Received: 24 May 2023 | Accepted: 13 July 2023

DOI: 10.1002/jvc2.217

ORIGINAL ARTICLE

JEADV CLINICAL PRACTICE
OPEN ACCESS JOURNAL OF THE EUROPEAN ACADEMY OF DERMATOLOGY & VENEREOLOGY

The DERMACLEAR study: Verification results of a natural language processing system in dermatology

Francisco J. Ortiz de Frutos[1] | Ana M. Giménez-Arnau[2] | Lluís Puig[3] | Juan F. Silvestre[4] | Esther Serra[3] | Laura Salgado-Boquete[5] | Vicente García-Patos[6] | Jose L. L. Estebaranz[7] | Jaime Notario[8,9] | Ana Martin-Santiago[10] | Gabriel M. Pontevia[11] | Víctor Martín[12] | Guillermo Guinea[12] | Pau Terradas[12] | Esteban Daudén[13]

[1]Hospital Universitario 12 de Octubre, Universidad Complutense, Madrid, Spain

[2]Department of Dermatology, Hospital del Mar-IMIM, Universitat Pompeu Fabra, Barcelona, Spain

[3]Servicio de Dermatología, IIB SANT PAU, Hospital de la Santa Creu i Sant Pau, Barcelona, Spain

[4]Hospital General Universitario Dr Balmis, Alicante, Spain

[5]Complejo Hospitalario Universitario de Pontevedra, Pontevedra, Spain

[6]Hospital Universitari Vall d'Hebron, Barcelona, Spain

[7]Hospital Universitario Fundación Alcorcón, Madrid, Spain

[8]Hospital de Bellvitge, Barcelona, Spain

[9]Hospitalet de Llobregat, Barcelona, Spain

[10]Hospital Universitari Son Espases, Palma de Mallorca, Spain

[11]IOMED Medical Solutions, Barcelona, Spain

[12]Novartis Farmacéutica S.A, Barcelona, Spain

[13]Department of Dermatology, Hospital Universitario de la Princesa, Instituto de Investigación Sanitaria (IIS-HP), Madrid, Spain

Abstract

Background: Accurately determining the epidemiology of dermatological diseases such as hidradenitis suppurativa (HS), psoriasis (PsO), chronic urticaria (CU) and/or atopic dermatitis (AD) is challenging due to variations in prevalence and disease severity in the reported literature.

Objectives: The DERMACLEAR study aims to use natural language processing (NLP) to assess the proportions of patients with HS, PsO, CU and/or AD, and obtain information on patient profiles, patient journeys, and disease and healthcare burden in Spain. Here, the study design and objectives of the DERMACLEAR study are described and the precision of the NLP system used is assessed.

Methods: This study will retrospectively collect patient information from electronic health records (EHRs) at dermatology departments from seven tertiary hospitals in Spain. The NLP system was developed by IOMED Medical Solutions and was verified internally (IOMED scientific team) and externally (principal investigators of each hospital) to determine its precision in identifying patients with HS, PsO, CU and/or AD. Furthermore, internal verification was performed on other medical variables relevant to the study.

Results: To date, the DERMACLEAR study has retrospectively collected data from 54,458 patients with HS, PsO, CU and/or AD (HS: 5045; PsO: 32,559; CU: 8397; AD: 12,492). The average precision of the NLP system to identify patients diagnosed with HS, PsO, CU, and/or AD across all hospitals exceeded 95% via external and internal verification.

Conclusions: Results from the DERMACLEAR study will increase the real-world evidence of clinical practice, obtaining a large amount of information on patients with the studied diseases. The NLP system used is precise in

Correspondence
Esteban Daudén, Department of Dermatology, Hospital Universitario de la Princesa, IIS-HP, Madrid, Spain.
Email: estebandauden@gmail.com

Funding information
Novartis Farmacéutica

identifying patients diagnosed with HS, PsO, CU and/or AD, and other medical variables from EHRs, highlighting that it is a valid system to use in the DERMACLEAR study.

KEYWORDS
deep learning, dermatology, machine learning, natural language processing, OMOP CSM, real world data

INTRODUCTION

Hidradenitis suppurativa (HS), psoriasis (PsO), chronic urticaria (CU) and atopic dermatitis (AD) are chronic, immune-mediated skin diseases with relapsing courses that require long-term management; these skin diseases are associated with substantial morbidity.[1] Accurately determining the epidemiology of HS, PsO, CU and/or AD is complex due to variations in prevalence and disease severity reported in the literature. Furthermore, these diseases can result in increased costs for patients, an increased burden on hospitals, and a diminished quality of life for patients.[2]

Big data, the analyses of large data sets, often utilizes machine learning algorithms to gain insights into large volumes of clinical data, including unstructured data (e.g., clinical notes from electronic health records [EHRs]), to accelerate innovation strategies in healthcare.[3] An EHR is the digital version of a paper clinical chart, which stores all digital documents related to a patient's health including medical history, diagnoses, medications, treatment plans, imaging data, and laboratory test results.

Artificial intelligence (AI) technology is increasingly used in healthcare to manage EHRs and improve patient management.[4] Reading, capturing and processing EHR data is useful for obtaining valuable epidemiological insights for various diseases, which may support timely treatment of patients and accelerate personalized care.[5] Real-world big data is increasingly being used in clinical settings to mine existing clinical and epidemiological data in various therapeutic domains,[6,7] including dermatology.[8–11]

Natural language processing (NLP) is an AI tool that analyses large amounts of text data, converting it into structured data that can be analysed more easily, providing meaningful insights.[12–15] NLP can extract large amounts of data from EHRs with unprecedented efficiency by reducing the time spent on manual searching.[5] NLP technology has been utilized in previous studies to extract information from unstructured clinical notes[16–18]; for example, to identify cases of psoriatic arthritis from a large volume of EHRs.[19] However, the real-world application of AI tools, including NLP, pose several challenges in terms of accuracy and reliability.[20] Therefore, a robust process to verify the clinical utility of big data analytics is required, and each NLP system should be verified for accuracy.

