Master of Public Health

Master de Santé Publique

Identification of Hydroxychloroquine-Induced Retinal Toxicity in Lupus Patients Using Rule-Based Natural Language Processing on French Electronic Health Record Data

SOLENE BOSSART
Master of Public Health
2023-2025

Practicum location:
Pitié-Salpêtrière Hospital
PEPITES research team
Institut Pierre Louis d'Epidémiologie et de Santé Publique
Paris, France

Professional advisor
Alexandre Teboul
Pharmaco-epidemiology department
Pitié-Salpêtrière Hospital
PEPITES research team
Institut Pierre Louis d'Epidémiologie et de Santé Publique

Academic advisor:
Rebecca Kehm
Columbia University
Mailman School of Public Health

Acknowledgements

I am extremely grateful to all the people that made the opportunity to carry out this work possible!

To my parents and sisters, I owe it all to y'all for supportive me through an international move to continue my education and always being my biggest fans. Love you!

To my supervisors Dr. Alexandre Teboul and Audrey Berges, I extend my deepest gratitude for being so incredibly supportive and kind throughout this internship. This work would not have been possible without your endless guidance!!

To my classmates at EHESP whom I had the pleasure of sharing a classroom with for the past 2 years, I'm so thankful for all the lessons I learned from y'all and all the incredible memories we made together!

To Juan and Éabha, who have been my rocks throughout this masters, I wouldn't be the person I am without y'all!

And finally, to Lulu and Kelly, for being the best roommates ever and always letting me rant about everything that went wrong with my code that day. Your support means the world to me!

Table of Contents

List of Acronyms

Acronym	Full Term
AP-HP	Assistance Publique- Hôpitaux de Paris
CLE	Cutaneous Lupus Erythematosus
CI	Contraindication
EHR	Electronic Health Record
EDS	Entrepôt de Données de Santé
HCQ	Hydroxychloroquine
ICD-10	International Classification of Diseases, 10th revision
LE	Lupus Erythematosus
NLP	Natural Language Processing
OPH	Ophthalmology (general entity)
OPH2	Filtered Ophthalmology (general entity)
SLE	Systemic Lupus Erythematosus

Abstract

Hydroxychloroquine (HCQ) is a cornerstone of lupus treatment but carries a risk of retinal toxicity with prolonged use. Monitoring for hydroxychloroquine-induced maculopathy is essential, yet documentation of related follow-up and treatment decisions is often embedded in unstructured electronic health records (EHRs), limiting systematic analysis. This project aimed to develop and validate a rule-based natural language processing (NLP) pipeline to extract clinically relevant information on HCQ exposure, maculopathy, ophthalmology follow-up, and treatment discontinuation from French EHR data.

A rule-based NLP pipeline was created using manually annotated clinical notes and domain-informed regular expressions. The pipeline was evaluated against a gold-standard reference set of 600 annotated clinical notes and then applied at scale to three lupus cohorts—systemic lupus erythematosus, cutaneous lupus erythematosus, and a broader lupus erythematosus group—identified from the Clinical Data Warehouse of the Greater Paris University Hospitals.

The pipeline demonstrated high performance for core entities, achieving F1 scores of 0.997 for HCQ and 0.957 for maculopathy. Contextual filtering improved precision for ophthalmology mentions (F1 = 0.921). Applied to over 10,000 lupus patients, the algorithm revealed that documentation of maculopathy and related care actions varied between lupus subtypes, with systemic lupus patients showing higher prevalence and greater intensity of follow-up and treatment discontinuation.

These results demonstrate that rule-based NLP can effectively extract relevant safety signals and care trajectories from unstructured clinical text. Despite ongoing challenges, such as limited access to treatment timelines and structured dosing data, this approach highlights the feasibility of using NLP for retrospective surveillance and sets the stage for more detailed risk modeling as data granularity improves.

Keywords: lupus erythematosus, hydroxychloroquine, maculopathy, electronic health records, natural language processing

1. INTRODUCTION

1.1 Background and Clinical Problem

Lupus erythematosus (LE) is a chronic autoimmune disease that can affect multiple organ systems, including the skin, joints, kidneys, cardiovascular system, and central nervous system. It presents in two main forms: systemic lupus erythematosus (SLE), which is multi-organ and often severe, and cutaneous lupus erythematosus (CLE), which is primarily limited to skin involvement but may precede systemic disease (1,2). LE disproportionately affects women of childbearing age and is more prevalent in non-white populations (1).

