nefrología
Revista de la Sociedad Española de Nefrología
www.revistanefrologia.com

Original article

# The big data era: The usefulness of folksonomy for natural language processing

Laia Sans[a], Ismael Vallvé[b], Joan Teixidó[b], Josep Manel Picas[b], Jordi Martínez-Roldán[c], Julio Pascual[a,*]

[a] Servicio de Nefrología, Hospital del Mar, Barcelona, Spain
[b] Bismart, Barcelona, Spain
[c] Dirección de Innovación y Transformación Digital, Hospital del Mar, Barcelona, Spain

ARTICLE INFO

*Article history:*
Received 28 May 2020
Accepted 15 September 2021
Available online 15 March 2023

*Keywords:*
Big data
Folksonomy
Natural language processing
Nephrology

ABSTRACT

*Background:* A huge amount of clinical data is generated daily and it is usually filed in clinical reports as natural language. Data extraction and further analysis requires reading and manual review of each report, which is a time consuming process. With the aim to test folksonomy to quickly obtain and analyze the information contained in media reports we set up this study.
*Methods and objectives:* We have used folksonomy to quickly obtain and analyze data from 1631 discharge clinical reports from the Nephrology Department of Hospital del Mar, without the need to create a structured database.
*Results:* After posing some questions related to daily clinical practice (hypoglycaemic drugs used in diabetic patients, antihypertensive drugs and the use of renin angiotensin blockers during hospitalization in the nephrology department and data related to emotional environment of patients with chronic kidney disease) this tool has allowed the conversion of unstructured information in natural language into a structured pool of data for its further analysis.
*Conclusions:* Folksonomy allows the conversion of the information contained in clinical reports as natural language into a pool of structured data which can be further easily analyzed without the need for the classical manual review of the reports.

DOI of original article:
https://doi.org/10.1016/j.nefro.2021.09.006.

[*] *Corresponding author.*
E-mail address: julpascual@gmail.com (J. Pascual).

## La era del big data: análisis del lenguaje natural mediante la aplicación de folksonomía

*Palabras clave:*
Big data
Folksonomía
Procesamiento del lenguaje natural
Nefrología

### RESUMEN

*Antecedentes y objetivo:* Gran parte de la información médica que se deriva de la práctica clínica habitual queda recogida en forma de lenguaje natural en los informes médicos. Clásicamente, la extracción de información clínica para su posterior análisis a partir de los informes médicos requiere de la lectura y revisión manual de cada uno de ellos con la consiguiente inversión de tiempo. El objetivo de este proyecto piloto ha sido evaluar la utilidad de la folksonomía para la extracción y análisis rápido de los datos que contienen los informes médicos.

*Material y métodos:* En este proyecto piloto hemos utilizado la folksonomía para el análisis y la rápida extracción de datos de 1.631 informes médicos de alta de hospitalización del Servicio de Nefrología del Hospital del Mar sin necesidad de crear una base de datos estructurada previamente.

*Resultados:* A partir de determinadas preguntas sobre la práctica médica habitual (tratamiento hipoglicemiante de los pacientes diabéticos, tratamiento antihipertensivo y manejo de los inhibidores del sistema renina angiotensina durante el ingreso en nefrología y análisis de datos relacionados con la esfera emocional de los pacientes renales) la herramienta ha permitido estructurar y analizar la información contenida en texto libre en los informes de alta.

*Conclusiones:* La aplicación de folksonomía a los informes médicos nos permite transformar la información contenida en lenguaje natural en una serie de datos estructurados y analizables de manera automática sin necesidad de proceder a la revisión manual de los mismos.

## Introduction

The term ***big data*** refers to a large amount of data whose volume, variability and necessary speed of processing make its analysis very complex using manual systems or standard software for its management.[1,2]

In the health field, millions of data derived from patient care are generated every day. Presently, in our environment, the use of the electronic medical record is widespread and technological advances can facilitate the analysis of the data collected. The analysis of data from medical records allows for quality control of medical actions, as well as obtaining observational data from patient cohorts to generate scientific evidence and select individuals with certain characteristics tha makes them suitable for participation in clinical trials.

