Received: 28 July 2021 | Revised: 24 February 2022 | Accepted: 28 March 2022

DOI: 10.1002/clt2.12144

ORIGINAL ARTICLE

Artificial intelligence processing electronic health records to identify commonalities and comorbidities cluster at Immuno Center Humanitas

Pierandrea Morandini[1] | Maria Elena Laino[1] | Giovanni Paoletti[2,3] | Alessandro Carlucci[1] | Tobia Tommasini[1] | Giovanni Angelotti[1] | Jack Pepys[2] | Giorgio Walter Canonica[2,3] | Enrico Heffler[2,3] | Victor Savevski[1] | Francesca Puggioni[2,3]

[1]Artificial Intelligence Center, IRCCS Humanitas Research Hospital, Milan, Italy

[2]Department of Biomedical Sciences, Humanitas University, Milan, Italy

[3]Personalized Medicine, Asthma and Allergy, IRCCS Humanitas Research Hospital, Milan, Italy

Correspondence

Victor Savevski, Artificial Intelligence Center, IRCCS Humanitas Research Hospital, Via Manzoni 56, Rozzano, Milan, 20089, Italy.
Email: victor.savevski@humanitas.it

Funding information

Internal project funded by IRCCS Istituto Clinico Humanita—5 × 1000

Abstract

Background: Comorbidities are common in chronic inflammatory conditions, requiring multidisciplinary treatment approach. Understanding the link between a single disease and its comorbidities is important for appropriate treatment and management. We evaluate the ability of an NLP-based process for knowledge discovery to detect information about pathologies, patients' phenotype, doctors' prescriptions and commonalities in electronic medical records, by extracting information from free narrative text written by clinicians during medical visits, resulting in the extraction of valuable information and enriching real world evidence data from a multidisciplinary setting.

Methods: We collected clinical notes from the Allergy Department of Humanitas Research Hospital written in the last 3 years and used it to look for diseases that cluster together as comorbidities associated to the main pathology of our patients, and for the extent of prescription of systemic corticosteroids, thus evaluating the ability of NLP-based tools for knowledge discovery to extract structured information from free text.

Results: We found that the 3 most frequent comorbidities to appear in our clusters were asthma, rhinitis, and urticaria, and that 991 (of 2057) patients suffered from at least one of these comorbidities. The clusters which co-occur particularly often are oral allergy syndrome and urticaria (131 patients), angioedema and urticaria (105 patients), rhinitis and asthma (227 patients). With regards to systemic corticosteroid prescription volume by our clinicians, we found it was lower when compared to the therapy the patients followed before coming to our attention, with the exception of two diseases: Chronic obstructive pulmonary disease and Angioedema.

Pierandrea Morandini and Maria Elena Laino contributed equally to this study.

Victor Savevski and Francesca Puggioni equally supervised this study.

Conclusions: This analysis seems to be valid and is confirmed by the data from the literature. This means that NLP tools could have significant role in many other research fields of medicine, as it may help identify other important, and possibly previously neglected clusters of patients with comorbidities and commonalities. Another potential benefit of this approach lies in its potential ability to foster a multidisciplinary approach, using the same drugs to treat pathologies normally treated by physicians in different branches of medicine, thus saving resources and improving the pharmacological management of patients.

KEYWORDS
allergy, artificial intelligence, asthma, clustering, natural language processing, urticaria

1 | INTRODUCTION

Comorbidity is common in autoimmune or inflammatory conditions, such as asthma,[1] chronic obstructive pulmonary disease (COPD),[2] rheumatoid arthritis,[3] psoriasis and psoriatic arthritis,[4] and inflammatory bowel disease (IBD)[5] with 30% of patients manifesting more than one condition and thus requiring a multidisciplinary approach.[6]

Assessing how and when comorbidities are associated with a major condition would provide a deeper understanding of the comorbidity itself and, at the same time, provide new insights for a better treatment strategy.

Based on this background and taking advantage of data warehouse (DWH) resources of the Humanitas Immuno Center, our aim is to evaluate the ability of NLP-based tools for knowledge discovery to detect information about pathologies in medical records collected from free text format. Medical records are written by clinical professionals in a narrative style during hospital visits. As a main outcome, we expect to use patients' data to identify the different pathologies treated in our Allergy Department, understand if there are any comorbidity associations, and extract positive feedback for the practical management of these patients.

Indeed, this would allow more precise patients' phenotyping and tailored therapies, reducing both active and passive costs related to poor control of the disease, and improving the quality of life of the patients.[7–9]

2 | MATERIALS AND METHODS

2.1 | Dataset

We retrospectively collected all the clinical notes written from January 2017 to September 2020 of patients with ongoing or terminated care process at the Allergy Department of Humanitas Research Hospital.

We included in our study only medical records from patients who gave their consent for the use of their data for research purposes.

We excluded the hospital records collected during encounters with only therapeutic purposes (i.e., visits for drug infusion), since these records do not contain relevant information for our analysis.

2.2 | Data selection

The clinical notes we processed present multiple layout structures, hence the information we collected is generally located in different paragraphs of the clinical notes.

In this regard, a normalization of the clinical notes was required in order to standardize the data for the downstream processes. An analysis of the used layouts led to the identification of the paragraphs containing the information we are interested in.

In particular, the only paragraphs we considered in our analysis were those related to the patient's *anamnesis*, in which the searched pathologies are considered as comorbidities and drugs are considered as previous therapy, the *conclusions* paragraph to extract the final diagnosis, and the *therapy* paragraph to extract the drugs prescribed by our clinicians.

This method of data analysis was selected following consultations with the allergy unit clinicians on their standardized method of reporting.

The complete list of considered comorbidities can be found in Table 1 and the list of systemic corticosteroids can be found in Table 2.

The selections of pathologies and drugs were carried out in consultation with ImmunoCenter experts and literature analysis. Finally, the list of the diseases (36) and drugs (10 active principles and 31 tradenames) was identified focusing on those treated/prescribed through the multidisciplinary approach within the Humanitas ImmunoCenter.

2.3 | Data pre-processing

The first data extraction step consisted in querying the data from DWH. We used Oracle SQL™ to gather the relevant data of patients examined at the Allergy Department. Consequently, a pre-process

TABLE 1 List of comorbidities reported in the anamnesis paragraph

List of comorbidities	
Anafilassi	Anaphylaxis
Angioedema	Angioedema
Arterite	Arteritis
Artrite psoriasica	Psoriatic arthritis
Artrite reumatoide	Rheumatoid arthritis
Asma	Asthma
Aspergillosi	Aspergillosis
Bronchiectasie	Bronchiectasis
Broncopneumopatia cronica ostruttiva (BPCO)	Chronic Obstructive Pulmonary Disease (COPD)
Churg strauss	Churg strauss
Colite indeterminata	Indeterminate colitis
Colite ulcerosa	Ulcerative colitis
Connettivite	Connectivitis
Dattilite	Dactylitis
Dermatite atopica	Atopic dermatitis
Esofagite eosinofila	Eosinophilic esophagitis
Interstiziopatia	Interstitial disease
Lupus	Lupus
Mastocitosi	Mastocytosis
Miosite	Myositis
Morbo di Crohn	Crohn's disease
Orticaria	Urticaria
Osteoporosi	Osteoporosis
Poliposi nasale	Nasal polyposis
Polmonite eosinofila	Eosinophilic pneumonia
Psoriasi	Psoriasis
Rinite	Rhinitis
Rinosinusite	Rhinosinusitis
Sacroileite	Sacroiliitis
Sclerosi sistemica	Systemic sclerosis
Sindrome orale allergica (SOA)	Oral allergic syndrome (SOA)
Sinusite	Sinusitis
Sjogren	Sjogren
Spondilite	Spondylitis
Spondiloartrite	Spondyloarthritis
Vasculiti	Vasculitis

pipeline was implemented to clean the text data from unwanted or unnecessary characters, returning a cleaned corpus ready to be processed. The pre-processing phase aimed to both normalize the characters to ASCII format, and remove all HTML special characters from the text.