Hydroxychloroquine (HCQ) is a cornerstone of lupus treatment and is recommended for all patients with SLE and CLE, unless contraindicated. It has been shown to reduce disease flares, delay organ damage, and improve long-term survival outcomes (3). However, prolonged use of HCQ carries a well-documented risk of retinal toxicity, which can lead to irreversible vision loss if not detected early (4,5). The risk is especially elevated in patients taking higher daily doses (>5.0 mg/kg/day) and those on treatment for over five years (4). Recent studies using advanced retinal imaging have revealed that the prevalence of HCQ-induced maculopathy may be significantly higher than previously thought, with early-stage retinal changes detectable before symptoms appear (4).

Because HCQ-induced maculopathy is irreversible, early detection through routine ophthalmologic screening is essential. Maculopathy refers to damage to the macula, the central part of the retina responsible for sharp, central vision. Current guidelines recommend baseline retinal examination and regular follow-up, especially after five years of therapy (4,5). However, real-world data suggests that adherence to these screening recommendations is variable, and monitoring practices are not systematically tracked across health systems (6). This gap presents a public health concern, as patients may remain on HCQ therapy without adequate ophthalmic surveillance.

Given that much of the relevant information in electronic health records (EHRs) is unstructured, including free-text clinical documentation such as hospitalization summaries, consultation notes, and prescription narratives, automated methods such as natural language processing (NLP) are increasingly used to extract clinically meaningful insights from free-text data (6–8). This is particularly important for detecting mentions of HCQ, ophthalmology visits, or maculopathy findings, which are often embedded in narrative text and not captured in structured fields. These

data points are typically absent from structured fields for several reasons: HCQ is often prescribed and dispensed in outpatient or private practice settings, so prescription data may not be captured within structured hospital records. Ophthalmology consultations are frequently conducted outside the hospital setting, and scheduled outpatient visits are often underrepresented in structured EHR fields. Maculopathy is also not coded with sufficient specificity to distinguish HCQ-induced cases, and relevant diagnoses may be documented in community-based ophthalmology settings. Leveraging NLP to extract this information from unstructured clinical notes enables a more complete view of HCQ prescribing and monitoring practices, ultimately supporting safer and more informed care for patients with lupus.

1.2 Current Applications of NLP in Clinical Research

Natural language processing and machine learning methods are increasingly applied to EHRs to support research and clinical surveillance in diseases like lupus, especially when structured data is incomplete or insufficient (6,7). NLP enables the automated extraction of key clinical information, such as medication mentions, test results, or treatment changes, from unstructured free-text notes (6,7).

In ophthalmology, NLP has been used to identify disease patterns, adverse effects, and follow-up behaviors across large volumes of clinical documentation (6). These methods can uncover critical information related to hydroxychloroquine (HCQ), including maculopathy diagnosis, ophthalmology visits, or treatment adjustments that are not consistently recorded in structured fields.

Despite these advances, relatively few studies have applied NLP to monitor treatment safety in lupus care. In particular, research on tracking HCQ-induced maculopathy or adherence to ophthalmologic screening guidelines remains limited, especially in French clinical settings. This gap restricts our ability to evaluate real-world safety practices and implement scalable monitoring systems.

1.3 Study Objectives and Hypothesis

We hypothesize that mentions of HCQ exposure, ophthalmologic follow-up, and HCQ-induced maculopathy can be accurately identified within EHRs using rule-based NLP methods. The primary objective of this project is to develop and validate an NLP algorithm capable of identifying confirmed cases of HCQ-induced maculopathy among a cohort of patients suspected to have

lupus in the Clinical Data Warehouse of the Assistance Publique des Hôpitaux de Paris (AP-HP), a network of 39 university hospitals in the Greater Paris area.

A secondary objective is to estimate the prevalence of maculopathy among HCQ-treated lupus patients identified by the pipeline and to compare this prevalence between patients with systemic lupus erythematosus (SLE) and cutaneous lupus erythematosus (CLE). This also supports the broader extraction of relevant follow-up and treatment information, enabling downstream analyses of care trajectories, screening practices, and treatment discontinuation patterns.

1.4 Student Contributions and Practicum Context

This work was carried out as part of an internship within the MAXYPLUS research project, hosted by the Pierre Louis Institute of Epidemiology and Public Health (IPLESP) and APHP. The project was supervised by Dr. Alexandre Teboul and Audrey Berges in collaboration with the PEPITES research team from IPSLEP. The project was conducted within the secure servers of the APHP Clinical Data Warehouse using pseudonymized patient data.

Contributions included designing the annotation schema, conducting manual annotations of clinical reports, developing and evaluating rule-based NLP algorithms, and performing data extraction and iterative refinement based on performance analysis.