The DERMACLEAR study aims to retrospectively collect patient information from EHRs using an NLP system in dermatology departments from seven tertiary hospitals in Spain. Here we present the study design and objectives of the DERMACLEAR study and the results of the verification of the NLP system used in the study.

METHODS

Study design

The DERMACLEAR study is a large, national, multicentre, noninterventional study that will retrospectively collect patients' information from EHRs in dermatology departments from seven tertiary hospitals in Spain. The primary aim of the study is to provide a comprehensive overview of the number and proportion of patients diagnosed with HS, PsO, CU and/or AD referred to these dermatology departments over 6 years (June 2015–June 2021). The study will further assess the patient journey and clinical management, the change in treatment patterns over time and the burden on the healthcare system. The objectives and endpoints are outlined in Table 1.

The DERMACLEAR study design is shown in Figure 1. Briefly, a data processing unit was installed in the seven hospitals, and all data collection were performed by IOMED Medical Solutions. Data were gathered, anonymized and stored on the servers of the hospitals for optimal security; IOMED's cloud servers will be used to analyse the EHRs. To select and include eligible patients, and to collect and organize data, EHRs will be processed by a data analytics platform which includes an NLP system[16–18] developed by IOMED. This medical NLP system was retrained for this study by IOMED, updating existing NLP models, and designing and training new models via data annotation performed by physicians. IOMED extracted and

TABLE 1 Objectives and related endpoints of the DERMACLEAR study.

	Objective	Endpoint
Primary		
1	• To determine the proportions of patients with HS, PsO, CU and/or AD, and any combination of them visiting dermatology departments in 7 tertiary hospitals in Spain during the last 6 years (June 2015–June 2021)	• The proportion of patients with HS, PsO, CU and/or AD, and any combination of them visiting dermatology departments in 7 tertiary hospitals in Spain during the last 6 years (June 2015–June 2021)
Secondary		
2	• To describe patient profiles	• Demographic characteristics • Clinical characteristics • Laboratory and other complementary tests at diagnosis, and during the disease course
3	• To describe the patient journey	• Referral patterns from other medical departments (including GPs) to dermatology departments • Referral patterns from dermatology departments to other medical departments (including GPs) • Time since symptom onset, first visit, diagnosis, treatment, first positive response/remission to relapse
4	• To describe the patient treatment patterns	• Sequence/duration of treatments • Reasons for discontinuing biologic/nonbiologic therapies • Restarting, discontinuing or switching therapies
5	• To describe patient disease activity	• Severity scores before, during and after treatment • Clinical response to therapies according to the natural language used in clinical practice
6	• To measure the healthcare burden	• Annual visits, hospitalizations, annual emergency visits, annual laboratory tests and other complementary tests • Annual prescription medication prescribed in each hospital
7	• To analyse the previous objectives in different patient subpopulations	• To describe endpoints 1–6 according to the different patient profiles for HS, PsO, CU and/or AD: demographic and clinical characteristics, laboratory and other complementary tests at diagnosis, and during the disease course
8	• To analyse the correlation between patient journey milestones and demographic, clinical and laboratory study variables	• Time since symptom onset, first visit, diagnosis, treatment, first positive response and remission to relapse • Demographic and clinical characteristics, and laboratory tests • Disease activity over time
Exploratory		
9	• To explore the proportions of patients with HS, PsO, CU and/or AD, and/or any combination of them who did not visit a dermatologist but visited other hospital medical specialists	• The proportions of patients with HS, PsO, CU and/or AD, and/or any combination of them who did not visit a dermatologist but visited other hospital medical specialists
10	• To explore the referral patterns of patients with HS, PsO, CU and/or AD, and/or any combination of them who were not referred to dermatologists but to other hospital medical specialists instead	• Referral patterns of patients with HS, PsO, CU and/or AD, and/or any combination of them who were not referred to dermatologists but to other hospital medical specialists instead

Abbreviations: AD, atopic dermatitis; CU, chronic urticaria; GP, general practitioner; HS, hidradenitis suppurativa; PsO, psoriasis.

standardized clinical data from participating hospitals, anonymized and processed clinical notes employing this NLP system, extracted study variables, assessed the quality of variables via manual annotation and retrained and updated the NLP system to improve output metrics (Figure 2a).

Figure 2b details the distribution of IOMED processing components between the hospital infrastructure and IOMED's cloud servers. Each hospital database was the clinical data source, from which a connector component extracted and transformed the data into an international standard

