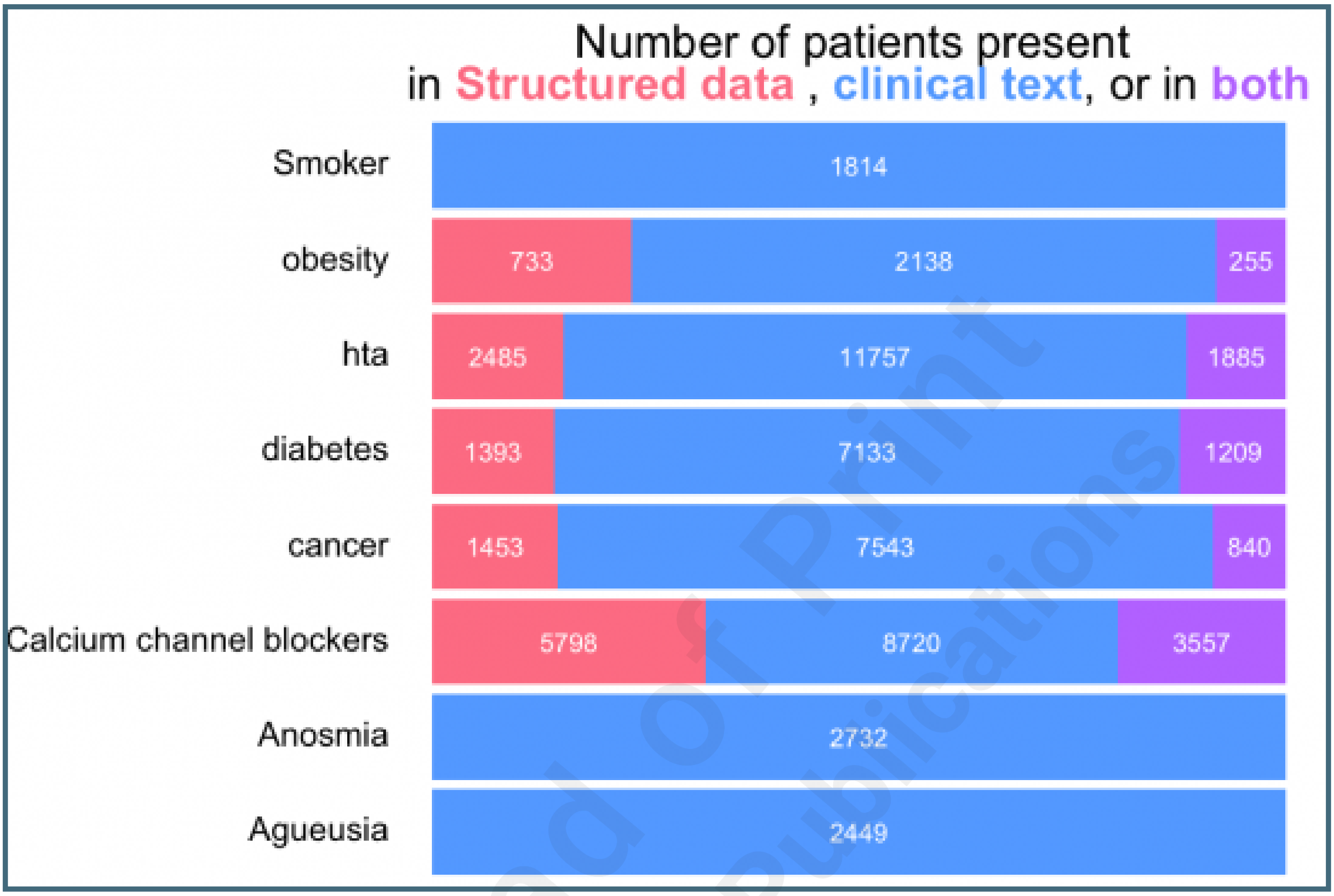
1. Snow J. On the mode of communication of cholera. John Churchill; 1855.
2. Chapman WW, Dowling JN, Ivanov O, Gesteland PH, Olszewski R, Espino JU, Wagner MM. Evaluating natural language processing applications applied to outbreak and disease surveillance.