Although some of the data obtained are numerical (laboratory data or the collection of constants), most of them are collected in the clinical history of the patients in the form of natural language (for example, the data obtained from the anamnesis of the patient, the physical examination, the treatment, the various complementary examinations or the diagnoses themselves). Transforming all this valuable data collected in natural language into a series of structured data implies a significant investment of time, since it requires manual work consisting on reading the medical history, identifying and obtaining the data that has previously been considered of interest, the generation and addition of data to databases that must be structured in quantitative data (it is required the transformation of the information collected in natural language into numerical variables) and, finally, the analysis of these data. This process, in addition to consuming a significant amount of time, does not allow the reanalysis of new data once the parameters considered of interest in the initial project have been collected, unless the review process and manual data collection is started again. This fact also makes it impossible to reanalyze in real time the new clinical histories that are generated and that are of interest for a particular project; any reconsideration entails redoing the entire manual process.

The technology of *natural language understanding* (NLU) or *natural language processing* (NLP) make it possible to quickly and automatically convert all the information collected in free text into an ordered structure, and thus be able to proceed to make a much faster analysis of all the information.

Most of the systems that perform NLU or NLP require an ontology or master entity to subsequently analyze the documents.[3] In other words, it is necessary to decide in advance which are the terms or labels of interest (before starting to obtain data) and, therefore, they do not allow the discovery of any term that is not already arranged in the ontology (*top down* distribution).

The use of **folksonomy (comes from the terms "*folk*" and "taxonomy")** allows information contained in free text to be obtained without the prior need to generate a master entity of terms of interest, which provides an obvious advantage over traditional systems of NLP. This advanced analytics transforms unstructured text documents into structured text documents, enabling the discovery of information without requiring an initial closed draft of search terms before starting the retrieval of information. Therefore, folksonomy would allow automatic highlighting of natural language concept labels to reveal the internal content. Folksonomy is an automatic classification system in real time, based on tags and the frequency with which they appear, and it is the only viable method to work with huge amounts of documents. The way this system works is known as *bottom up* and the *Bismart Folksonomy solution* is the first software that can manage this type of classification (https://bismart.com/es/inicio/).

The use of NLP algorithms together with folksonomy in the medical field would make it possible to invest no more time in the generation of databases than that required for the usual clinical care activity and the analysis in real-time of the new data collected. In this manner, *big data* would bring significant benefits to the medical sector.[4]

Although there are publications on the use of NLP in the medical field, to our knowledge there are no previous experiences on the application of folksonomy to obtain data in the field of medicine in general, nor in the specific field of Nephrology. In this article we report the first pilot experience in the use of folksonomy together with artificial intelligence in NLP to analyze clinical data of hospital discharge reports during in a given period from the Nephrology Service of the Hospital del Mar de Barcelona, based on some questions eminently related to usual medical practice and examine the performance of this system for automatic data analysis.

## Methods

### Data

A total of 1631 hospital discharge reports were collected from the Hospital del Mar Nephrology Service between 2016 and 2018. The documents were anonymized in PDF format from a computer located at Hospital del Mar, where a Bismart programmer had physical access without internet connection to proceed with the elimination of the headers containing patients affiliation data through an automated process developed by Bismart (Barcelona) using Python language, being able to only identify the gender, necessary for the subsequent application of glomerular filtration estimation formulas. Once the headers of the medical discharge documents were removed, ensuring their anonymity, they were uploaded to the Bismart Folksonomy portal, proceeding to their conversion to text using an OCR system from Microsoft Cognitive Systems, and using pattern detection algorithms, the different fields based on the sections of the reports (diagnoses, reason for consultation, personal history, usual treatment, complementary examinations, evolution and treatment at discharge), to later store the already anonymized information in the Hospital del Mar data cloud. Next, data normalization and lemmatization processes were carried out, algorithms were applied, folksonomization was performed, and the Bismart Folksonomy web portal was installed. Bismart has an automatic process that installs in the chosen data cloud a virtual machine with the database and all the necessary services for the Bismart Folksonomy portal; so that the technical deployment is a relatively simple process. (Fig. 1). The portal complies with all GDPR regulations *(general data protection regulation)* with a registry of accesses and modifications or inquiries made to the data; the time, the user and the IP address is recorded in a *log* of the device.

Medical terms and acronyms specific to the specialty appeared in the documents. Furthermore, the reports were written indistinctly in two languages (Catalan and Spanish) and terms in both languages could even appear in the same report. This added more complexity to the data extraction, but since folksonomy works with terms and not with languages, the creation of synonyms of words between both languages or of words and acronyms allows the identification of the search term regardless of the language.