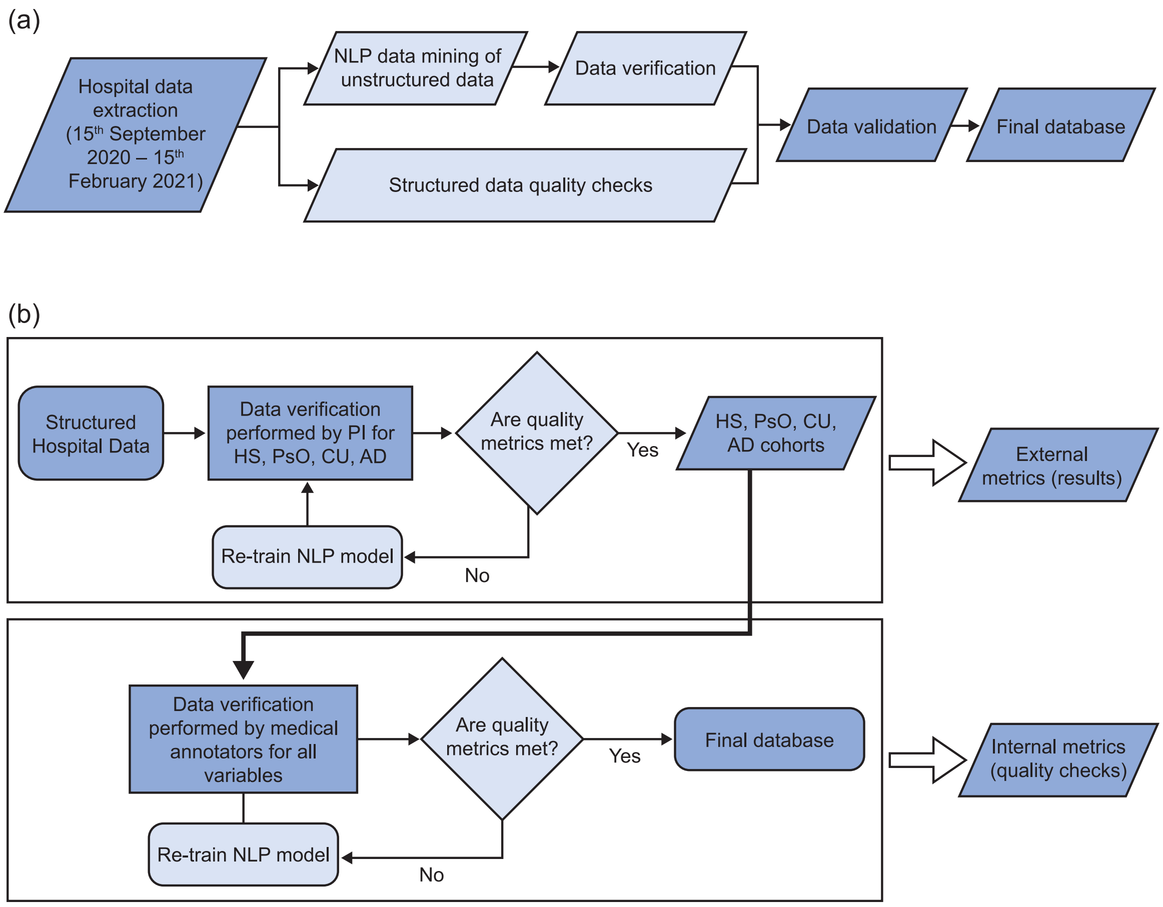
(a)
Hospital data extraction (15th September 2020 – 15th February 2021)
NLP data mining of unstructured data
Data verification
Structured data quality checks
Data validation
Final database
(b)
Structured Hospital Data
Data verification performed by PI for HS, PsO, CU, AD
Are quality metrics met?
Yes
HS, PsO, CU, AD cohorts
External metrics (results)
Re-train NLP model
No
Data verification performed by medical annotators for all variables
Are quality metrics met?
Yes
Final database
Internal metrics (quality checks)
Re-train NLP model
No

FIGURE 1 Study design of the DERMACLEAR study. Flow chart detailing the DERMACLEAR study design. (a) Data extraction (both structured and unstructured data) from hospital sites to the final database lock; data will be extracted from hospital sites from 15 September 2020 to 15 February 2021. Structured data will undergo quality checks and unstructured data (clinical notes) will undergo data mining by the NLP system, before data verification. Both structured and unstructured data, will be validated to yield the final database lock. (b) Structured hospital data will be verified by the PI for the diseases of interest (HS, PsO, CU and/or AD). Should these data not meet quality metrics, the NLP will be retrained. If these data meet quality metrics, the external metrics will be available (study results). Internal metrics (for quality checks) will also be verified through the same process. AD, atopic dermatitis; CU, chronic urticaria; HS, hidradenitis suppurativa; NLP, natural language processing; PI, principal investigator; PsO, psoriasis.

(Observational Medical Outcomes Partnership Common Data Model). The clinical text was anonymized, normalized and stored in a database managed by IOMED at each hospital. Clinical texts, which may have been in a variety of different formats (including tags inside the text, headings and sections) were converted to plain text, streamlining it to remove these elements and yield consistent text from all hospitals. Anonymized and normalized texts were processed with the NLP system using the cloud infrastructure; the NLP results were stored on-site inside the hospital, with no data stored in the cloud infrastructure.

Inclusion and exclusion criteria

Eligibility criteria are individuals aged ≥18 years with a diagnosis of HS, PsO, CU and/or AD. Patients with a diagnosis of HS, PsO, CU and/or AD needed to visit ≥1 outpatient visit at a dermatology clinic or other department of the corresponding hospital during the previous 6 years. Patients were excluded if no EHR data were available.

Data sources, management and collection process

The data source for the DERMACLEAR study was the EHR for each patient pertaining to the diseases that were being investigated. To identify and extract information from EHRs, the NLP system 'Medical Language API' was used. This software has successfully passed due diligence, information technology security assurance and data protection impact assessment processes to adhere to data

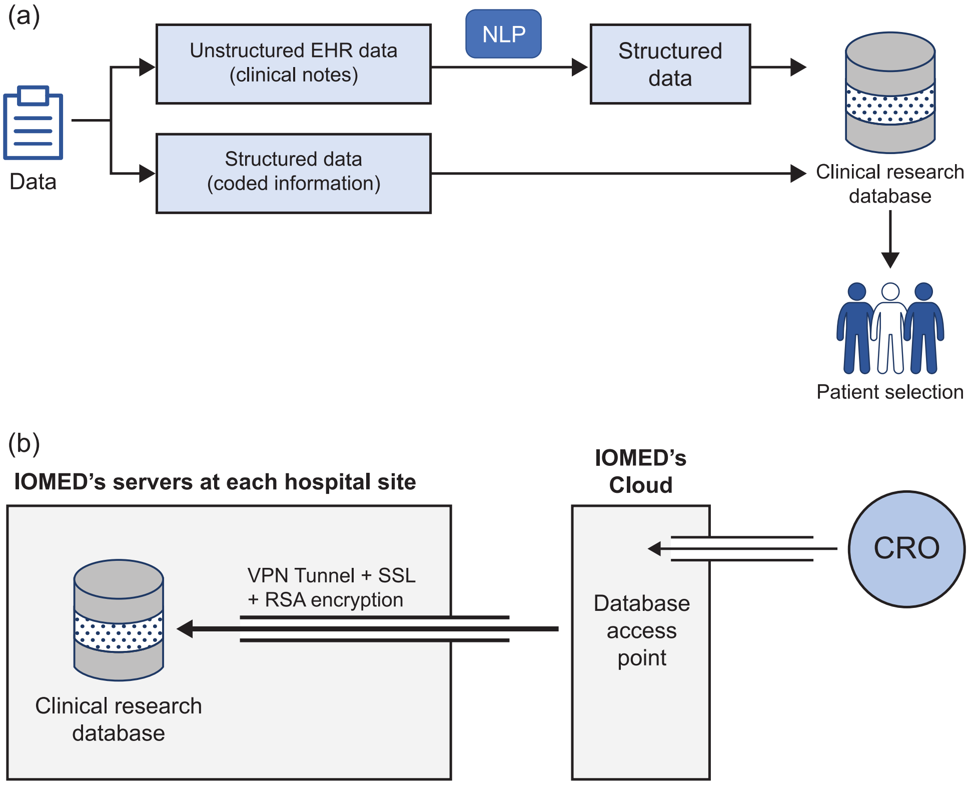
(a)
Data
Unstructured EHR data (clinical notes)
Structured data (coded information)
NLP
Structured data
Clinical research database
Patient selection
(b)
IOMED's servers at each hospital site
Clinical research database
VPN Tunnel + SSL + RSA encryption
IOMED's Cloud
Database access point
CRO