Proceedings of 36th symposium on the interface: computing science and statistics Citeseer; 2004.
3. Elkin PL, Froehling DA, Wahner-Roedler DL, Brown SH, Bailey KR. Comparison of natural language processing biosurveillance methods for identifying influenza from encounter notes. Annals of Internal Medicine American College of Physicians; 2012;156(1_Part_1):11–18.
4. Zhang L, Sun Y, Zeng H-L, Peng Y, Jiang X, Shang W-J, Wu Y, Li S, Zhang Y-L, Yang L, Chen H, Jin R, Liu W, Li H, Peng K, Xiao G. Calcium channel blocker amlodipine besylate is associated with reduced case fatality rate of COVID-19 patients with hypertension. medRxiv Cold Spring Harbor Laboratory Press; 2020 Apr 14;2020.04.08.20047134. [doi: 10.1101/2020.04.08.20047134]
5. Devlin J, Chang M-W, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:181004805 [cs] [Internet] 2018 Oct 10 [cited 2018 Nov 17]; Available from: http://arxiv.org/abs/1810.04805
6. Lindberg DA, Humphreys BL, McCray AT. The Unified Medical Language System. Methods Inf Med 1993;32:281–291.
7. Okazaki N, Tsujii J. Simple and Efficient Algorithm for Approximate Dictionary Matching. Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010) [Internet] Beijing, China: Coling 2010 Organizing Committee; 2010 [cited 2020 May 8]. p. 851–859. Available from: https://www.aclweb.org/anthology/C10-1096
8. Lample G, Ballesteros M, Subramanian S, Kawakami K, Dyer C. Neural Architectures for Named Entity Recognition. Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies [Internet] San Diego, California: Association for Computational Linguistics; 2016 [cited 2018 Sep 24]. p. 260–270. [doi: 10.18653/v1/N16-1030]
9. Jouffroy J, Feldman SF, Lerner I, Rance B, Burgun A, Neuraz A. MedExt: combining expert knowledge and deep learning for medication extraction from French clinical texts (Preprint) [Internet]. Journal of Medical Internet Research; 2020 Jan. [doi: 10.2196/preprints.17934]
10. Akbik A, Bergmann T, Blythe D, Rasul K, Schweter S, Vollgraf R. FLAIR: An Easy-to-Use Framework for State-of-the-Art NLP. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations) [Internet] Minneapolis, Minnesota: Association for Computational Linguistics; 2019 [cited 2020 Jun 30]. p. 54–59. [doi: 10.18653/v1/N19-4010]
11. Garcelon N, Neuraz A, Benoit V, Salomon R, Burgun A. Improving a full-text search engine: the importance of negation detection and family history context to identify cases in a biomedical data warehouse. J Am Med Inform Assoc 2017 May 1;24(3):607–613. PMID:28339516
12. Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, Suchard MA, Park RW, Wong ICK, Rijnbeek PR, van der Lei J, Pratt N, Norén GN, Li Y-C, Stang PE, Madigan D, Ryan PB. Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. Stud Health Technol Inform 2015;216:574–578. PMID:26262116
13. Cox DR. Regression Models and Life-Tables. Journal of the Royal Statistical Society Series B (Methodological) [Royal Statistical Society, Wiley]; 1972;34(2):187–220.
14. European Medicines Agency. Strategic reflection. EMA Regulatory Science to 2025 2018;79.
15. Pizer SD. Falsification Testing of Instrumental Variables Methods for Comparative Effectiveness Research. Health Services Research 2016;51(2):790–811. [doi: 10.1111/1475-6773.12355]