### The degree of chronic kidney disease

In the field of Nephrology, the classification of the degree of chronic kidney disease (CKD) is very important[5] since it has prognostic and therapeutic implications. The review of the classification of the disease or renal situation according to the information collected in the "diagnoses" section of the discharge reports, only allowed us to identify the degree of CKD in some 300 reports. Due to the fact that the tool allows the addition of synonyms, the words "grau", "estádio" and "estadi" were assigned to the word "grado", which made it possible to find the degree of CKD in 768 reports. To classify the degree of CKD in the rest of the reports, as well as to identify reports with the diagnosis of acute renal failure, algorithms with heuristic rules were generated for the correct identification of the renal disease situation based on: a) the presence of the words "acute renal failure" and synonyms in the diagnosis section, which implied the label of acute renal failure, b) the presence of the words "admission for kidney transplant recipient" and synonyms in the reason for consultation implied the label CKD grade 5, c) the identification of the words "chronic kidney disease grade X" and synonyms among the personal history allowed the reports to be labeled as CKD grades 1–5, d) the use of creatinine in the admission analysis with age and gender (data collected between the anthropometric variables) allowed the calculation of the estimated glomerular filtration rate by entering the CKD-EPI formula[6] in the software *(chronic kidney disease epidemiology collaboration)*. Despite this, the algorithms did not allow the classification of 79 documents in terms of renal status, so the renal status was manually reviewed and assigned in the remaining patients that were not classified. Thus, the application of the software tools (creation of labels, synonyms and algorithms with heuristic rules) allowed the automatic classification of the degree of renal disease despite not being included in the "diagnoses" section of the medical records in 95% of all reports, manual review was required only in 5%. In this way, all reports were classified as: acute renal failure, CKD grades 1–5 or without renal disease (Fig. 2).

### Questions posed as a pilot study

*As a pilot study, three questions were raised*

1 How do we treat diabetic patients with kidney disease? Bearing in mind that metformin is the oral hypoglycemic agent mainly used in the diabetic population, what is our attitude

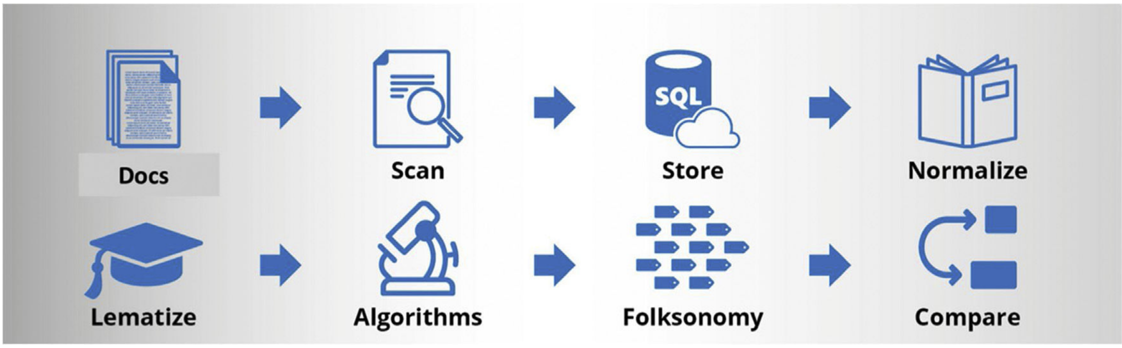

**Fig. 1 – The solution proposed by Bismart is based on a flow of data that begins with the incorporation into the database knowledge of the PDF documents provided by Hospital del Mar. On these documents we apply OCR processes and the definition of the fields that we want to import from each document. Once the fields are stored in the database and the fields have been identified, the system starts the folksonomy process, detecting important words or groups of words in the collection of documents. Once the Folksonomy tool has extracted the information that is needed to work, it is presented in a Web so that it can be consulted, modified or to execute the process again upon request.**