FIGURE 2 Electronic health record data extraction and data access. (a) Illustration of the role of the NLP system in the conversion of EHRs into structured data. Each hospital database was the source of the clinical data, from which a connector component extracted and transformed the data into an international standard (OMOP-CDM). Structured data (coded information) will be incorporated directly into the clinical research database. Unstructured data (e.g., clinical notes from EHRs) will be converted into structured data using NLP system before incorporation into the clinical research database. The database will be used to select eligible patients. (b) Illustration of the access to clinical data by third parties such as a CRO or IOMED staff. Each hospital will be securely connected to IOMED's cloud, where the cloud database (termed "database access point") serves as a point of access to IOMED's databases for each hospital. This unique access point will ensure control of access and facilitate the distribution of database queries to the hospital sites. CRO, contract research organization; EHR, electronic health record; NLP, natural language processing; OMOP-CDM, Observational Medical Outcomes Partnership-Common Data Model; RSA, Rivest-Shamir-Adleman encryption; SSL, Secure Sockets Layer; VPN, virtual private network.

privacy policy according to the General Data Protection Regulation.

IOMED sent the database to the contract research organization (CRO), Dynamic Science S.L., in a pre-specified format to perform the statistical analysis (Figure 3). The sponsor (Novartis Farmacéutica S.A.) has no access to the overall database or any hospital database.

The data anonymization process was divided into anonymization of structured data (e.g., laboratory tests) and of unstructured data (e.g., clinical notes). Structured data were anonymized via the removal of personal information from tables and by replacing personal identifiers with hashes. Clinical notes were anonymized via NLP. All anonymization processes were performed at each individual hospital site (Figure 3) using the IOMED server installed onsite.

Verification of the NLP system

Verifying NLP systems implies evaluating the precision and recall. However, in this study, where millions of clinical notes were processed by the NLP system, measuring recall was not possible, as the total number of mentions of diseases in the clinical notes would need to be known in advance; this would require manual revision of all notes, which would be intractable. A sample of notes could be reviewed to estimate the total number of mentions, but due to the sparsity of disease mentions, this estimation would not be reliable. Due to this limitation, recall was not calculated and only precision metrics are provided. However, a previously reported strategy was used during the NLP system's development to reduce false negatives and improve recall.[21]

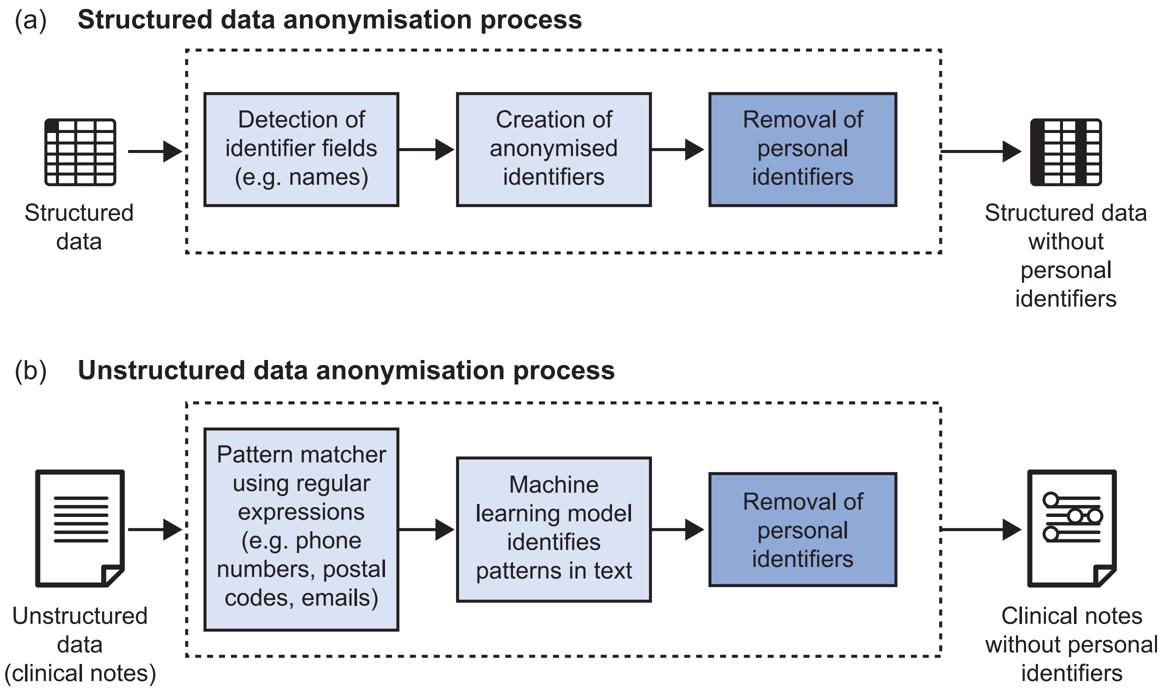
(a) Structured data anonymisation process
Structured data
Detection of identifier fields (e.g. names)
Creation of anonymised identifiers
Removal of personal identifiers
Structured data without personal identifiers
(b) Unstructured data anonymisation process
Unstructured data (clinical notes)
Pattern matcher using regular expressions (e.g. phone numbers, postal codes, emails)
Machine learning model identifies patterns in text
Removal of personal identifiers
Clinical notes without personal identifiers