2. METHODS

2.1 Data Source

This project uses data from the AP-HP Clinical Data Warehouse (Entrepôt de Données de Santé, EDS), which includes electronic health record (EHR) data from over 11 million patients across 39 university hospitals in the Greater Paris area. Available data include structured fields (ICD-10 diagnoses, CCAM procedures, laboratory results, drug prescriptions) and unstructured clinical narratives (hospitalization summaries, consultation notes, prescriptions, imaging, and pathology reports). All data are pseudonymized and accessible for research under an approved protocol validated by the AP-HP's independent Scientific and Ethics Committee (Comité Scientifique et Éthique, CSE; approval number: CSE-230003).

2.2 Study Population

The study focused on a cohort of possible lupus cases, encompassing both SLE and CLE. These individuals were previously identified through a systematic query of the APHP's comprehensive EHR system.

1. **Possible lupus cases data mart (N = 16,304):**

   This initial cohort consisted of patients identified through a broad query combining structured and unstructured data. The cohort extraction covered the period from 2017 to April 2023. The lower bound reflects the creation of the AP-HP Clinical Data Warehouse (EDS), which includes patients who presented at any AP-HP hospital since 2017, regardless of whether they had earlier care. For these patients, all clinical documents stored in the Orbis system were made available, including those dated as far back as 2012, when Orbis was first implemented at AP-HP and gradually expanded to all hospitals. Patients were included if they met at least one of the following criteria (inclusion was based on OR logic, not AND):

   - **Structured inclusion:** at least one ICD-10 diagnosis code related to systemic or cutaneous lupus erythematosus, as listed in the Appendix
   - **Unstructured inclusion:** presence of lupus-related keywords (e.g., "lupus", "lupique") in free-text clinical notes, detected via regular expression filters. These expressions included terms corresponding to SLE and CLE variants derived from the Unified Medical Language System (UMLS).

2. Predicted Lupus cases (LE, SLE, CLE subcohorts):

From the possible cases data mart, three subcohorts were derived using separate logistic regression models with LASSO penalization, developed and validated by the supervisory team. Each algorithm, combining structured EHR fields and NLP-derived features, was designed to capture a specific lupus phenotype. The distinction between systemic and cutaneous lupus was guided by clinical practice: SLE is typically classified using the 2019 ACR/EULAR criteria, which incorporates both clinical and immunological features (9). In contrast, patients with CLE often do not meet these criteria, except in cases with isolated biological markers. This clinical and immunological differentiation informed the logic behind the SLE and CLE subcohort definitions.

- **Lupus erythematosus (LE)**: N = 10,352 (63.5% of possible cases)
- **Systemic lupus erythematosus (SLE):** N = 7279 (44.6%)
- **Cutaneous lupus erythematosus (CLE):** N = 4643 (28.5%)

Because the three algorithms were applied independently and optimized for different case definitions, patients could be assigned to more than one cohort. As a result, the broader LE group is not equivalent to the sum of the SLE and CLE cohorts.

