Official journal
of the Spanish Society
of Chemotherapy

Original

Revista Española de Quimioterapia
doi:10.37201/req/032.2023

Adrián Téllez Santoyo[1,4*]
Carlos Lopera[2]
Andrea Ladino Vásquez[3]
Ferran Seguí Fernández[3]
Ignacio Grafiá Pérez[1]
Mariana Chumbita[2]
Tomasso Francesco Aiello[2]
Patricia Monzó[2]
Olivier Peyrony[2,5]
Pedro Puerta-Alcalde[2,4]
Celia Cardozo[2]
Nicole Garcia-Pouton[2]
Pedro Castro[1,4]
Sara Fernández Méndez[1,4]
José María Nicolas Arfelis[1,4]
Alex Soriano Viladomiu[2,4]
Carolina Garcia-Vidal[2,4*]

Identifying the most important data for research in the field of infectious diseases: thinking on the basis of artificial intelligence

[1]Medical Intensive Care Unit, Hospital Clínic-IDIBAPS, Barcelona, Spain
[2]Department of Infectious Diseases, Hospital Clínic-IDIBAPS, Barcelona, Spain
[3]Department of Internal Medicine, Hospital Clínic-IDIBAPS, Barcelona, Spain
[4]University of Barcelona, Spain
[5]Emergency Department, Hôpital Saint Louis, Assistance Publique-Hôpitaux de Paris, Paris, France.

Article history
Received: 1 April 2023; Revision Requested: 15 May 2023; Revision Received: 7 June 2023; Accepted: 14 June 2023;
Published: 12 August 2023

ABSTRACT

Objectives. Clinical data on which artificial intelligence (AI) algorithms are trained and tested provide the basis to improve diagnosis or treatment of infectious diseases (ID). We aimed to identify important data for ID research to prioritise efforts being undertaken in AI programmes.

Material and methods. We searched for 1,000 articles from high-impact ID journals on PubMed, selecting 288 of the latest articles from 10 top journals. We classified them into structured or unstructured data. Variables were homogenised and grouped into the following categories: epidemiology, admission, demographics, comorbidities, clinical manifestations, laboratory, microbiology, other diagnoses, treatment, outcomes and other non-categorizable variables.

Results. 4,488 individual variables were collected, from the 288 articles. 3,670 (81.8%) variables were classified as structured data whilst 818 (18.2%) as unstructured data. From the structured data, 2,319 (63.2%) variables were classified as direct—retrievable from electronic health records—whilst 1,351 (36.8%) were indirect. The most frequent unstructured data were related to clinical manifestations and were repeated across articles. Data on demographics, comorbidities and microbiology constituted the most frequent group of variables.

Conclusions. This article identified that structured variables have comprised the most important data in research to generate knowledge in the field of ID. Extracting these data should be a priority when a medical centre intends to start an AI programme for ID. We also documented that the most important unstructured data in this field are those related to clinical manifestations. Such data could easily undergo some structuring with the use of semi-structured medical records focusing on a few symptoms.

Keywords: artificial intelligence, structured data, natural language processing, semi-structured medical reports.

Identificación de los datos más importantes para el desarrollo de inteligencia artificial en el campo de las enfermedades infecciosas

Objetivos. Los datos clínicos sobre los que se entrenan y prueban los algoritmos de inteligencia artificial (IA) proporcionan la base para mejorar el diagnóstico o el tratamiento de las enfermedades infecciosas (EI). Nuestro objetivo es identificar datos importantes para la investigación de las enfermedades infecciosas con el fin de priorizar los esfuerzos realizados en los programas de IA.

Material y métodos. Se buscaron 1.000 artículos de revistas de EI de alto impacto en PubMed, seleccionando 288 de los últimos artículos en 10 revistas de primer nivel. Los clasificamos en datos estructurados o no estructurados. Las variables se homogeneizaron y agruparon en las siguientes categorías: epidemiología, ingreso, demografía, comorbilidades, manifestaciones clínicas, laboratorio, microbiología, otros diagnósticos, tratamiento, desenlace y otras variables no categorizables.