2.4 | Marker extraction

For the whole of the following analysis, we used Python (ver. 3.6.9), including multiple libraries: *pandas*[10] *kmodes*[11] *regex* (*re*)[12] and *scikit-learn*[13] among others.

TABLE 2 Patients treated with a specific systemic corticosteroid drug

Corticosteroid drugs	Prescription	Anamnesis
Cortisone	2	31
Prednisone	131	152
Prednisolone	0	0
Methylprednisolone	9	30
Beclomethasone	5	22
Triamcinolone	2	15
Budesonide	13	20
Betamethasone	66	59
Dexamethasone	0	2
Hydrocortisone	1	13
Not specified corticosteroid	0	203

Note: In column "Prescription", we included the number of patients prescribed with a specific or not specified corticosteroid in our center. In the column "Anamnesis", we included the number of patients previously treated with a specific or not specified corticosteroid.

The marker extraction was performed entirely with Regular Expressions (RegEx), described in detail in Supporting information S1. The marker extraction process helped define reliable patterns used to detect the presence of the considered pathologies and therapies in the retrieved text. From this point onwards, we refer to those pathologies as ***entities***.

2.5 | Evaluation of marker extraction process

We sampled a subset of sentences to manually evaluate the goodness of the extracted markers. For each pathology, we validated the extracted marker of 20 sentences. The first 10 sentences were presumed to express the presence of the pathology, 6 were supposed to express negations of pathology, and 4 were control sentences in which the pathology was not detected. We evaluated a binary outcome depending on the presence or absence of the disease. This allowed us to evaluate our algorithm with the indexes of Recall, Precision and F1 score.

2.6 | Clustering

The dataset underwent a clustering process to explore optimal grouping arrangements of the gathered entities. The aim is to find the main families of clinical conditions considering, for each patient, both comorbidities and diagnosis.

All markers related to each hospital encounter were aggregated, resulting in a list of all the different autoimmune pathologies present along the care process for every patient.

Since the used data are composed only by binary flags, the clustering was performed with the k-modes algorithm.[14] This is a variation of the well-known k-means algorithm[15] specifically fitted to work with binary data.

To define the optimal number of clusters traditional methods, base the clustering evaluation on metrics regarding the spatial distances between observations and their cluster centroids. Since it is not possible to define a spatial distance between categorical data, we relied on the cost function defined by the k-modes algorithm, to find the optimal number of clusters.

The cost function is defined as:

$$P(W,\ Q)=\sum_{l=1}^{k}\sum_{i=1}^{n}\sum_{j=1}^{m}w_{i,l}\delta\left(x_{i,j},q_{l,j}\right)\ (1)$$

where

$$\delta\left(x_j,\ y_j\right)=\begin{cases}0\ \left(x_j=y_j\right)\\1\ \left(x_j\neq y_j\right)\end{cases}$$

Subject to

$$\sum_{l=1}^{k}w_{i,l}=1,\ 1\leq i\leq n$$

$$w_{i,l}\in\{0,1\},\ 1\leq i\leq n,\ 1\leq l\leq k$$

These equations define the cost function as the sum of dissimilarities between a data point X, composed of *m* categorical attributes and *n* observations, and a matrix $Q=[q_1,q_2,\cdots,q_k]$ defining the modes of k clusters. These dissimilarities are weighted by the coefficients of a matrix W.

As suggested by Huang et al.[14] to solve the Equation (1) an iterative process is needed. In particular, the values of W and Q are found by following these steps:

1. Fix $\widehat{Q}=Q_t$ and solve $P\left(W,\widehat{Q}\right)$ to obtain W_t
2. Fix $\widehat{W}=W_t$ and solve $P\left(\widehat{W},Q\right)$ to obtain Q_{t+1}

3 | RESULTS

3.1 | Marker extraction evaluation performance

To validate the performances of regular expression a total of 720 sentences were manually annotated. The values for recall, precision and F1 score were respectively 0.97, 0.84, and 0.90.

3.2 | Hospital encounters analysis

In our Allergy Department, 3162 patients had 7827 visits from January 2017 to September 2020. Of these, we enrolled 2057 patients [887 (43.1%) men with the median age of 48.07 years (+20.59 s.d.)] after screening for the type of hospital visit and excluding