FIGURE 3 Data anonymization process. Illustration of the anonymization process to be conducted by IOMED. (a) Anonymization of structured data will involve detection of identifier fields, such as patient names and surnames, followed by the creation of anonymized identifiers and the removal of all patient personal identifiers to yield structured data without any personal identifiers. (b) Anonymization of unstructured data (clinical notes) will involve identifying regular expressions (e.g., phone numbers, postal codes and email addresses) using a pattern matcher. A machine learning model known as a probabilistic pattern matcher will be applied to identify patterns in the clinical notes before all personal identifiers are removed to yield clinical notes without any personal identifiers.

In this study, a data point relates to a specific mention of a variable of interest (e.g., in an unstructured data source; this could be mention of a medical term such as 'psoriasis' in a clinical note). Unstructured data points captured via a NLP system are subject to error and need to be verified for precision. The occurrence of HS, PsO, CU and/or AD in EHRs, and other medical variables of interest (e.g., type of drugs), were searched using the NLP system.

Both external and internal verification of the NLP system were performed; external verification was performed by the principal investigator of each hospital to verify the precision of the NLP system to identify patients with HS, PsO, CU and/or AD. Internal verification was performed by the IOMED team to verify the precision of the NLP system to identify patients with HS, PsO, CU and/or AD, as well as other medical variables of interest. Further methodological information is detailed in the Supporting Information: Methods.

The NLP system included models for named entity recognition (NER), named entity liking (NEL) and context detection. The NER model identifies a medical term (e.g., 'psoriasis') in an EHR and subsequently assigns a semantic category to the medical term such as 'disease'. The NEL model then associates each specific medical term to a specific code under a medical coding system to represent the specific meaning to the term in the given context. Subsequently, context detection models determine whether the mention of a variable was negated, uncertain or referred to the past, present or future. The result of the NLP system is a series of mentions in the text of each of the desired variables (e.g., disease type), where each entity includes a series of contextual attributes determining its temporality (present/future/past), negation (negative/positive) or certainty (certain/uncertain). Precision was calculated for each variable of interest per hospital by using the formula: (1 − NER error rate) × number of annotations in that variable.

Data analysis

All information from patients who fulfill the selection criteria will be included in the data set for analysis. A descriptive analysis of the variables included in the study will be performed. For continuous variables, the mean (SD) or median (first quartile, third quartile), and minimum and maximum values will be reported. Categorical variables will be described by absolute and relative frequencies. For categorical and continuous variables, the number of observations and missing data will be specified. To minimize this risk of bias, unavailable data/assessments will be stated as 'not available' and be noted as 'missing' in the statistical analyses. It is expected, however, that most data of interest in this study will be registered in medical records, as they are part of patients' routine clinical follow-ups.

Where appropriate, linear regression and logistical regression models will be used to assess correlations between patient journey milestones and demographic, clinical and laboratory study variables. All analyses will be performed using the designated CRO and will be performed using the SAS system package (version 9.4 or higher).

RESULTS

Patients

To date, the DERMACLEAR study has retrospectively collected data from 54,458 patients with HS, PsO, CU and AD (HS: 5045; PsO: 32,559; CU: 8397; AD: 12,492) attending dermatology clinics across seven tertiary hospitals in Spain in the previous 6 years (June 2015-June 2021).

Precision of the NLP system to identify patients with HS, PsO, CU and/or AD

Based on a sample size calculation for each hospital site, 683 patients with HS, 756 patients with PsO, 821 patients with CU and 896 patients with AD were selected as a sample for verification (Table 2). The average precision of the NLP system across all seven hospitals exceeded 95% in identifying all four dermatological diseases via both external and internal verification processes (Figure 4). The external precision was 99.9% for HS, 98.8% for PsO, 100.0% for CU and 98.3% for AD. The internal precision was 100.0% for HS, CU and/or AD, and 96.7% for PsO.

Precision of the NLP system in identifying all medical variables of interest

The precision of the NLP system, based on internal verification, to identify all variables of interest was high (≥95%) for most variables. A total of 367 variables of interest were categorized into seven groups: type of disease/disorder (33.2%, 122/367), type of drug (29.2%, 107/367), body area affected (11.4%, 42/367), outcome measure/diagnostic test (8.7%, 32/367), symptom (7.9%, 29/367), miscellaneous (7.9%, 29/367) and nonpharmacological treatment (1.9%, 7/367). The precision of each variable is detailed in Supporting Information: Table S1.

DISCUSSION

Obtaining accurate epidemiological insights and disease characteristics from large patient cohorts remains a challenge in dermatological diseases such as HS, PsO, CU and/or AD due to variations in prevalence and disease severity in the reported literature.[1] The use of NLP in processing information from EHRs represents a unique opportunity to gain insights into disease epidemiology and improve patient management.[14]

The current study investigated the precision of the NLP system that was used to extract data from EHRs in dermatology departments across seven tertiary hospitals in Spain as part of the DERMACLEAR study. Overall, both external and internal verification of the NLP system's precision to identify all four dermatological diseases of interest was high, with an average precision ≥95%. For all variables of interest ($N = 367$), the precision remained high (≥95%) for most variables.

TABLE 2 The proportion of patients included in the random sample for verification in each dermatological disease.