Resultados. Se recogieron 4.488 variables individuales, procedentes de 288 artículos. 3670 (81,8%) variables se clasificaron como datos estructurados, mientras que 818 (18,2%) como datos no estructurados. De los datos estructurados, 2.319 (63,2%) variables se clasificaron como directas -recuperables a partir de historias clínicas electrónicas-, mientras que 1.351 (36,8%) fueron indirectas. Los datos no estructurados más frecuentes estaban relacionados con las manifestaciones clínicas y se repetían en todos los artículos. Los datos sobre demografía, comorbilidades y microbiología constituyeron el grupo más frecuente de variables.

Correspondence:
Carolina Garcia-Vidal, MD, PhD. Infectious Diseases Department, Hospital Clínic-IDIBAPS, Barcelona, Spain. Carrer de Villarroel 170, 08036, Barcelona, Spain.
E-mail: cgarciav@clinic.cat and carolgv75@hotmail.com

*Both authors equally contributed to the paper.

Conclusiones. Este artículo identificó que las variables estructuradas han constituido los datos más importantes en la investigación para generar conocimiento en el campo de la EI. La extracción de estos datos debería ser una prioridad cuando un centro médico pretende iniciar un programa de IA para la EI. También hemos documentado que los datos no estructurados más importantes en este campo son los relacionados con las manifestaciones clínicas. Estos datos podrían estructurarse fácilmente con el uso de historias clínicas semiestructuradas centradas en unos pocos síntomas.

Palabras clave: inteligencia artificial, datos estructurados, procesamiento del lenguaje natural, informes médicos semiestructurados.

INTRODUCTION

Artificial intelligence (AI) and personalised clinical care will be a forthcoming revolution in medicine [1–4]. The key to making this event possible is having high-quality data that can feed AI algorithms to achieve personalised diagnoses and help treat diseases. In the area of healthcare, an exponential amount of data is generated as each second passes and the potential for its use becomes greater [5]. However, there is little information on what data are integral to developing medical research. This point is of vital importance for several reasons: 1) to focus initial efforts on building AI programmes in Medicine with highly significant data; 2) to choose the best data extraction strategies for electronic health records (EHRs); and 3) to analyse the difficulties involved in assessing data quality. After analysing thousands of articles from the most important journals in our area of expertise—infectious diseases (ID)—our aim was to explore what kind of data are relevant for ID research to prioritise our retrieving EHR data model, establish the best systematic collection of such information, and determine the order of importance when developing semi-structured EHR systems.

MATERIAL AND METHODS

We perform a transversal and descriptive study to identify the most relevant data used in 10 different journals.

Screening. We screened 1,000 articles (a hundred of which were the latest in published articles across 10 different, high-impact journals specialised in either infectious diseases, tropical medicine, general medicine or multidisciplinary with articles published in the field of infectious diseases). High-impact journal was defined as a journal with the highest Journal Impact Factor. The Journal Impact Factor is a metric that quantifies the average number of citations received per article published in a specific journal within a designated period. The journals screened were The Lancet, The Lancet Infectious Diseases, The New England Journal of Medicine, Journal of the American Medical Association, Clinical Infectious Diseases, Clinical Microbiology and Infection, Emerging Infectious Diseases, The Journal of Infectious Diseases, PLOS Neglected Tropical Diseases and PLOS One.

We used the Web of Science Database with the next strategy search for each journal: SO= ("JOURNAL X") Refined by: RESEARCH AREAS: (INFECTIOUS DISEASES) AND DOCUMENT TYPES: (ARTICLE OR CLINICAL TRIAL) Timespan: 2018-2019. Databases: MEDLINE Search language=Auto. The search was conducted on 19/05/2019.

Inclusion criteria. Articles included for variable recollection had to fulfil the following inclusion criteria: 1) related to ID area in humans and 2) original studies with either clinical trials, observational/case-control design, or propensity score analysis, which presented clinical or epidemiological outcomes. A consensus team (AT, AL, FS, IG, CL and CGV) made decisions regarding the inclusion of certain articles when discrepancies arose. Exclusion criteria included non-related ID articles, animal studies, basic or microbiologic research, case reports, reviews, guidelines, letters or other types of non-original editorial research.