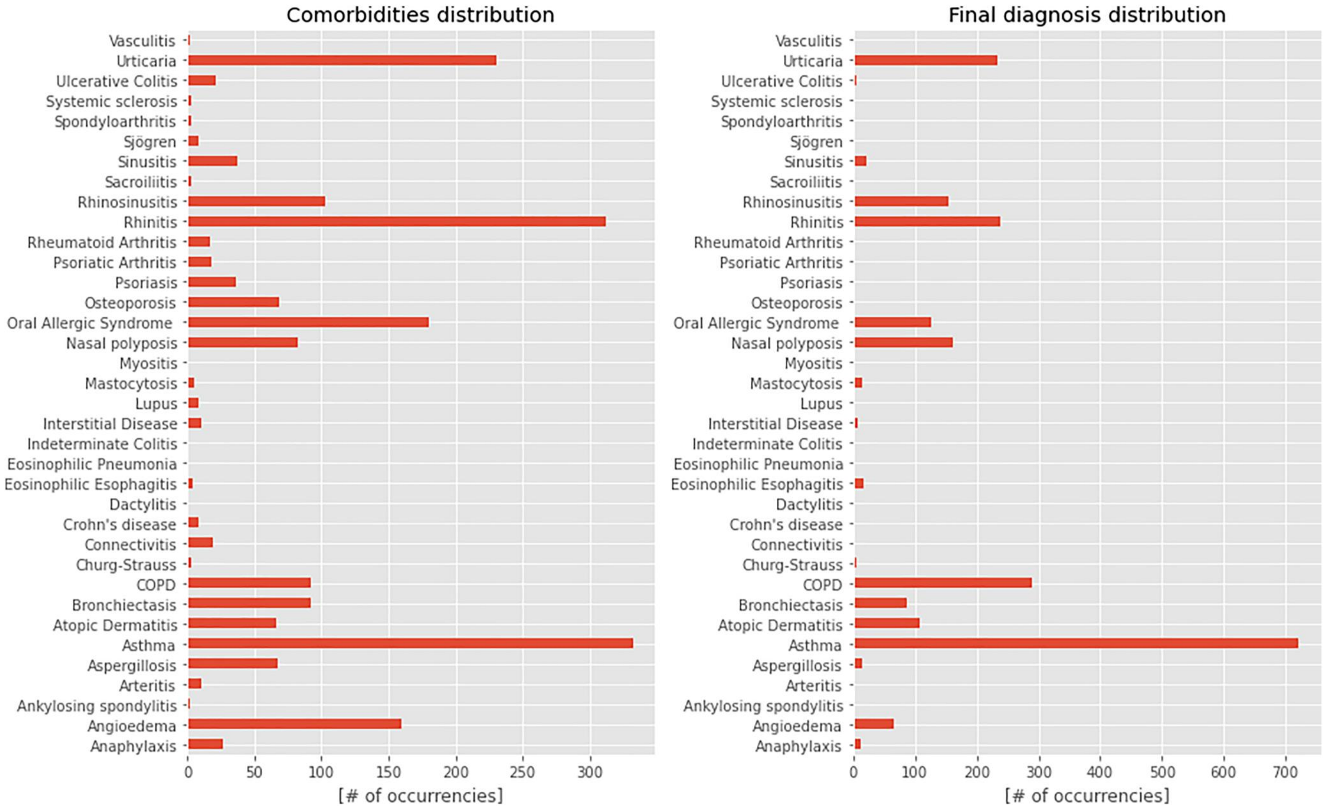

FIGURE 1 Distribution of diseases after the marker extraction considering a single hospital encounter. The first plot shows all the pathologies contained in the paragraph "anamnesis"—which we considered as comorbidities—and the second plot shows all the pathologies contained in the paragraph "conclusions"—which we considered as final diagnosis after the hospital encounter. The most commonly occurring comorbidities were asthma (detected in 332 hospital encounters, 10.2% of considered visits), rhinitis (312, 9.6%), and urticaria (230, 7.1%). Asthma is the most frequent pathology diagnosed (719 hospital encounters, 22.2%), followed by COPD (290, 8.9%) and rhinitis (239, 7.4%)

patients who came only for therapeutic purposes. In this way, the number of visits was reduced to 3226 (average of 1.57 visits per patient).

Figure 1 shows the distribution of diseases after the marker extraction step, before the aggregation to the patient level. In particular, it indicates how many clinical notes reported the pathologies of interest during a single hospital encounter.

As previously mentioned, we considered the diseases reported in the paragraphs "anamnesis" as comorbidities (left side of the figure) and those cited in the paragraph "conclusions" as diagnosis (right side of the figure).

In our series, the three most frequent comorbidities are *asthma* (detected in 332 hospital encounters, 10.2% of considered visits), *rhinitis* (312, 9.6%) and *urticaria* (230, 7.1%).

Moreover, *asthma* is the most frequent pathology diagnosed (719 hospital encounters, 22.2%), followed by *COPD* (90, 8.9%) and *rhinitis* (239, 7.4%).

After data aggregation and selection steps described in the section "Clustering", we analyzed how many comorbidities were reported in clinical notes for each patient, considering each hospital encounter, as shown in Figure 2. We found that 991 out of 2057 patients suffered from at least one of the considered comorbidities, for a total of 1465 over 3226 hospital encounters, while 1066 patients were considered as not presenting the searched pathologies as we did not find the relevant comorbidities in their anamnesis paragraph.

Furthermore, we investigated differences or similarities between the two categories of patients (with or without comorbidities reported in the *anamnesis* paragraph). For this purpose, we compared the distribution of pathologies found in the *conclusion* paragraph between the two categories of patients, as shown in Figure 3.

In particular, we noticed differences in the volume of found markers for Nasal *Polyps*, *Rhinosinusitis*, *Bronchiectasis* and *Asthma*. The latter remains the most diagnosed pathology in both categories but clinical notes which did not contain the searched comorbidities closely follows.

In Table 3, we show the *p*-values for the correlation between the presence of at least one comorbidity and the diagnosis of asthma, rhinosinusitis, polyps, or bronchiectasis. The correlation is statistically significant with the diagnosis of all of them. This could be explained by the fact that patients with these four diseases frequently have a comorbidity as widely demonstrated in the literature.[16]

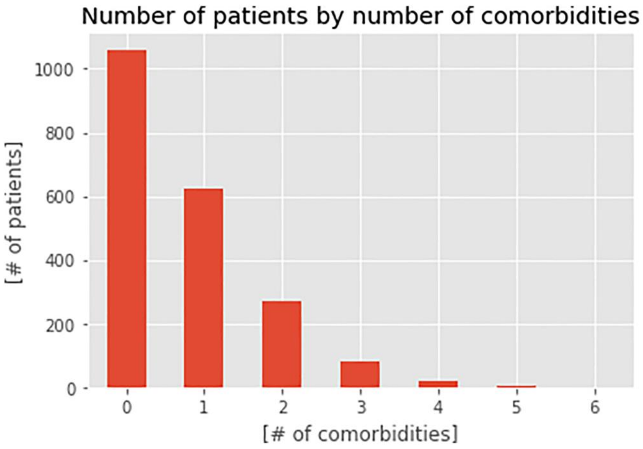

FIGURE 2 Volume of patients with comorbidities. This bar plot shows the volumes of patients as a function of the number of comorbidities found in the paragraph "anamnesis." In particular, in our series 1066 did not show any comorbidities, 629 showed 1 comorbidity, 260 showed 2 comorbidities, 77 showed 3 comorbidities, 20 showed 4 comorbidities, 4 showed 5 comorbidities, 1 showed 6 comorbidities

4 | PATIENTS ANALYSIS

From this point onwards, the analysis was conducted with focus on the whole patients' cure process, instead of considering single hospital encounters.

In order to define the optimal number of clusters to use in our study, we analyzed the elbow plot (Figure 4), which suggests that $N = 6$ is the best option for our series.

To confirm the goodness of clustering with $N = 6$, we analyzed the silhouette for each observation included in the clustering process.

As can be seen in Figure 5, for $N = 6$ all the observations have a silhouette relatively close to 1. We then characterized the different clusters in terms of comorbidities presence and numerosity in order to understand by which comorbidities they are defined.

Figure 6 shows, for each cluster, the importance that the specific comorbidities have in characterizing the clusters (ratio between the cluster population and the number of those patients who experienced a certain comorbidity).

1. out of 6 clusters (clusters 1–5) showed a strong recurrence of a specific pathology.

Furthermore, in the above-mentioned clusters, at least one secondary comorbidity seems to be correlated to the main one.

Specifically, *rhinitis* and *asthma* co-occur in cluster 1 (388 patients), *angioedema* and *urticaria* in cluster 2 (258 patients), *asthma, rhinosinusitis and polyps* in cluster 3 (122 patients), *chronic obstructive pulmonary disease (COPD) and bronchiectasis in cluster 4 (187 patients), asthma and oral allergy syndrome in cluster 5 (268 patients).*

As can be seen, cluster 0 is not as defined as the others: no comorbidity is present in the majority of the population.

In cluster 0, the most common pathologies are *oral allergy syndrome, angioedema and atopic dermatitis.*

Furthermore, we analyzed how corticosteroid therapy correlated with assigned clusters. We collected all the information about prescription of systemic corticosteroids both in the anamnesis paragraph—which represented the therapy that patients followed before coming to our centre- and in the conclusive therapy paragraph—which included the drugs prescribed by our clinicians.