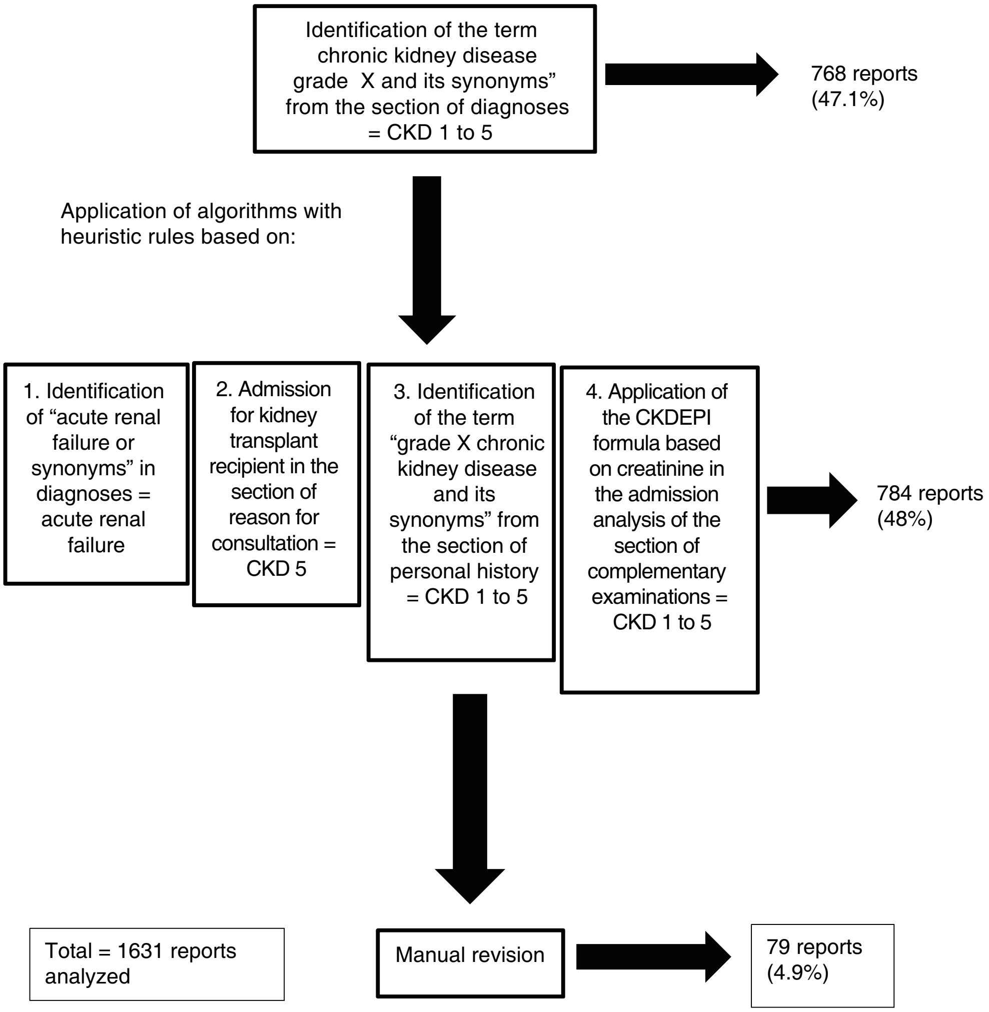
Identification of the term chronic kidney disease grade X and its synonyms" from the section of diagnoses = CKD 1 to 5
768 reports (47.1%)
Application of algorithms with heuristic rules based on:
1. Identification of "acute renal failure or synonyms" in diagnoses = acute renal failure
2. Admission for kidney transplant recipient in the section of reason for consultation = CKD 5
3. Identification of the term "grade X chronic kidney disease and its synonyms" from the section of personal history = CKD 1 to 5
4. Application of the CKDEPI formula based on creatinine in the admission analysis of the section of complementary examinations = CKD 1 to 5
784 reports (48%)
Total = 1631 reports analyzed
Manual revision
79 reports (4.9%)

**Fig. 2 – Algorithm applied to classify the kidney disease status of the reports included in the analysis.**

regarding the prescription of this drug in situations of kidney disease? How many and what are the characteristics of patients diagnosed with lactic acidosis due to metformin?

2 How do we treat arterial hypertension in patients with chronic kidney disease? Are renin angiotensin system inhibitors the most widely used type of drugs considering their benefits on nephroprotection? What is the attitude of

the nephrologists of the Hospital del Mar Nephrology Service in relation to the withdrawal or maintenance of **renin angiotensin system inhibitors** in patients admitted to the Nephrology Service?

3 What is the percentage of nephrology admissions that receive some **hypnotic/sedative/antidepressant treatment** despite not being included diagnosis of this pathology in the patient's clinical history?

## Results

### *Treatment of diabetes mellitus*

Despite the fact that metformin continues to be the most widely used oral hypoglycemic agent with the greatest evidence of efficacy in the treatment of type 2 diabetes due to its benefits in terms of morbidity and mortality, its use in patients with kidney disease is restricted.[7] In CKD, the dose must be adjusted and its use is contraindicated in advanced CKD, as it may be associated with the presence of lactic acidosis, especially when administered to patients with a glomerular filtration rate less than 30 mL/min/1.73 m$^2$.[8] Thus, in a situation of advanced CKD, only other types of oral hypoglycemic agents can be used or it may be considered the use of insuline. The treatment received by diabetic patients admitted to the Hospital del Mar Nephrology service and the characteristics of patients diagnosed with metformin-induced lactic acidosis were analyzed.