Variables and definitions. Variables from each article were introduced into a new database manually. Dichotomic, ordinal, discrete and continuous variables were considered as one variable, whilst nominal variables as one different variable for each instance. Variables were classified into either structured or unstructured data depending on whether the variable-in-question was retrievable in a structured table from our EHR. Additionally, when we could obtain data directly from our EHR, such structured data was classified as direct. In cases when an algorithm was needed, structured data was classified as indirect.

Lastly, variables were homogenised according to the nature of the variable. Therefore, homogenized variables were defined as variables that share the same data, but they are expressed differently in the articles. For example, here is a series of variables re-grouped within a category: age ≥65, age of children, age at delivery, age at diagnosis, etc, all of them were homogenized within the variable age. Finally, variable groups were rearranged according to the healthcare workflow within ID processes. The healthcare workflow was defined as the usual workflow used in the care of patients in the clinical field. It were defined by consensus eleven different healthcare workflow categories: epidemiology (e.g., incidence, prevalence, mortality, etc.), admission (e.g., admission unit, length of stay, etc.), demographics (e.g., age, sex, etc.), comorbidities, clinical manifestations, laboratory (e.g., blood test, serology, etc.), microbiology (e.g., isolated microorganism, antibiotic susceptibility, etc.) other diagnoses (e.g., pathology, images, electrocardiogram, etc.), treatment (e.g., antibiotics, antibiotic duration, etc.), outcomes (e.g., survival, cure rate, etc.) and other non-categorizable variables.

Statistical analysis. The qualitative variables were described as absolute and relative frequencies. The analysis was performed using SPSS version 24.0 software.

Ethics. No ethics approval was necessary due to the nature of our study.

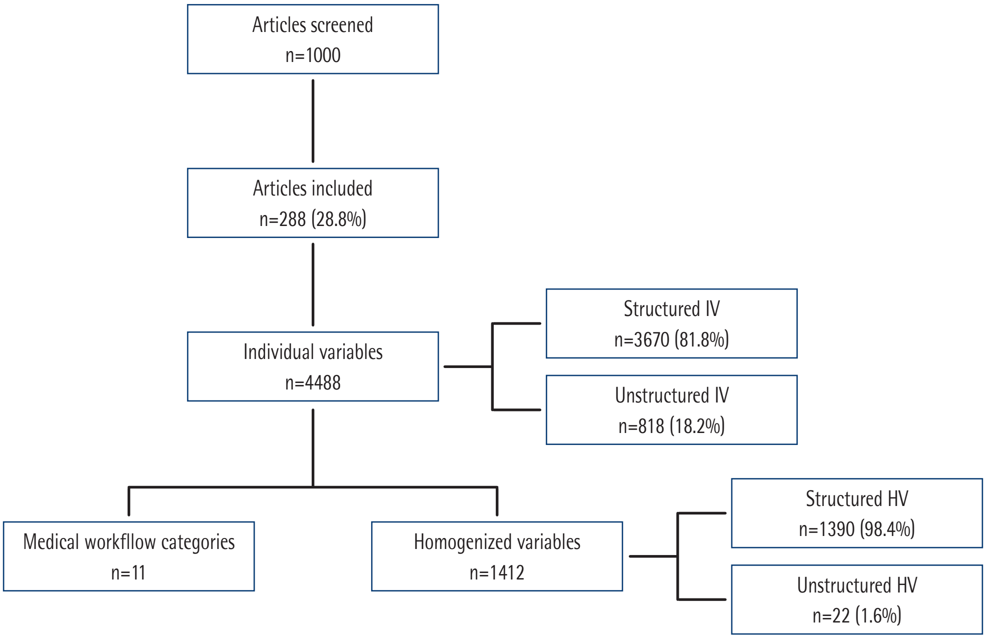

Figure 1 Workflow and processing of variables.

Table 1 Articles included from each journal.

Journal	n (%)
Clinical Infectious Diseases	38 (38)
Clinical Microbiology and Infection	26 (26)
Emerging Infectious Diseases	23 (23)
Journal of the American Medical Association	18 (18)
Journal of Infectious Diseases	41 (41)
The Lancet	25 (25)
The Lancet Infectious Diseases	38 (38)
The New England Journal of Medicine	42 (42)
PLOS Neglected Tropical Diseases	19 (19)
PLOS One	18 (18)

RESULTS

Screening. Of the 1,000 articles screened, 288 (28.8%) articles were selected for the study per criteria. From these articles, we collected 4,489 variables from our database (Figure 1) for final analysis. Table 1 shows those articles included, whilst Supplementary Table 1 provides specifications about each one included. From the 288 included articles, 110 (38%) were clinical trials, 149 (52%) observational studies, 14 (5%) case-control studies, 10 (3%) propensity score analysis and 5 (2%) other studies.