Analyzing the volumes of drugs prescribed in Humanitas and taken by patients before visits to our center, two differences are noticeable. The first one is that the volumes of drugs prescribed by our clinicians is lower than the volume reported in patients' anamnesis prior to treatment in our ImmunoCenter, as shown in Table 2. The reduction of prescriptions of these drugs is, indeed, an advantage of our center. Secondly, we found a difference between the patients treated with corticosteroids prior to visiting our clinicians and the patients to whom corticosteroids were prescribed by our clinicians—as shown in Table 4.

Moreover, after analysing the correlation between drugs prescribed in our centre and patients' clusters, we found a significant correlation between prednisone and cluster 2 and betamethasone and clusters 2 and 4. As shown in Table 5, prednisone and betamethasone were the drugs that there was less of a reduction, or in the case of betamethasone, an increase in the prescription by our clinicians.

On the other side, there is no correlation between drugs found in the anamnesis and patients' clusters, as shown in Table 5.

5 | DISCUSSION

We built a framework to extract structured information from free text through NLP, which can eventually be transposed to other types of clinical notes to extract valuable information to enrich real world evidence data.

After establishing the patients' inclusion criteria and the pathologies of interest for the study, we queried the data from the hospital's DWH. Subsequently, a pre-processing pipeline was implemented in order to clean the text data from unwanted or unnecessary characters. Finally, the clinical notes underwent the marker extraction step, which consisted in the detection of the pathologies of interest (entities) in the analyzed clinical notes. The marker extraction process was carried out with regular expressions. This approach proved to be an efficient tool for the entity detection task on medical texts in an initially unsupervised fashion.

As stated by Wang,[17] in medicine the information extraction tasks are mainly left to techniques that make use of empirical rules (as per regex) to obtain the requested results. One of the reasons justifying this preference is that rule-based information extraction can incorporate domain knowledge from knowledge bases or experts, which is essential for clinical applications. In our study, we chose to use Regular Expressions instead of building an Entity Recognition model, which is a very time-consuming option,[18] mostly because of

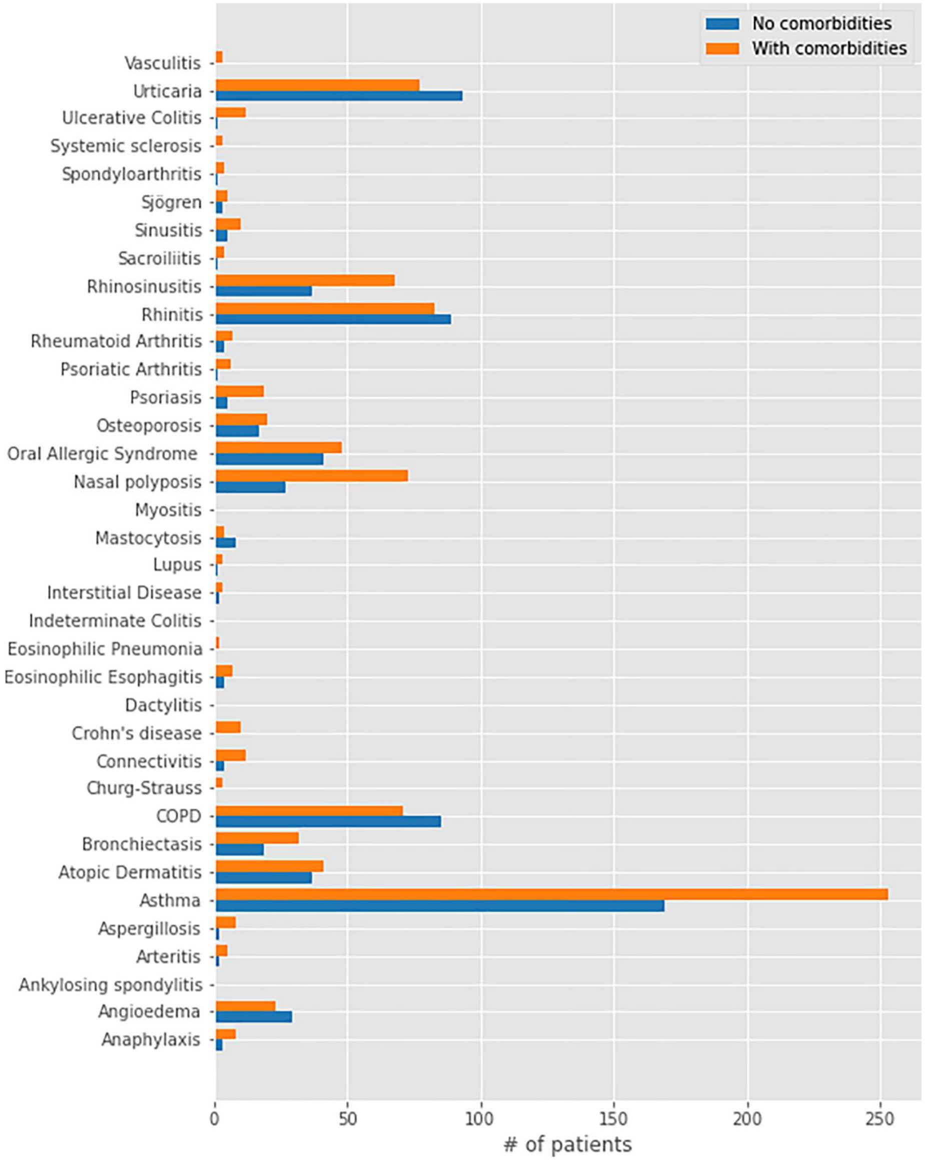

FIGURE 3 Distribution of the final diagnosis between patients with at least one comorbidity and patients with no comorbidity. We compared the raw counting of the distribution in the two groups of patients since they are composed of a very similar number of observations (1066 without comorbidities vs. 991 with comorbidities). Between the two groups, the most important differences in volume of found markers between the two groups were for nasal polyps (73 in the first group and 27 in the second group), rhinosinusitis (68 vs. 37), bronchiectasis (32 vs. 19), and asthma (253 vs. 169)

the high volume of annotated data required. Moreover, the task is suitable for RegEx approach because searched expressions of the considered pathologies are well defined as specific nouns due to the specificity of the medical terminology and presence of abbreviations. For these reasons, the action that may produce an alteration of the expressions can be caused only by misspelling or typing errors which are eventualities that could be handled by RegEx.

To test this hypothesis in our series, we evaluated our performances on a subset of sentences and obtained very good results. The high recall, in particular, can be explained by the method we used to validate. Sampling the sentences to annotate stratifying on the extracted marker is crucial to get a balanced set, but might introduce a bias. A more interesting parameter is the precision, which is still good, but not as good as the recall: this is caused by missed negations that precede or follow the mentioned pathology. This means that generally, it is possible to extract entities from clinical notes using RegEx being aware that it is crucial to focus also on the negations detection. With these data, it is possible to say that the marker extraction algorithm has acceptable performance, although a more in-depth evaluation is required to better evaluate the performances of our algorithm.

Of note in our results is the fact that through regular expressions we retrieved epidemiologic data about our Allergy Department patients' phenotype.