Diabetic patients were identified based on the presence of this diagnosis in the "diagnoses" section of the discharge reports. Thus, it was identified a percentage of reports with a diagnosis of diabetes lower than expected, for which reason the search was expanded with new heuristic rules, assigning the diagnosis of diabetes to those reports in which some hypoglycemic drug was found in the treatment. Given the large number of hypoglycemic agents available on the market, the inclusion of each of them individually in the searches generated greater complexity for the project. Thus, to label each drug, it was decided to use the ATC (*Anatomical Therapeutic Chemical classification system*) classification of the Spanish Medicines Agency (AEMPS), which groups these compounds by the active ingredient. The ATC classification provides information on both the trade name and the active ingredient, and in the data analysis process the entire text was searched for either one. In the case of identifying a trade name, thanks to the ATC classification, the active principle and the ATC group to which it corresponds can be inferred by applying graph analysis algorithms. It must be taken into account that there are ATC groups that can contain two or more active ingredients, so this logic was added to the detection algorithm.

In the case of drugs for the treatment of diabetes, these correspond to group A, subgroup A10 of the ATC classification.

Finally, there were identified 637 of the 1631 reports with the diagnosis of diabetes (39.05% of the reports). Table 1 shows the hypoglycemic treatments received by these patients according to their degree of renal disease (subgroups were A10A for insulin, A10BA for metformin, A10BH for DPP4 inhibitors, A10BX for repaglinide and transporter the inhibitor SGLT2 –requiring search by trade name and active principle-, A10BB for derivatives of sulfonylureas and A10BG for thiazolidinediones). Thus, the most widely used treatment in these renal patients with diabetes is insulin (337 reports contained insulin in the usual treatment on admission), followed by metformin (85 reports contained metformin in the medication on admission). In five of these 85 reports, the term "lactic acid" and its synonyms were identified in the diagnostic section, thus revealing five cases of metformin-induced lactic acidosis (three episodes in the context of acute renal failure, one patient with stage 3 CKD, and one patient with stage 4 CKD). In all cases, the reason for consultation turned out to be acute gastroenteritis (four with diarrhea and one with emetic syndrome). In addition, in all cases, except for the patient with stage 3 CKD, hemodialysis was required in the context of impaired renal function and lactic acidosis due to metformin.

In 102 reports, it was not detected any hypoglycemic treatment, so it was concluded that 16% of the patients followed only dietary treatment for their diabetes.

### *Treatment of hypertension*

The prevalence of arterial hypertension in kidney disease is high; both pathologies coexisting in 80%–85% of patients with kidney disease.[9]

The renal and cardiovascular benefits of renin angiotensin system inhibitors (RAS inhibitors) in patients with CKD have been widely demonstrated[10] and they should be part of the treatment of arterial hypertension in renal patients. However, in situations of acutly decompensated renal function, they are usually withdrawn. The delay in their reintroduction once the decompensating episode has resolved could imply a worsening of the prognosis of our patients.[11]

It was analyzed the antihypertensive treatment received by patients diagnosed with arterial hypertension and admitted to the Hospital del Mar Nephrology Service. The diagnosis of arterial hypertension was identified in 1520 of the 1631 available reports (93.19% of the reports). For this, the synonyms "arterial hypertension" and "HTN" were included under the label "arterial hypertension". The antihypertensive drugs that patients received separated by groups (atc C02A were centrally acting antihypertensives, atc C02C for doxazosin, atc C02DB for hydralazine, atcC03A for diuretics: 03AA hydrochlorothiazide, 03BA chlorthalidone and indapamide, 03CA loop diuretics and 03DA for antialdosterones; atc C07 for beta-blockers, atc C08CA dihydropyridine calcium antagonists and atc C08D non-dihydropyridines, and atc C09 for renin angiotensin system inhibitors). On admission, diuretics were the most commonly used drugs (in 754 reports there was at least one diuretic in the usual medication, mostly [549] loop diuretics). It should be noted that 34 of the 126 reports that recorded the administration of thiazide and related diuretics corresponded to patients with stage 5 CKD, a degree of kidney disease in which the diuretic effect of these drugs is lost. The second most widely used antihypertensives were the family of beta-blockers (611 reports), followed by dihydropyridine calcium antagonists (532 reports) and then renin-angiotensin system inhibitors (437 reports). Admission modified the pattern of antihypertensive administration at discharge, both in terms of the total number of antihypertensive drugs prescribed (2588 at admission and 2758 at discharge) and in the