The most frequent data in the whole database. A total of 4,488 individual variables were collected from the 288 articles. Of these, 3,670 (81.8%) variables were classified as structured data and 818 (18.2%) as unstructured data. From the structured data, 2,319 (63.2%) variables were classified as direct, whilst the other 1,351 (36.8%) variables as indirect. Supplementary Table 2 describes all data collected, as well as homogenised variables (HV) and grouped variables per medical workflow details.

Homogenised variables. After homogenising the 4,488 variables, we obtained a total of 1,412 HV. When each of these HV were considered as one independent variable, 1,390 (98.4%) were structured data and 22 (1.6%) unstructured data. The twenty most frequent HV comprises 32.2% of all HV. Table 2 shows these twenty most frequent HV and their structured or unstructured data classification status.

Medical workflow categories. A rearrangement of the individual variables according to the medical workflow was conducted, and a total of 11 medical workflow categories (MWC) were created. These MWC were also classified into structured and unstructured data. From these MWC, demographics, comorbidity and microbiology were the most frequent variables. Table 3 shows the most common MWC and describes the frequency of each MWC as structured or unstructured data.

Table 2 The twenty most frequent homogenised grouped variables

Ranking	Grouped variable	Structured data in our EHR	Frequency n (%)
1	Clinical manifestations	No	325 (7.2)
2	Age	Yes	251 (5.6)
3	Gender	Yes	212 (4.7)
4	Race	No	90 (2.0)
5	Economic or work demographics	No	54 (1.2)
6	Other (non-homogenising variables)	No	47 (1.0)
7	Antimicrobial susceptibility	Yes	46 (1.0)
8	Body mass index	Yes	46 (1.0)
9	Housing characteristic demographics	No	45 (1.0)
10	Sexual behaviour demographics	No	44 (1.0)
11	Diabetes	Yes	40 (0.9)
12	Other comorbidities	Yes	37 (0.8)
13	CD4 count	Yes	36 (0.8)
14	Education demographics	No	28 (0.6)
15	Region	Yes	28 (0.6)
16	White blood cell count	Yes	28 (0.6)
17	Creatinine	Yes	26 (0.6)
18	Microorganisms	Yes	25 (0.6)
19	Country	Yes	24 (0.5)
20	Diagnosis	Yes	24 (0.5)

DISCUSSION

To our knowledge, this is the first study to detail which variables were the most important in the best recently published medical research to improve the management, diagnosis, treatment and/or outcomes of patients with ID.

These insights are essential when considering different objectives within the field of ID. Firstly, developing an AI programme is expensive and much of the cost may come from data collection [6]. Identifying which variables have contributed to generating the most current knowledge on the most important research topics in the field can help prioritise data extraction. It can also result in conceiving optimal strategies for retrieving such data from EHR.

Secondly, our study has identified that the most important variables in developing ID research are those that are structured. In fact, they represent more than 81% of the variables used in the articles reviewed. This information is of vital importance, because it facilitates the initial steps when creating and implementing AI programmes in hospitals. Obtaining structured data is less costly in terms of time and money than retrieving unstructured data. Moreover, finding objective criteria to check data quality is more feasible.

Lastly, our study provides new information on the most important unstructured variables used in advancing the field of ID. Currently, the only way to collect these variables is through natural language processing (NLP). NLP is complex and have significant shortcomings in medicine [7]. An objective analysis of this information arises due to the subjective perception of some clinical manifestations by patients; physicians' varying ways of writing clinical courses and discharge reports, including multiple acronyms and/or different languages; and the unpredictable and ambiguous nature of medical records. Therefore, the role of NLP has become extremely limited. However, our study has determined that many unstructured variables used in the ID research are related to clinical manifestations. This finding may help create a semi-structured clinical course for physicians.