Disease	Number of patients overall across the 7 hospital sites	Number of patients in the random sample for verification	Proportion of patients overall in the random sample for verification (%)
HS	5045	683	13.5
PsO	32,559	756	2.3
CU	8397	821	9.8
AD	12,492	896	7.2

Note: Patients with ≥1 dermatological disease of interest were included in each disease row but only included once in the total number of patients.

Abbreviations: AD, atopic dermatitis; CU, chronic urticaria; HS, hidradenitis suppurativa; PsO, psoriasis.

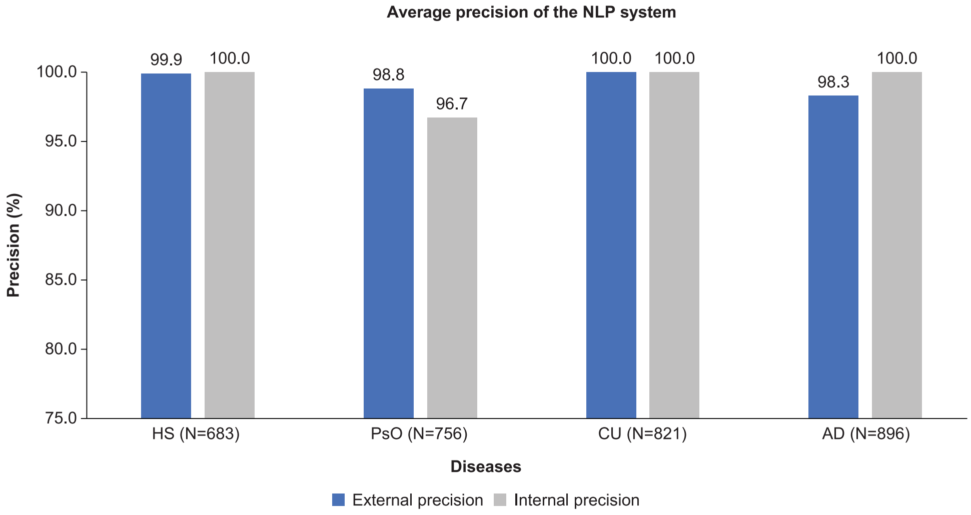

FIGURE 4 Precision of the NLP system. The average external and internal precision of the NLP system to identify patients with the four dermatological diseases of interest based on a random sample from the seven tertiary hospitals. AD, atopic dermatitis; CU, chronic urticaria; HS, hidradenitis suppurativa; N, number of patients included in the random sample; NLP, natural language processing; PsO, psoriasis.

The results of the DERMACLEAR study are reliant on the NLP system being precise when collecting and processing patient data from EHRs. The high precision observed in this study is reassuring and enables physicians to trust the system to extract and process health information accurately. The precision of the NLP system used in the DERMACLEAR study is higher than that reported in an onco-dermatology NLP study that assessed basal cell carcinoma[22] and in a neurology study in patients with epilepsy,[23] potentially due to the different NLP systems used.

NLP systems have the potential to improve epidemiology disease awareness; NLP has been utilized in previous studies to extract information from unstructured clinical notes, including the dermatology setting.[16–18,24] Morandini et al.[16] used NLP to assess clinical notes to identify diseases that present together as comorbidities. Through NLP, they reported that, of 2057 patient records, the most common conditions that occur together were oral allergy syndrome and urticaria, angioedema and urticaria, and rhinitis and asthma. A large study in the United States ($n = 133{,}025$) used NLP to extract information from clinical notes of patients with AD and identified unmet needs in physician care; the study revealed that while physician notes for patients with AD recorded symptoms and treatment strategies, the notes had no mention of quality of life and burden of disease.[24] A recent study by Malden et al.[25] has shown that NLP can be used to identify coronavirus disease 2019 (COVID-19) symptoms from the unstructured EHRs of 359,938 patients with COVID-19. Although NLP has been previously employed to extract information from clinical notes of patients with AD,[24] the current study will be the first, to the authors' knowledge, that uses NLP to extract information from the clinical records of patients with HS, PsO, CU and/or AD.

This first-in-kind study will implement AI through NLP to process EHRs and determine the proportions of patients with HS, PsO, CU and/or AD in Spain, as well as describe patient profiles, patient journeys and the disease and healthcare burden. The DERMACLEAR study will include a large volume of EHRs from patients attending multiple hospitals across Spain over 6 years, making the upcoming results representative of the population of interest in Spain. Using EHRs is difficult for conducting clinical research, as EHR data are typically presented as unstructured clinical narratives; results from the DERMACLEAR study will help address machine-readability problems and increase awareness of the prevalence, clinical unmet needs and patient profiles of each of the studied diseases. The results will also highlight the importance of a comprehensive follow-up and appropriate disease management for patients suffering from HS, PsO, CU and/or AD.

Study limitations

The real-world data collected in the DERMACLEAR study will be generated from EHRs from dermatologists in tertiary hospitals, which has some potential limitations. Selection bias may produce false associations if the study population does not reflect the population of interest. To avoid this, included patients will comprise a heterogeneous population with the only selection criteria being adults with a diagnosis of HS, PsO, CU and/or AD and ≥1 outpatient visit during the last 6 years. The difficulty of extracting clinical data from unstructured EHRs by NLP and their conversion into research variables may represent a quality limitation. In clinical databases, data from EHRs can be inaccurate, incomplete, fragmented and inconsistent, and are not subject to a quality procedure to ensure data integrity and accuracy. Furthermore, some variables expressed as acronyms, scores, or text with numeric values (e.g., PASI75) can be difficult for the AI system to capture without context.

A further limitation is that the proportion of patients included in the random sample for verification across the four dermatological diseases was not equal and was small relative to the total population enroled. Finally, data access and data sources are not homogeneous between hospital sites and there may be some duplication; these reports must be de-duplicated, which may be time consuming.