TABLE 3 Statistical correlation between the indicated disease and the presence of at least one comorbidity, expressed as *p*-value calculated using the Pearson's χ^2 test

Disease	*p*-Value
Asthma	<0.001
Rhinosinusitis	<0.001
Polyps	<0.001
Bronchiectasis	0.049

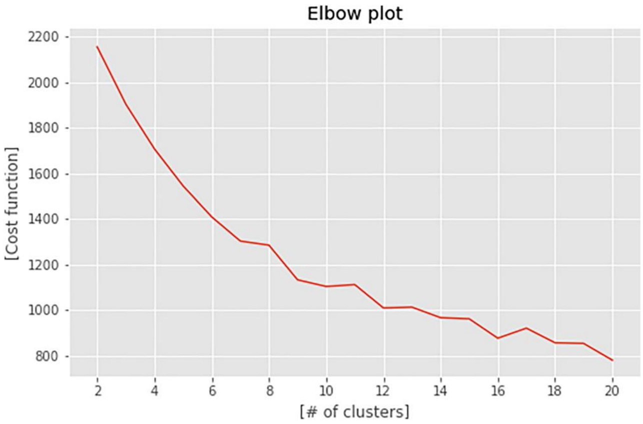

FIGURE 4 Clustering cost measure curve. Using the elbow method, $N = 6$ was the optimal number of clusters chosen in this setup

We found that asthma is the pathology most frequently diagnosed. This data is due to different factors, as asthma affects up to 18% of the population[19] and Humanitas' Allergy Unit is a world-renowned center of excellence for asthma management and has performed several international clinical trials on asthma and comorbidities this is to be expected.

Similarly, Chronic rhinosinusitis with nasal polyps (CRSwNP) affects 5%–12% of the general population[20] and is the second most frequent pathology managed by the Humanitas Allergy Unit, which is unsurprising, as it is often associated with severe asthma. The disease management in a multidisciplinary rhinology clinic by allergists and ENTs is another explanation for the frequency in which we encounter it.

Furthermore, as Figure 1 shows, in the first plot there is a more homogeneous distribution of the pathologies we took into consideration compared to the second plot. We expected this kind of distribution since we assumed that the anamnesis paragraph contained the information about comorbidities and the conclusions paragraph contained those about the diagnosis. This means that the second plot shows all the diseases treated in our Allergy Department, while in the first plot we can find diseases not directly treated in the Allergy Department but rather generally treated in our ImmunoCenter.

Retrieving information about the most frequent comorbidities in our series helped us get a more complete picture of our patients, which is very important in a multidisciplinary context.

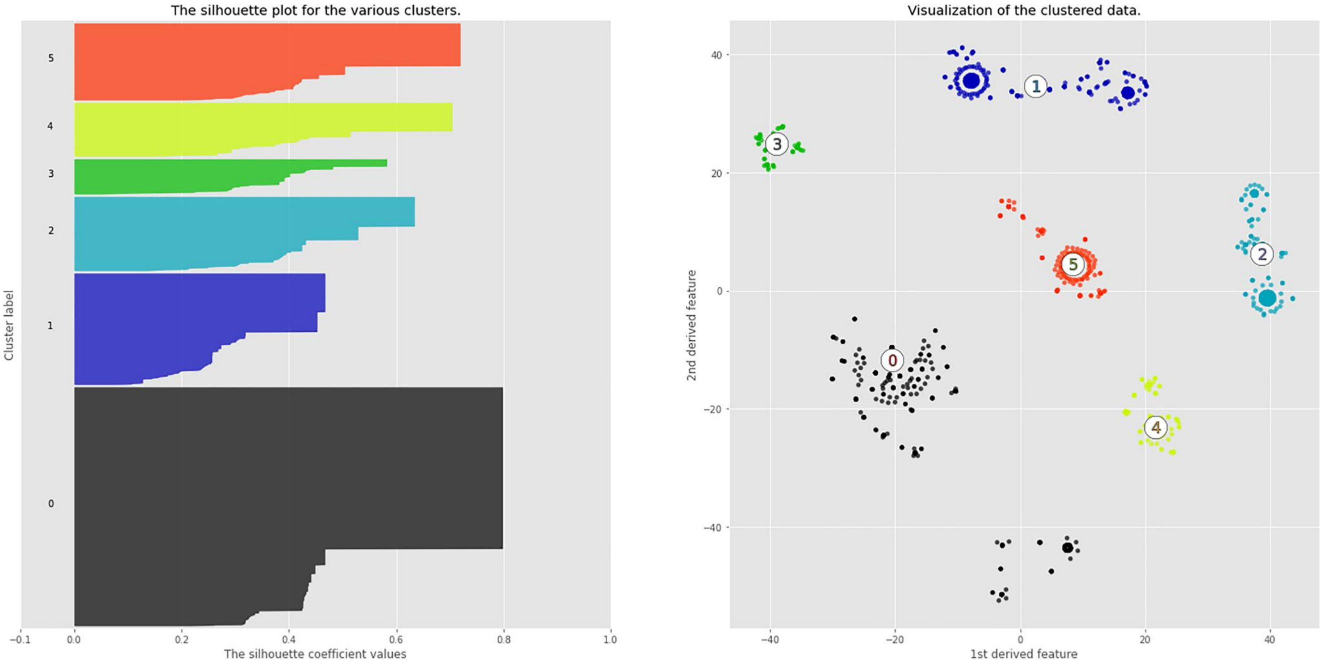

FIGURE 5 Patients' clusters in a fictitious 2D space. This figure represents the plots of the clustered patients in a fictitious 2D space (right side) and a plot in which the silhouette measure is calculated for each patient and plotted as part of the belonging cluster's silhouette. The silhouette measure indicates the goodness of clustering for each observation and its value ranges in the interval [−1, +1]. The larger the measure, the better the observation clustering. Based on this measure, all the different clusters have good ratings