Our study has some limitations. We performed the study in a single hospital, which has an EHR system based on SAP. Other hospitals with other EHR programmes may have structured data unavailable to us or, inversely, have data as unstructured that would otherwise be structured in our case.

Table 3 Frequency of medical workflow categories and classification status as structured and unstructured data.

Medical workflow categories Total variables = 4,488	Structured data in our EHR n (%)	Unstructured data in our EHR n (%)
Epidemiology	203 (4.5)	185 (4.1)
Admission	84 (1.9)	0
Demographics	664 (14.8)	251 (5.6)
Comorbidities	547 (12.2)	9 (0.2)
Clinical manifestations	195 (4.3)	325 (7.2)
Laboratory	317 (7.1)	0
Microbiology	513 (11.4)	13 (0.3)
Other diagnosis	477 (10.6)	11 (0.2)
Treatment	487 (10.9)	2 (0)
Outcomes	180 (4)	21 (0.5)
Other	1 (0)	3 (0.1)

Furthermore, our study focuses on what data are necessary to generate knowledge in the field of ID. Determining how to ensure that these high-quality data are retrievable should be the subject of future research.

To conclude, our study identified the most important variables that have been used in research to build knowledge in the field of ID. As methodologic approaches for obtaining unstructured data improves, healthcare programmes aimed at implementing AI in ID can work on extracting high-quality structured data. With these data, computer scientists and clinicians could strengthen a powerful base by which to develop potentially useful artificial intelligence algorithms in current medical practice. We also documented that the most significant unstructured data in ID are related to clinical manifestations and could be easily structured with the use of semi-structured medical records focusing on a few symptoms.

ACKNOWLEDGEMENTS

Authors would like to thank for OP post-doctoral fellow financial support: La Ligue Nationale contre le Cancer (convention number: AAPMRC 2022/OP) and La Direction de l'Assistance Publique - Hôpitaux de Paris (APHP). We also would like to thank Anthony Armenta for providing medical editing assistance for the manuscript at hand.

FUNDING

PP-A [JR20/00012] and CG-V [Proyecto FIS PI21/01640, funded by Instituto de Salud Carlos III (ISCIII) and co-funded by the European Union] have received research grants. The funders had neither a specific role in study design or collection of data, nor in the writing of the paper or the decision to submit.

CONFLICT OF INTERESTS

PP-A has received honoraria for talks on behalf of Merck Sharp and Dohme, Gilead, Lilly, ViiV Healthcare and Gilead Science. AS has received honoraria for talks on behalf of Merck Sharp and Dohme, Pfizer, Novartis, Menarini, Angellini, as well as grant support from Pfizer. PC has received honoraria for talks on behalf of Gilead Science, MSD, Pfizer, Janssen and Alexion. CG-V has received honoraria for talks on behalf of Gilead Science, MSD, Novartis, Pfizer, Janssen, GSK, and Menarini, as well as a grant from Gilead Science, Pfizer and MSD.

REFERENCES

1. Reddy S, Fox J, Purohit MP. Artificial intelligence-enabled healthcare delivery. J Roy Soc Med. 2018;112(1):22–8. DOI: 10.1177/0141076818815510
2. Desai AN. Artificial Intelligence: Promise, Pitfalls, and Perspective. JAMA. 2020;323(24):2448–9. DOI: 10.1001/jama.2020.8737
3. Garcia-Vidal C, Sanjuan G, Puerta-Alcalde P, Moreno-García E, Soriano A. Artificial intelligence to support clinical decision-making processes. Ebiomedicine. 2019;46:27–9. DOI: 10.1016/j.ebiom.2019.07.019
4. Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. FuturHealthc J. 2019;6(2):94–8. DOI: 10.7861/futurehosp.6-2-94
5. Coughlin S, Roberts D, O'Neill K, Brooks P. Looking to tomorrow's healthcare today: a participatory health perspective. Intern Med J. 2018;48(1):92–6. DOI: 10.1111/imj.13661
6. "Analytics Insight" [cited 2022 Jul 25]. Available from: https://www.analyticsinsight.net/how-much-does-artificial-intelligence-cost-in-2021/
7. Tseng YH, Lin CJ, Lin YI. Text mining techniques for patent analysis. Inform Process Manag. 2007;43(5):1216–47. DOI: 10.1016/j.ipm.2006.11.011