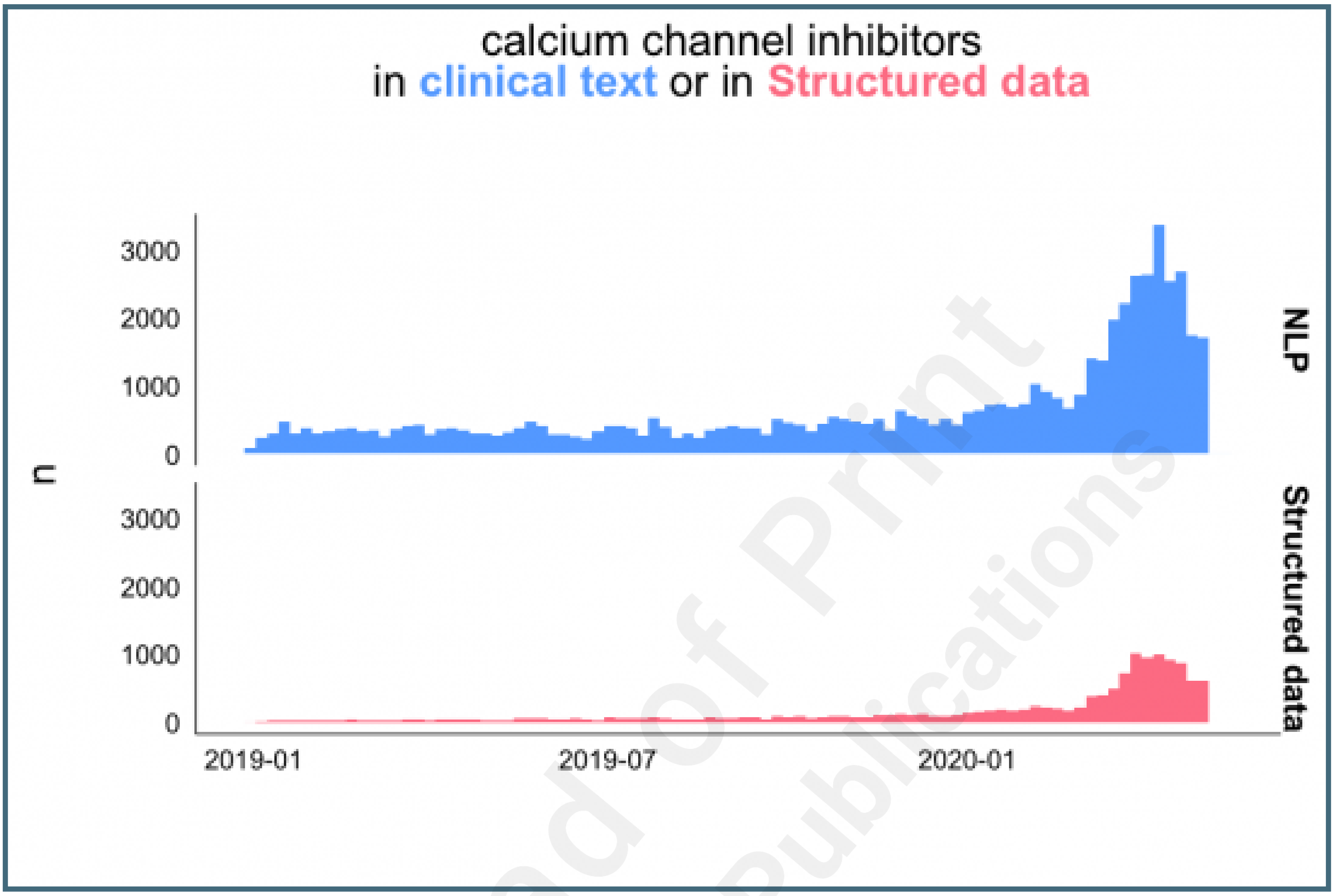
Supplementary Files

Figures

Quantity of patients with information for a selection of items depending on the source of data.



Quantity of information about calcium channel blockers over time.



Multimedia Appendixes

Supplementary methods.
URL: http://asset.jmir.pub/assets/aa72272b73a9727890f894aab42ada71.docx

Definition of calcium channel blockers (name, ATC).
URL: http://asset.jmir.pub/assets/9583a5dacce5a18ebaedfa1b44544a61.docx

Definition of phenotypes (name, ICD10 codes).
URL: http://asset.jmir.pub/assets/7376a5006c1f7057da354d71e4007453.docx

Characteristics of the population of COVID positive patients with hypertension in EDS-COVID.
URL: http://asset.jmir.pub/assets/b7534d7ed84d14b15ac3ae12dda30eca.docx

Flowchart of the use case: COVID positive patients with hypertension.
URL: http://asset.jmir.pub/assets/0054ce1f22c849eefb9a61863232f41d.docx

Description of the natural language processing pipeline.
URL: http://asset.jmir.pub/assets/b5cc7f1a3f062e8d09170d6e28366129.docx

Examples of regular expression for the extraction of phenotypes.
URL: http://asset.jmir.pub/assets/401d711a44362c374ee42048164237b6.docx

Performances of the medication information extraction model before and after normalization of the entities.
URL: http://asset.jmir.pub/assets/ebf1648dddc7addbe51ef818e572b857.docx