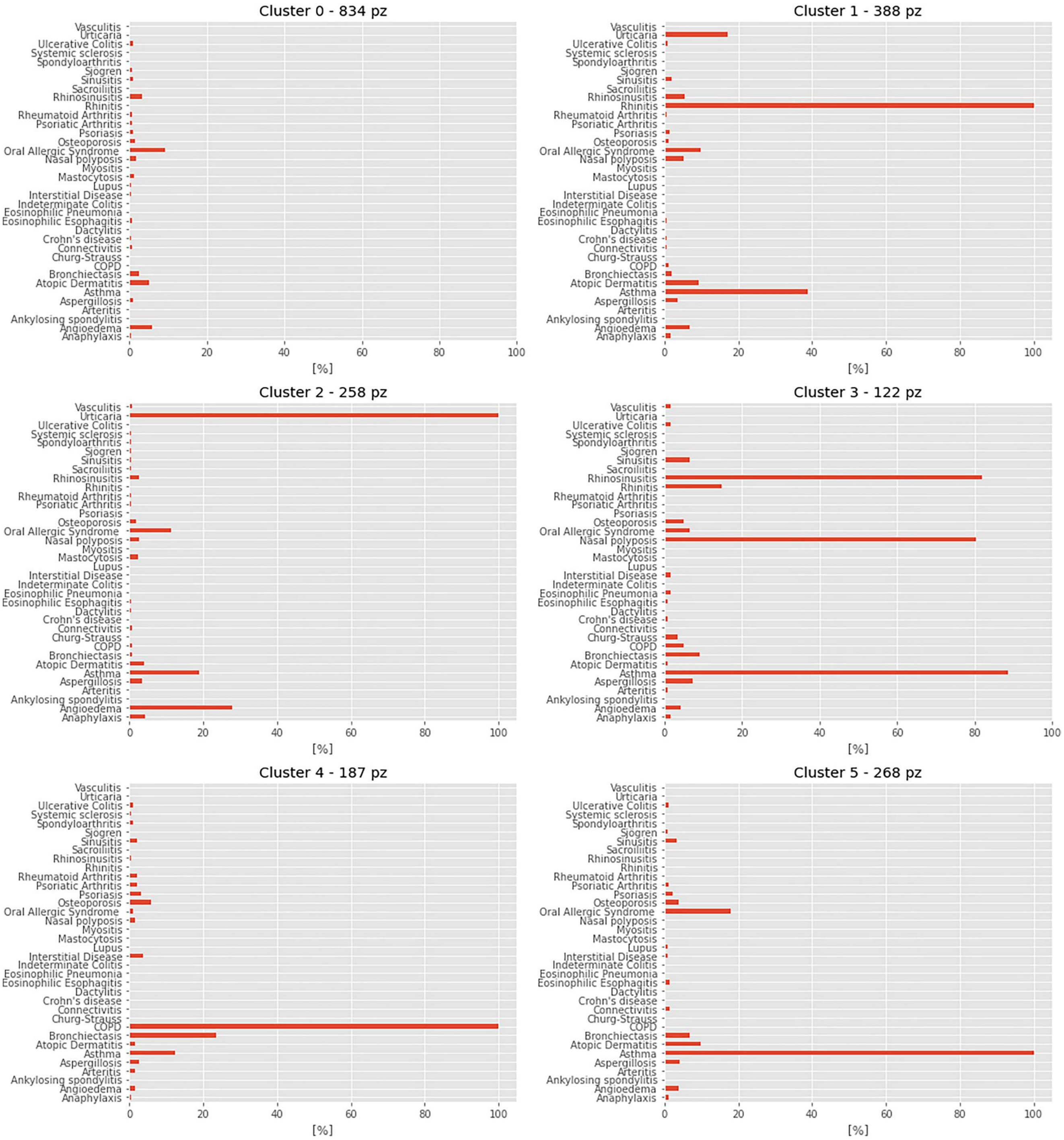

FIGURE 6 Cluster characterization by comorbidities. Each plot shows a specific cluster of patients in which we divided our series. The clusters have a different numerosity, as can be seen in the title of each plot. Furthermore, in each plot the percentage of patients with a specific comorbidity is represented

Moreover, we found that four pathologies—asthma, rhinosinusitis, polyps, and bronchiectasis (Figure 3)—occur more frequently within the group of patients with at least one comorbidity, as a result of what is shown in Figure 3 and Table 3. This could suggest that the above-mentioned diseases are more frequently associated with other diseases. This can be explained by the fact that asthma and rhinosinusitis with or without polyps can be driven by a common molecular mechanism, namely type 2 inflammation. This inflammatory response

TABLE 4 Measures of overlap between newly prescribed and previous use of corticosteroids expressed as the number of patients who were prescribed corticosteroid along their care process

Corticosteroid drug	Before only	After only	Both
Cortisone	31	2	0
Prednisone	145	124	7
Prednisolone	0	0	0
Methylprednisolone	30	9	0
Beclomethasone	22	5	0
Triamcinolone	15	2	0
Budesonide	20	13	0
Betamethasone	58	65	1
Dexamethasone	2	0	0
Hydrocortisone	13	1	0
Not specified corticosteroid	189	215	14

Note: The column "Before only" contains the number of patients who were reported to use corticosteroids in the anamnesis and not in the prescriptions, the "After only" column contains the number of patients who were prescribed corticosteroids at our center, and in "Both" are the numbers of patients who were reported to use corticosteroids both before and after visiting the ImmunoCenter. In the last row "Not specified corticosteroid" we considered the previous use of a not specified corticosteroid (which means that there was no mention of the commercial name or of the active substance) and all the prescriptions made by our center.

is emerging as a unifying feature of classically defined allergic diseases, such as asthma, and a range of other inflammatory diseases, such as rhinosinusitis[21]

For the other diseases, we did not find any substantial differences in the distribution of the final *diagnosis* between the group of patients with at least one comorbidity and the group with no comorbidities (Figure 3).

Interestingly, we found associations between different comorbidities, as shown in Figure 6. Specifically, in our clusters we found a co-occurrence of:

- *Rhinitis* and *Asthma* (cluster 1),
- *Angioedema* and *Urticaria* (cluster 2)
- *Asthma*, *Rhinosinusitis*, and *Polyps* (cluster 3)
- *COPD* and *Bronchiectasis* (cluster 4)
- *Asthma* and *Oral Allergy Syndrome* (cluster 5)

When analysing these associations from a medical and pathophysiological point of view it is unsurprising to find them in the same patients, since they have the same endotype.

- Allergic rhinitis and asthma (cluster 1) are common diseases frequently occurring together. This association is known as "united airway disease." Epidemiological studies have shown that the majority of patients with asthma have concomitant rhinitis and the presence of rhinitis is an increased risk factor for development of asthma[22–26]
- The underlying mechanism of the second cluster is mast cell degranulation, they are the primary effectors in urticaria and in many cases of angioedema,[27] resulting in skin-limited manifestations in urticaria while affecting the deeper layers in angioedema. It is therefore unsurprising that approximately 40% of patients with urticaria experience angioedema (cluster 2)[28]
- Patients with chronic rhinosinusitis with nasal polyps (CRSwNP) characterized by a type 2 immune inflammation often have severe and recurrent symptoms. Lower airway conditions such as asthma are common comorbidities and share similar pathophysiology (Cluster 3). CRSwNP with asthma is characterized by tissue eosinophilia and high local IgE levels. These conditions are correlated with more severe sinonasal symptoms and worse quality of life and clinical outcomes control[29]
- The prevalence of bronchiectasis in patients with COPD is high (cluster 4), especially in advanced stages. Some of the etiological factors for bronchiectasis are also present in patients with COPD and may be responsible for its development. Similarly, presence of a chronic bronchitis phenotype determines recurrent infective exacerbations, which perpetuate chronic inflammation, and tissue destruction[30]
- *Oral allergy syndrome* is a hypersensitivity reaction to plant-based foods, manifesting most commonly with pruritus of the lips, tongue, and mouth. Unlike simple food allergy, this disease requires prior sensitization to a cross-reacting inhalant allergen rather than direct sensitization to a specific food protein. However, a proportion of patients with oral allergic syndrome sensitized to certain pollens may have asthma (cluster 5) as an additional co-morbidity[31]

One of the most interesting aspects, which we shall investigate in future research, is the correlation between pathology, treatment, clinical personal response to therapy and modification of the therapeutic approach in our multidisciplinary ImmunoCenter compared to what happens in a simple allergy unit. Results show that the corticosteroid prescription volume from our clinicians is lower compared to the therapy that patients followed prior to coming to our attention, except for two diseases: COPD and Angioedema.