CONCLUSION

The DERMACLEAR study will increase the real-world evidence of clinical practice, obtaining a large amount of information on patients with HS, PsO, CU and/or AD attending hospitals in Spain. Results from the DERMACLEAR study will provide insight into the disease prevalence, clinical unmet needs and patient profile of patients with HS, PsO, CU and/or AD. This may further elucidate the role of a comprehensive follow-up and disease management of patients suffering from these dermatological diseases. The precision of the NLP system used in the DERMACLEAR study is high (≥95%) in identifying HS, PsO, CU and/or AD in EHRs, verifying that it is a valid instrument with clinical utility for the DERMACLEAR study.

AUTHOR CONTRIBUTIONS

All authors were involved in the conception and design or analysis and interpretation of the data, drafting of the manuscript or revising it critically, and read and approved the final manuscript.

ACKNOWLEDGEMENTS

The authors thank Alvaro Abella and Anna Nicolau Torra (IOMED Medical Solutions, Barcelona, Spain) for providing support for data extraction and data management in this study. The authors also thank Philip O'Gorman, PhD, Ellen McKenna, MSc (Novartis Ireland Ltd, Dublin, Ireland) and Asma S (Novartis Healthcare Pvt Ltd, Hyderabad, India) for providing medical writing support and assistance, which was funded by Novartis Pharma AG, Basel, Switzerland, in accordance with Good Publication Practice (GPP 2022) guidelines (https://www.ismpp.org/gpp-2022). This work was supported by Novartis Farmacéutica S.A, Madrid, Spain.

CONFLICT OF INTEREST STATEMENT

Francisco J. O. de Frutos has participated as a medical advisor and received funding for his collaboration in different research projects for Sanofi Genzyme, Novartis, Astellas Pharma, MSD and Uriach y Laboratorios Viñas. Ana M. Giménez-Arnau participated as a medical advisor for Uriach Pharma, Genentech, Novartis, FAES, GSK, Sanofi–Regeneron, Amgen, Thermo Fisher Scientific, Almirall, received research grants from Uriach Pharma, Novartis, Instituto Carlos III- FEDER and has carried out training activities for Uriach Pharma, Novartis, Genentech, Menarini, LEO-PHARMA, GSK, MSD, Almirall and Sanofi. Lluís Puig has received remuneration for consultancy/conferences from Abbvie, Almirall, Amgen, Baxalta, Biogen, Boehringer Ingelheim, Celgene, Gebro, Janssen, LEO Pharma, Lilly, Merck-Serono, MSD, Mylan, Novartis, Pfizer, Regeneron, Roche, Sandoz, Samsung-Bioepis, Sanofi and UCB, and has participated in meetings for Celgene, Janssen, Lilly, MSD, Novartis and Pfizer. Juan F. Silvestre has served as a consultant and received speaker fees at educational events for Sanofi Genzyme, Regeneron, Abbvie, Eli Lilly, Galderma, LEO Pharma, Novartis and Pfizer, and has served as the principal investigator in clinical trials sponsored by AbbVie, Amgen, Bristol Meyer Squibb, Eli-Lilly, Incyte, LEO Pharma, Novartis, Pfizer and Sanofi Genzyme outside of submitted work. Esther Serra participated in conferences, meetings and as a researcher for Almirall, Leo, Faes, Novartis, Sanofi, Galderma, Lilly and Pfizer. Laura Salgado is a board member, served as a consultant, received grants, supported research, participated in clinical trials and/or received speaking fees for Abbvie, Almirall, Janssen-Cilag, LEO Pharma, Novartis, Pfizer, MSD-Schering-Plough, Celgene, Lilly, UCB, Eucerin, Bristol-Myers Squibb, Biogen and Amgen. Vicente García-Patos served as a consultant, received grants, participated in clinical trials and/or received speaking fees from Almirall, Janssen, LEO Pharma, Pfizer, Lilly, Abbvie, Sanofi Genzyme, Novartis, MSD, Laboratorios Viñas, Pierre-Fabre and Isdin. Jose L. L. Estebaranz served as a consultant, participated in clinical trials and/or received speaking fees from Almirall, Janssen, LEO Pharma, Lilly, Abbvie, Bioderma, Novartis, Pierre-Fabre and Isdin. Jaime Notario has perceived consultancy/speakers' honoraria and/or participated in clinical trials sponsored by AbbVie, Almirall, Celgene, Gebro, Janssen, LEO Pharma, Lilly, MSD, Novartis and Pfizer. Ana Martin-Santiago served as a consultant or has received

speaking fees at educational events for AbbVie, Amgen, Janssen, LEO Pharma, Leti, Lilly, Mylan, Novartis, Pierre-Fabre, Pfizer, Sanofi-Genzyme and UCB. Gabriel M. Pontevia is an employee at IOMED Medical Solutions, Barcelona, Spain. Victor Martin, Guillermo Guinea and Pau Terradas are full-time employees of Novartis Pharmaceutical S.A., Madrid (Spain). Esteban Daudén has the following conflict of interests: Advisory board member, consultant, grants, research support, participation in clinical trials, honorarium for speaking, research support, with the following pharmaceutical companies: Abbvie/Abbott, Almirall, Amgen, Janssen-Cilag, LEO Pharma, Novartis, Pfizer, MSD-Schering-Plough, Celgene, Lilly and UCB.

DATA AVAILABILITY STATEMENT

Novartis is committed to sharing with qualified external researchers, access to patient-level data and supporting clinical documents from eligible studies. These requests are reviewed and approved by an independent review panel on the basis of scientific merit. All data provided are anonymized to respect the privacy of patients who have participated in the trial in line with applicable laws and regulations.

ETHICS STATEMENT

The DERMACLEAR study was a retrospective study based on secondary data that was conducted according to Ministerial Order SAS/3470/2009 on post-authorization observational studies for drugs for human use. This study complies with the general data protection regulation (GDPR), and the Container-Based Systems for Big data and Distributed and Parallel Computing (CBDP). For each hospital site, ethical approval was obtained by the Clinical Research Ethics Committee before the start of data collection.