**Table 1 – Number of reports containing the different therapeutic options for the treatment of diabetes in patients admitted to the Nephrology Department.**

| | INS | MTF | – DPP4 | RGL | SLN | –SGLT2 _ | GLP1 | TZN |
|---|---|---|---|---|---|---|---|---|
| FRA | 78 | 16 | 5 | 0 | 3 | 0 | 0 | 0 |
| CKD 1 | 12 | 10 | 5 | 0 | 1 | 0 | 0 | 0 |
| CKD 2 | 15 | 21 | 2 | 2 | 0 | 0 | 0 | 0 |
| CKD 3 | 46 | 32 | 8 | 8 | 1 | 2 | 2 | 0 |
| CKD 4 | 74 | 6 | 8 | 13 | 0 | 0 | 0 | 1 |
| CKD 5 | 182 | 0 | 27 | 24 | 1 | 0 | 0 | 0 |
| Total | 337 | 85 | 55 | 47 | 6 | 2 | 2 | 1 |

ARF: acute renal failure; CKD: chronic kidney disease; INS: insulin; MTF: metformin; –DPP4: dipeptyl peptidase 4 inhibitors; RGL: repaglinide; SLN: sulfonylurea derivatives; –SGLT2: sodium glucose transporter inhibitors; GLP1: glucagon-like peptide type 1 agonists; TZN: thiazolidinediones.

**Table 2 – Prescription of types of antihypertensive drugs in medication on admission and on discharge.**

| | Medication on admission | Discharge medication |
|---|---|---|
| Loop diuretics | 549 | 585 |
| Thiazide and related diuretics | 126 | 112 |
| Potassium-sparing diuretics | 79 | 94 |
| Beta blockers | 611 | 724 |
| Dihydropyridine calcium antagonists | 532 | 639 |
| Non-dihydropyridine calcium antagonists | 18 | 10 |
| RAS inhibitors | 437 | 317 |
| Alpha blockers | 153 | 177 |
| Hydralazine | 78 | 97 |
| Central acting antiadrenergics | 5 | 3 |
| Total | 2588 | 2758 |

ARS: renin angiotensin system.

**Table 3 – Renal status and number of reports of hypertensive patients receiving treatment with RAS inhibitors on admission (column 2) and discharge (column 3) and the percentage reduction in the prescription at discharge.**

| Total reports of patients with RAS inhibitors at admission (n = 437) | n (on admission) | n (on discharge) | % reduction |
|---|---|---|---|
| Acute renal failure | 24 (5.5%) | 18 | 25% |
| Grade 1 chronic kidney disease | 30 (6.9%) | 20 | 33.3% |
| Grade 2 chronic kidney disease | 22 (5.0%) | 19 | 13.6% |
| Grade 3 chronic kidney disease | 93 (21.3%) | 68 | 26.9% |
| Grade 4 chronic kidney disease | 70 (16%) | 48 | 31.4% |
| Grade 5 chronic kidney disease | 193 (44.2%) | 76 | 60.6% |
| Without chronic kidney disease [*] | 5 (1.1%) | 5 | 0% |

ARS: renin angiotensin system.

[*] Admission for adrenal catheterization as a study of primary hyperaldosteronism.

increase or reduction of prescription of certain families of antihypertensives, as shown in Table 2. In 437 of the 1520 reports with a diagnosis of arterial hypertension (28.75%), a drug belonging to the C09 group according to the ATC classification (ARS inhibitors) was identified in the usual treatment. The renal situation of the patients receiving treatment with this group of drugs on admission is shown in Table 3 (column 2). The same table (column 3) shows the number of reports that continued to receive RAS inhibitors at discharge, while the last column of the table shows the percentage reduction in the prescription of renin-angiotensin system inhibitors at the time of discharge from Nephrology. A withdrawal of RAS inhibitors stood out in a high percentage of discharge reports, showing an increasing trend depending on the degree of chronic kidney disease (from stages 3–5). In this study, the causes of withdrawal of these drugs were not analyzed, although knowing the usual clinical practice, it was probably related to the acute deterioration of renal function, more likely to observe with more advanced the renal disease. Taking this reasoning into account, it did not seem plausible with routine clinical practice that the percentage of reports maintaining treatment with RAS inhibitors at discharge among those classified as “acute renal failure” would be so high (only 25% reduction in prescription at discharge). For this reason, these reports were manually reviewed. The manual review made it possible to detect words such as “suspend” or “modify” in front of RAS inhibitor drugs, so that, in the acute renal failure group, only seven reports actually maintained the treatment at discharge, having therefore been erroneously detected and as false positives 11 reports.