There was a significant correlation between the prescription of prednisone by our clinicians and cluster 2 and betamethasone and clusters 2 and 4 (see Table 5). The explanation is that these drugs are recommended through an action plan as rescue medication in case of the appearance of severe angioedema[32] or during severe exacerbation of COPD.[33]

5.1 | Limitations of the study

A limitation of our study might be that most physicians have the tendency to focus on the pathologies of interest of their department.

TABLE 5 *p*-values indicating statistical correlation between patients' clusters (rows 0–5) and systemic corticosteroid

	Patients' clusters											
	0		1		2		3		4		5	
Corticosteroid drugs	**Prescription**	**Anamnesis**	**Prescription**	**Anamnesis**	**Prescription**	**Anamnesis**	**Prescription**	**Anamnesis**	**Prescription**	**Anamnesis**	**Prescription**	**Anamnesis**
Cortisone	0.654	0.980	0.824	0.278	0.595	0.832	0.253	0.612	0.434	0.841	0.615	0.432
Prednisone	0.007	0.479	0.229	0.542	0.000	0.157	0.071	0.863	0.898	0.432	0.135	0.500
Prednisolone	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Methylprednisolone	0.208	0.804	0.306	0.586	0.708	0.335	0.962	0.180	0.712	0.884	0.745	0.823
Beclomethasone	0.631	0.855	0.612	0.722	0.864	0.867	0.700	0.859	0.000	1.000	0.840	0.687
Triamcinolone	0.654	0.759	0.824	0.827	0.595	0.765	0.253	0.669	0.434	0.902	0.615	0.050
Budesonide	0.662	0.858	0.973	0.321	0.913	0.501	0.750	0.514	0.509	0.303	0.135	0.944
Betamethasone	0.850	0.133	0.987	0.900	0.002	0.247	0.454	1.000	0.017	0.950	0.435	0.391
Dexamethasone	NaN	0.654	NaN	0.042321	NaN	0.594596	NaN	0.253342	NaN	0.433624	NaN	0.614832
Hydrocortisone	0.847	0.897	0.426	0.165	0.258	0.465	0.062	0.750	0.155	0.509	0.272	0.505
Not specified corticosteroid	NaN	0.549	NaN	0.405	NaN	0.450	NaN	0.742	NaN	0.667	NaN	0.959

Note: Under the semi-columns "Prescription", we included the *p*-values related to the drugs prescribed in our center and each cluster; under the semi-columns "Anamnesis", we included the *p*-values related to the drugs reported in the anamnesis paragraph (which were previously prescribed) and each cluster.

Thus, even assuming a correct extraction of the markers, we cannot exclude the omission of information relevant to the global health status of the patient.

Another limitation of the study can be the fact that we selected the pathologies of interest (comorbidities and diagnosis) before the marker extraction step. Therefore, the data we obtained may overlook useful information on the global health of a patient.

Furthermore, since we started from the analysis of free text, bias related to errors in sentence formatting (i.e., lack of punctuation) or spelling errors which may have influenced the marker extraction process cannot be excluded, even if the use of regular expressions aims at limiting this occurrence.

6 | CONCLUSIONS

Regular expressions were proven as an effective tool for entity recognition to extract medical information from free text data and to retrieve epidemiological data in our ImmunoCenter and Allergy Department.

This analysis seems to be valid and is confirmed by data from the literature. This could have significant implications for many other clusters of patients in other fields of medicine, as it may help identify other important, and possibly previously neglected clusters, but above all to be able to identify new unknown clusters of patients affected by immune system's diseases.

Another potential benefit of this approach lies in its potential ability to save resources and improve pharmacological management of patients by using the same drugs[34–38] to treat pathologies normally treated by physicians in different branches of medicine.

AI-based methods of processing electronic medical records can contribute, as we have shown, to the creation of a new patient journey based on real word evidence Data Driven approach.

ACKNOWLEDGMENTS

This work was supported by IRCCS Istituto Clinico Humanitas – 5 × 1000.

CONFLICT OF INTEREST

Francesca Puggioni received reimbursements for lectures, presentations, speakers bureaus, manuscript writing or educational events from AstraZeneca, Mundipharma, Menarini, Almirall, Chiesi, Valeas, Malesci Guidotti, Boehringer Ingelheim, Sanofi, GSK, Novartis, Stallergenes-Greer; for Consulting fees from Sanofi, Novartis, Stallergenes-Greer. Giovanni Paoletti received reimbursements for lectures, presentations, speakers bureaus, manuscript writing or educational events from Lusopharma and Novartis. Enrico Heffler received reimbursements for lectures, presentations, speakers bureaus, manuscript writing or educational events from AstraZeneca, Sanofi, GSK, Novartis, Circassia, Nestlè Purina, Stallergenes-Greer; for Consulting fees from AstraZeneca, Sanofi, GSK, Novartis, Circassia, Nestlè Purina, Stallergenes-Greer. Giorgio Walter Canonica received reimbursements for lectures, presentations, speakers bureaus, manuscript writing or educational events from AstraZeneca, Sanofi, GSK, Novartis, Chiesi Farmaceutici, Hal Allergy, Menarini, Stallergenes-Greer; for Consulting fees from AstraZeneca, Sanofi, GSK, Novartis, Chiesi Farmaceutici, Hal Allergy, Menarini, Stallergenes-Greer; for Participation on a Data Safety Monitoring Board or Advisory Board from AstraZeneca, Sanofi, GSK, Novartis, Chiesi Farmaceutici, Hal Allergy, Menarini, Stallergenes-Greer. The other authors declare that they have no conflict of interest to disclose regarding the publication of this manuscript.

ORCID

Maria Elena Laino https://orcid.org/0000-0001-7964-8798
Giovanni Paoletti https://orcid.org/0000-0003-3953-9225

REFERENCES

1. Lenferink A, van der Palen J, van der Valk PDLPM, et al. Exacerbation action plans for patients with COPD and comorbidities: a randomised controlled trial. *Eur Respir J*. 2019;54(5):1802134. https://doi.org/10.1183/13993003.02134-2018
2. Wang E, Wechsler ME, Tran TN, et al. Characterization of severe asthma worldwide: data from the International Severe Asthma Registry. *Chest*. 2020;157(4):790-804. https://doi.org/10.1016/j.chest.2019.10.053
3. Nikiphorou E, Nurmohamed MT, Szekanecz Z. Editorial: comorbidity burden in rheumatic diseases. *Front Med*. 2018;5:197. https://doi.org/10.3389/fmed.2018.00197
4. Kaushik SB, Lebwohl MG. Psoriasis: which therapy for which patient: Psoriasis comorbidities and preferred systemic agents. *J Am Acad Dermatol*. 2019;80(1):27-40. https://doi.org/10.1016/j.jaad.2018.06.057
5. Taleban S. Challenges in the diagnosis and management of inflammatory bowel disease in the elderly. *Curr Treat Options Gastroenterol*. 2015;13(3):275-286. https://doi.org/10.1007/s11938-015-0059-6
6. Deshmukh F, Vasudevan A, Mengalie E. Association between irritable bowel syndrome and asthma: a meta-analysis and systematic review. *Ann Gastroenterol*. 2019;32(6):570-577. https://doi.org/10.20524/aog.2019.0426
7. Bayliss EA, Bayliss MS, Ware JE, Jr, Steiner JF. Predicting declines in physical function in persons with multiple chronic medical conditions: what we can learn from the medical problem list. *Health Qual Life Outcome*. 2004;2:47. https://doi.org/10.1186/1477-7525-2-47
8. Makovski TT, Schmitz S, Zeegers MP, Stranges S, van den Akker M. Multimorbidity and quality of life: systematic literature review and meta-analysis. *Ageing Res Rev*. 2019;53:100903. https://doi.org/10.1016/j.arr.2019.04.005
9. Wang L, Si L, Cocker F, Palmer AJ, Sanderson K. A systematic review of cost-of-illness studies of multimorbidity. *Appl Health Econ Health Pol*. 2018;16(1):15-29. https://doi.org/10.1007/s40258-017-0346-6
10. McKinney W. Data Structures for Statistical Computing in Python. In: Proceedings of the 9th Python in Science Conference. Proceedings of the Python in Science Conference. SciPy; 2010:56-61.
11. Huang Z. Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data Mining Knowl Discov*. 1998;2:283-304.
12. Van Rossum G. The Python Library Reference, Release 3.8.2. Python Software Foundation; 2020.
13. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine Learning in Python. *J Mach Learn Res*. 2011;12(85):2825-2830.
14. Huang Z. Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data Mining Knowl Discov*. 1998;2:283-304.