ORCID

Ana M. Giménez-Arnau http://orcid.org/0000-0001-5434-7753
Lluís Puig http://orcid.org/0000-0001-6083-0952
Esteban Daudén http://orcid.org/0000-0002-0676-1260

REFERENCES

1. Karimkhani C, Dellavalle RP, Coffeng LE, Flohr C, Hay RJ, Langan SM, et al. Global skin disease morbidity and mortality: an update from the global burden of disease study 2013. JAMA Dermatol. 2017;153:406–12.
2. Lim HW, Collins SAB, Resneck Jr. JS, Bolognia JL, Hodge JA, Rohrer TA, et al. The burden of skin disease in the United States. J Am Acad Dermatol. 2017;76:958–72.
3. Habl C, Renner A-T, Bobek J, Laschkolnig A. Study on Big Data in public health, telemedicine and healthcare: final report, European Commission Directorate-General for Health and Food Safety Publications Office; 2016.
4. Wang Y, Wang L, Rastegar-Mojarad M, Moon S, Shen F, Afzal N, et al. Clinical information extraction applications: A literature review. J Biomed Inf. 2018;77:34–49.
5. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. 2019;25:44–56.
6. Wehner MR, Levandoski KA, Kulldorff M, Asgari MM. Research techniques made simple: an introduction to use and analysis of big data in dermatology. J Invest Dermatol. 2017;137:e153–8.
7. Okon E, Rachakonda V, Hong HJ, Callison-Burch C, Lipoff JB. Natural language processing of Reddit data to evaluate dermatology patient experiences and therapeutics. J Am Acad Dermatol. 2020;83:803–8.
8. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017;542:115–8.
9. Ogdie A, Rozycki M, Arndt T, Shi C, Kim N, Hur P. Longitudinal analysis of the patient pathways to diagnosis of psoriatic arthritis. Arthritis Res Ther. 2021;23:252.
10. Jalali-najafabadi F, Stadler M, Dand N, Jadon D, Soomro M, Ho P, et al. Application of information theoretic feature selection and machine learning methods for the development of genetic risk prediction models. Sci Rep. 2021;11:23335.
11. Han SS, Moon IJ, Lim W, Suh IS, Lee SY, Na JI, et al. Keratinocytic skin cancer detection on the face using region-based convolutional neural network. JAMA Dermatol. 2020;156:29–37.
12. Jha AK. The promise of electronic records: around the corner or down the road? JAMA. 2011;306:880–1.
13. Cohen KB, Hunter L. Natural anguage processing and systems biology. In: Dubitzky W, Azuaje F, editors. Artificial Intelligence Methods And Tools For Systems Biology. Dordrecht, The Netherlands: Springer; 2004. p. 147–73.
14. Koleck TA, Dreisbach C, Bourne PE, Bakken S. Natural language processing of symptoms documented in free-text narratives of electronic health records: a systematic review. J Am Med Inform Assoc. 2019;26:364–79.
15. Kaufman DR, Sheehan B, Stetson P, Bhatt AR, Field AI, Patel C, et al. Natural language processing-enabled and conventional data capture methods for input to electronic health records: a comparative usability study. JMIR Med Inform. 2016;4:e35.
16. Morandini P, Laino ME, Paoletti G, Carlucci A, Tommasini T, Angelotti G, et al. Artificial intelligence processing electronic health records to identify commonalities and comorbidities cluster at immuno center humanitas. Clin Transl Allergy. 2022;12:e12144.
17. Meskers CGM, van der Veen S, Kim J, Meskers CJW, Smit QTS, Verkijk S, et al. Automated recognition of functioning, activity and participation in COVID-19 from electronic patient records by natural language processing: a proof- of- concept. Ann Med. 2022;54:235–43.
18. Ducrot YM, Bruno E, Franco JM, Raffray L, Beneteau S, Bertolotti A. Scabies incidence and association with skin and soft tissue infection in Loyalty Islands Province, New Caledonia: a 15-year retrospective observational study using

electronic health records. PLoS Neglect Trop Dis. 2022; 16:e0010717.

19. Love TJ, Cai T, Karlson EW. Validation of psoriatic arthritis diagnoses in electronic medical records using natural language processing. Semin Arthritis Rheum. 2011;40:413–20.
20. Lee CH, Yoon HJ. Medical big data: promise and challenges. Kidney Res Clin Pract. 2017;36:3–11.
21. Quijada M, Vivó M, Abella-Bascarán Á, Chocrón P, Maeztu Gd. A framework for false negative detection in NER/NEL. In: Rosso P, Basile V, Martínez R, Métais E, Meziane F, editors. Natural Language Processing and Information Systems. Cham: Springer International Publishing; 2022. p. 323–30.
22. Ali SR, Strafford H, Dobbs TD, Fonferko-Shadrach B, Lacey AS, Pickrell WO, et al. Development and validation of an automated basal cell carcinoma histopathology information extraction system using natural language processing. Front Surg. 2022;9:870494.
23. Fonferko-Shadrach B, Lacey AS, Roberts A, Akbari A, Thompson S, Ford DV, et al. Using natural language processing to extract structured epilepsy data from unstructured clinic letters: development and validation of the ExECT (extraction of epilepsy clinical text) system. BMJ Open. 2019;9:e023232.
24. Pierce EJ, Boytsov NN, Vasey JJ, Sudaria TC, Liu X, Lavelle KW, et al. A qualitative analysis of provider notes of atopic dermatitis-related visits using natural language processing methods. Dermatol Ther. 2021;11:1305–18.
25. Malden DE, Tartof SY, Ackerson BK, Hong V, Skarbinski J, Yau V, et al. Natural language processing for improved characterization of COVID-19 symptoms: observational study of 350,000 patients in a large integrated health care system. JMIR Public Health Surveill. 2022;8:e41529.

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: Ortiz de Frutos FJ, Giménez-Arnau AM, Puig L, Silvestre JF, Serra E, Salgado-Boquete L, et al. The DERMACLEAR study: Verification results of a natural language processing system in dermatology. JEADV Clin Pract. 2023;2:775–785. https://doi.org/10.1002/jvc2.217