### *Emotional health*

There are studies showing that the prevalence of depression symptoms in patients with CKD is high and that psychosocial variables play an important role in the perception of quality of life in kidney patients.[12,13] However, in the the clinical care of nephrology departments, psychological cures of our patients continue to be relegated to a secondary priority.

A search was made for those reports that contained in the usual treatment at admission a drug from group N05 or N06 according to the classification (psycholeptic and psychoanaleptic drugs). There were identified 402 reports including some of these drugs (24.6% of the reports). However, if the search was based on the presence of a diagnosis of the anxious-depressive field in the sections of "diagnoses" or "personal history", only 45 (2.75%) and 192 (11.77%) were identified respetively. These data show that the physician's awareness regarding the prevalence of anxiety-depressive disorders in renal patients is low despite a high prescription of drugs to treat their symptoms.

## Discussion

In this pilot study, we have evaluated the usefulness of applying folksonomy and artificial intelligence techniques, such as NLP, for the analysis of data from the hospital discharge reports of the Nephrology Department aiming to respond merely clinical questions. The application of this technology has allowed us to significantly reduce the time expended to extract information. Only based on the usual structure of the discharge reports and their writing in natural language, it has been possible to extract relevant information that, if the tool was not available, would have required the manual review of the discharge reports and the generation of databases.

One of the lessons learned after having performed this pilot project is that the clear expression of relevant medical information in the field of nephrology (such as the classification of kidney disease) would have facilitated and accelerated the data collection. Despite a not completely uniform and structured expression of hospital discharges, often with a lack of relevant information (such as the adequate classification of the renal situation of the patients), the tool has allowed the inclusion of algorithms and heuristic rules to solve these initial difficulties.

The installation of the Bismart Folksonomy portal is a relatively quick process; however, the application of modifications and algorithms necessary for a specific project require more time to initiate the study. Searching for tags in documents is automatic, and once launched, results are obtained in less than a minute. Subsequently, the more laborious exercise of creating rules and synonyms may require 2–3 h of work to provide data analysis. However, this will also be an automatic process for all those documents that are subsequently incorporated, allowing a real-time analysis of all hospital discharges that are incorporated into the system and, therefore, allowing a real-time analysis of any issue to be explored, as well as creating alarms that would allow us to detect and select patients with certain characteristics of interest.

This tool could also be used in other healthcare settings, such as outpatient clinics or nephrology short stay hospital activity, where a significant volume of information is generated in natural language. In addition, having the possibility of crossing data from reports with laboratory results or other complementary examinations not directly included in medical reports, would exponentially increase the information extracted with the application of folksonomy. This is a relevant aspect of the analysis we have carried out of the treatments contained in the report at the time of discharge. Since the official document of treatment for a patient is the electronic prescription, it would be of great interest to be able to compare the pharmacological treatment contained in a discharge report with the data from the electronic prescription of the same patient. Although it is technically feasible to apply folksonomy to the electronic prescription, as has been done with discharge reports, in this project this process could not be carried because there was complete anonymization of the patient's affiliation.

As far as we know, there are no previous published experiences that have worked with folksonomy in the medical field. This technology allows to identified any search term without the need of having previously defined an ontology or master entity and there is no possibility of error in the search for the terms of interest. However, it exist the possibility of misclassifying search terms because they are false positives or false negatives (for example, terms like "no" or "discontinue" in front of our search terms would represent false positives). In this pilot experience, given inconsistent results, these reports have been manually reviewed and reclassified, but the tool allows the inclusion of rules that detect negative statements, thus avoiding the task of manual review. Finally, an aspect to improve is the fact that the search tool of the Bismart Folksonomy portal (*easy query section*) allows the addition of words search (use of the term "and") but currently does not allow the search for a term or another (use of the term "or"), which represents a certain limitation in obtaining information. In this protocol, this limitation regarding the term "or" was solved with the creation of "categories" (a term that groups a collection of terms). An example of this would be the ATC classification (the term ATC09 was associated with all the drugs that inhibit the renin angiotensin system).