15. Lloyd S. Least squares quantization in PCM. *IEEE Trans Inform Theory*. 1982;28(2):129-137.
16. Heffler E, Blasi F, Latorre M, et al. The severe asthma Network in Italy: findings and perspectives. *J Allergy Clin Immunol Pract*. 2019; 7(5):1462-1468. https://doi.org/10.1016/j.jaip.2018.10.016
17. Wang Y, Wang L, Rastegar-Mojarad M, et al. Clinical information extraction applications: a literature review. *J Biomed Inf*. 2018; 77:34-49. https://doi.org/10.1016/j.jbi.2017.11.011
18. Perera N, Dehmer M, Emmert-Streib F. Named entity recognition and relation detection for biomedical information extraction. *Front Cell Dev Biol*. 2020;8:673. https://doi.org/10.3389/fcell.2020.00673
19. Global Initiative for Asthma. *Global Strategy for Asthma Management and Prevention*; 2020. https://ginasthma.org/wp-content/uploads/2020/04/GINA-2020-full-report_-final-_wms.pdf
20. Fokkens W, Desrosiers M, Harvey R, et al. EPOS2020: development strategy and goals for the latest European Position Paper on Rhinosinusitis. *Rhinology*. 2019;57(3):162-169. https://doi.org/10.4193/Rhin17.253
21. Gandhi NA, Bennett BL, Graham NM, Pirozzi G, Stahl N, Yancopoulos GD. Targeting key proximal drivers of type 2 inflammation in disease. *Nat Rev Drug Discov*. 2016;15(1):35-50. https://doi.org/10.1038/nrd4624
22. Khan DA. Allergic rhinitis and asthma: epidemiology and common pathophysiology. *Allergy Asthma Proc*. 2014;35(5):357-361. https://doi.org/10.2500/aap.2014.35.3794
23. Brożek JL, Bousquet J, Agache I, et al. Allergic rhinitis and its impact on asthma (ARIA) guidelines-2016 revision. *J Allergy Clin Immunol*. 2017;140(4):950-958. https://doi.org/10.1016/j.jaci.2017.03.050
24. Leynaert B, Neukirch F, Demoly P, Bousquet J. Epidemiologic evidence for asthma and rhinitis comorbidity. *J Allergy Clin Immunol*. 2000;106(5 Suppl l):S201-S205. https://doi.org/10.1067/mai.2000.110151
25. Bergeron C, Hamid Q. Relationship between asthma and rhinitis: epidemiologic, pathophysiologic, and therapeutic aspects. *Allergy Asthma Clin Immunol*. 2005;1(2):81-87. https://doi.org/10.1186/1710-1492-1-2-81
26. Heffler E, Brussino L, Del Giacco S, et al. New drugs in early-stage clinical trials for allergic rhinitis. *Expet Opin Invest Drugs*. 2019; 28(3):267-273. https://doi.org/10.1080/13543784.2019.1571581
27. Amar SM, Dreskin SC. Urticaria. *Prim Care*. 2008;35(1):141-157. https://doi.org/10.1016/j.pop.2007.09.009
28. Kanani A, Betschel SD, Warrington R. Urticaria and angioedema. *Allergy Asthma Clin Immunol*. 201814(Suppl 2):59. https://doi.org/10.1186/s13223-018-0288-z
29. Laidlaw TM, Mullol J, Woessner KM, Amin N, Mannent LP. Chronic rhinosinusitis with nasal polyps and asthma. *J Allergy Clin Immunol Pract*. 2021;9(3):1133-1141. https://doi.org/10.1016/j.jaip.2020.09.063
30. Martinez-Garcia MA, Miravitlles M. Bronchiectasis in COPD patients: more than a comorbidity. *Int J Chronic Obstr Pulm Dis*. 2017;12:1401-1411. https://doi.org/10.2147/COPD.S132961
31. Price A, Ramachandran S, Smith GP, Stevenson ML, Pomeranz MK, Cohen DE. Oral allergy syndrome (pollen-food allergy syndrome). *Dermatitis*. 2015;26(2):78-88. https://doi.org/10.1097/DER.0000000000000087
32. Bernstein JA, Cremonesi P, Hoffmann TK, Hollingsworth J. Angioedema in the emergency department: a practical guide to differential diagnosis and management. *Int J Emerg Med*. 2017;10(1):15. https://doi.org/10.1186/s12245-017-0141-z
33. Ko FW, Chan KP, Hui DS, et al. Acute exacerbation of COPD. *Respirology*. 2016;21(7):1152-1165. https://doi.org/10.1111/resp.12780
34. Abuzakouk M, Ghorab OKHA, Wahla AS, et al. Efficacy and safety of biologic agents in chronic urticaria, asthma and atopic dermatitis - a real-life experience. *Open Respir Med J*. 2020;14:99-106. https://doi.org/10.2174/1874306402014010099
35. Heffler E, Saccheri F, Bartezaghi M, Canonica GW. Effectiveness of omalizumab in patients with severe allergic asthma with and without chronic rhinosinusitis with nasal polyps: a PROXIMA study post hoc analysis. *Clin Transl Allergy*. 2020;10(25). https://doi.org/10.1186/s13601-020-00330-1
36. Damask CC, Ryan MW, Casale TB, et al. Targeted molecular therapies in allergy and rhinology. *Otolaryngol Head Neck Surg*. 2021;164(1_Suppl l):S1-S21. https://doi.org/10.1177/0194599820965233
37. Pelaia C, Paoletti G, Puggioni F, et al. Interleukin-5 in the pathophysiology of severe asthma. *Front Physiol*. 2019;10:1514. https://doi.org/10.3389/fphys.2019.01514
38. Roufosse F. Targeting the interleukin-5 pathway for treatment of eosinophilic conditions other than asthma. *Front Med*. 2018;5:49. https://doi.org/10.3389/fmed.2018.00049

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Morandini P, Laino ME, Paoletti G, et al. Artificial intelligence processing electronic health records to identify commonalities and comorbidities cluster at Immuno Center Humanitas. *Clin Transl Allergy*. 2022;e12144. https://doi.org/10.1002/clt2.12144