In conclusion, the use of *big data* in the medical field, in this specific case of folksonomy and NLP, can allow significant time savings without detriment to the quality and truth of the information obtained for research purposes and quality management of the care activity being carried out.

## Key concepts

- A large amount of clinical data is generated every day, much of it being collected in the form of natural language.
- Classically, the extraction and analysis of data from medical records is being done through a manual process that requires a significant investment of time.
- The use of *big data tools*, specifically *natural language processing* (NLP), makes it possible to speed up this process.
- The application of folksonomy as an NLP tool does not require the prior creation of a master entity that collects

the search terms of interest, and this fact provides a clear advantage over other NLP tools.

- Based on certain clinical questions in the field of nephrology and by using the *Bismart Folksonomy software,* folksonomy has been applied to automatically extract and analyze data from discharge reports from a nephrology department.

## Conflict of interests

The authors declare that they have no conflict of interest.

## Thanks

This pilot project with the *Bismart Folksonomy tool* from the company Bismart, has been carried out through the financing of Laboratorios Ferrer, without having obtained any clinical data and did not guide this research project. Laia Sans, Jordi Martínez-Roldan and Julio Pascual have no professional relationship with Bismart. Ismael Vallvé and Joan Teixidó work for Bismart and Josep Manel Picas acts as a consultant for Bismart.

The authors thank Laboratorios Ferrer for their support in carrying out this study.

## REFERENCES

1. Palanisamy V, Thirunavukarasu R. Implications of big data analytics in developing healthcare frameworks – a review. J King Saud Univ - Comp Inform Sci. 2019;31:415–25.
2. Vigilante K, Escarvage S, Mc Connel M. Big data and the Intelligence Community — lessons for health care. N Engl J Med. 2019;380:1888–90.
3. Wen A, Fu S, Moon S, El Wazir M, Rosenbaum A, Kaggal VC, et al. Desiderata for delivering NLP to accelerate healthcare AI advancement and a Mayo Clinic NLP-as-a-service implementation. npj Digit Med. 2019;2:130, 17.
4. Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, et al. Artificial intelligence in healthcare: past, present and future. Stroke Vasc Neurol. 2017;2:230–43.
5. Levey A, de Jong P, Coresh J, El Nahas M, Astor B, Matsushita K, et al. The definition, classification, and prognosis of chronic kidney disease: a KDIGO Controversies Conference report. Kidney Int. 2011;80:17–28.
6. Levey AS, Stevens LA, Schmid CH, Zhang YL, Castro AF 3rd, Feldman HI, et al. A new equation to estimate glomerular filtration rate. Ann Intern Med. 2009;150:604–12.
7. Crowley M, Diamantidis C, McDuffie J, Cameron B, Stanifer J, Mock C, et al. Clinical outcomes of metformin use in populations with chronic kidney disease, congestive heart failure, or chronic liver disease: a systematic review. Ann Intern Med. 2017;166:191–200.
8. Lazarus B, Wu A, Shin JI, Sang Y, Alexander GC, Secora A, et al. Association of metformin use with risk of lactic acidosis across the range of kidney function: a community-based cohort study. JAMA Intern Med. 2018;178:903–10.
9. Kalaitzidis RG, Elisaf MS. Treatment of hypertension in chronic kidney disease. Curr Hypertens Rep. 2018;20:64.
10. Xie X, Liu Y, Perkovic V, Li X, Ninomiya T, Hou W, et al. Renin-angiotensin system inhibitors and kidney and cardiovascular outcomes in patients with CKD: a bayesian network meta-analysis of randomized clinical trials. Am J Kidney Dis. 2016;67:728–41.
11. Bhandari S, Ives N, Brettell EA, Valente M, Cockwell P, Topham PS, et al. Multicentre randomized controlled trial of angiotensin-converting enzyme inhibitor/angiotensin receptor blocker withdrawal in advanced renal disease: the STOP-ACEi trial. Nephrol Dial Transplant. 2016;31:255–61.
12. Cangini G, Rusolo D, Cappuccilli M, Donati G, La Manna G. Evolution of the concept of quality of life in the population in end stage renal disease. A systematic review of the literature. Clin Ther. 2019;170:e301–20.
13. Wang WL, Liang S, Zhu FL, Liu JQ, Wang SY, Chen XM, et al. The prevalence of depression and the association between depression and kidney function and health-related quality of life in elderly patients with chronic kidney disease: a multicenter cross-sectional study. Clin Interv Aging. 2019;14:905–13.