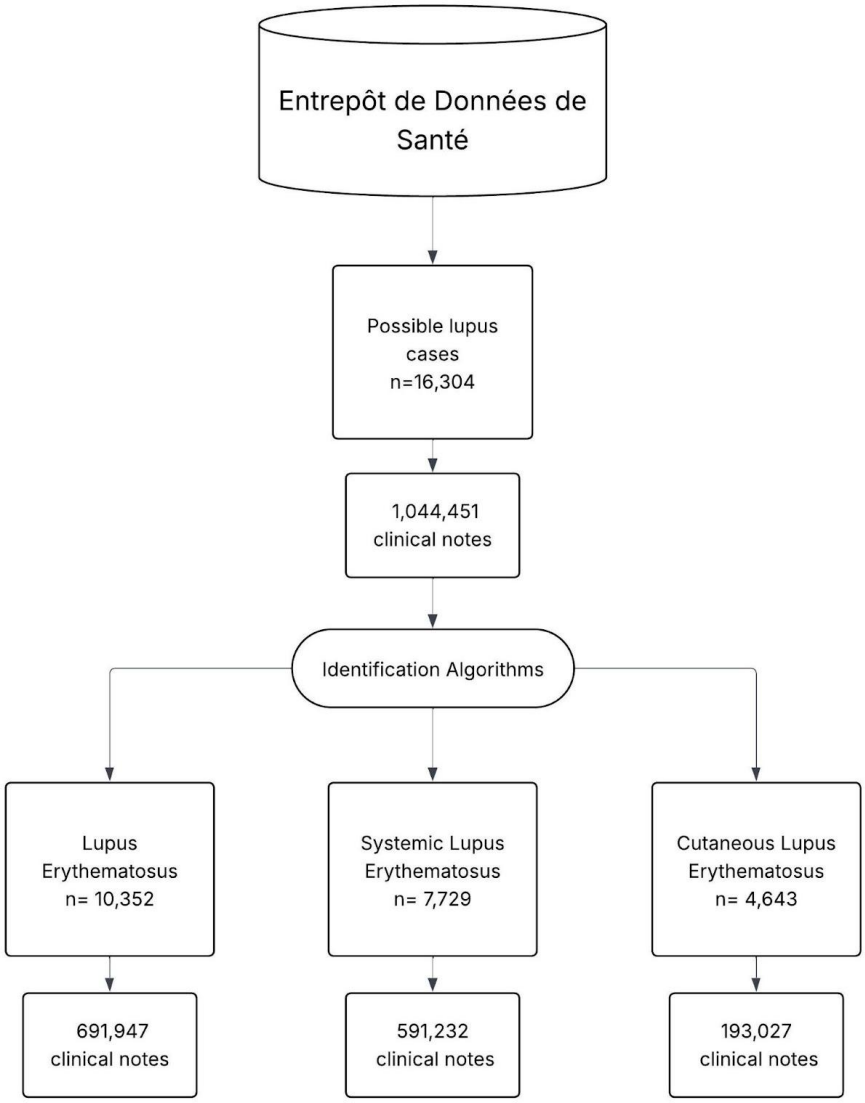

Figure 1. Identification of lupus subcohorts from the AP-HP Clinical Data Warehouse (Entrepôt de Données de Santé). A total of 16,304 possible lupus cases were identified and classified into three predicted subcohorts—lupus erythematosus (LE), systemic lupus erythematosus (SLE), and cutaneous lupus erythematosus (CLE)—using previously validated identification algorithms. The number of clinical notes associated with each subcohort is also shown.

These three subcohorts together represent over 1.8 million clinical notes: more than one million notes in the LE cohort (n = 1,044,451), 591,232 in SLE, and 193,027 in CLE, illustrating the scale and richness of the unstructured data used in this analysis.

2.3 Clinical Test Annotation for NLP Development

Manual annotation is the process of labeling specific pieces of information in clinical text so that they can be used to train, develop, or evaluate an automated system. In this study, annotation involved reading clinical reports and identifying phrases that referred to relevant clinical concepts.

We began with an exploratory review of over one hundred clinical reports drawn from patients identified as possible lupus cases. This initial review helped us understand how clinical events are typically described in free text, including variations in phrasing, context, and documentation style. Based on these insights, a custom annotation schema was developed to identify and categorize mentions of HCQ exposure, ophthalmology follow-up, maculopathy status, and HCQ treatment adjustments. The schema was developed in close collaboration with supervising clinicians and data scientists to ensure that it was both clinically meaningful and feasible for rule-based NLP extraction. The complete schema is presented in Table A1 in the Appendix.

To create a focused training set, we first filtered the clinical notes to retain only those likely to contain relevant clinical information. Note types such as anesthesiology reports, administrative documents, radiology reports, and scanned files were excluded based on metadata, as they were unlikely to mention HCQ exposure, ophthalmology follow-up, or maculopathy findings. From the remaining pool, we selected a subset of notes from patients with at least one mention of HCQ, identified using regular expression filters applied to unstructured text. This filtering step was necessary due to the large volume of available EHR data and the high proportion of notes unrelated to our target concepts. Apache Spark was used to efficiently process the data at scale. A final sample of 377 relevant clinical reports was extracted and stored securely within the AP-HP research environment for manual annotation.

Annotations were carried out using the Metanno interface on the sample of 377 clinical reports (10). The annotated sample served as the foundation for designing the rule-based NLP pipeline. Reviewing these clinical notes allowed for the identification of key terms, contextual patterns, and qualifiers necessary to capture relevant clinical concepts. The manual annotations made it possible to observe how maculopathy, HCQ exposure, treatment actions, and ophthalmology follow-up were typically expressed in free text. These insights directly informed the creation of regular expression patterns, the structure of contextual windows, and the prioritization rules used in the automated extraction pipeline.

2.4 NLP Pipeline Development

A natural language processing (NLP) pipeline is a sequence of automated steps used to process unstructured text and extract structured information. In the context of electronic health records (EHRs), NLP pipelines identify clinical concepts, assign relevant attributes, and handle contextual nuances such as negation or temporality. Rule-based pipelines, as opposed to machine learning approaches, rely on manually defined patterns such as regular expressions. They are particularly suited for applications where interpretability and transparency are important and they require significantly fewer computational resources.

For this project, a rule-based NLP pipeline was developed using the edsnlp library to extract relevant clinical information from unstructured clinical notes (11). The pipeline targeted mentions of four main clinical entities: hydroxychloroquine (HCQ), maculopathy, ophthalmology follow-up, and HCQ treatment changes, along with their associated attributes and qualifiers. It consisted of flat term matchers for entity recognition (e.g. detecting terms like *hydroxychloroquine* or *ophthalmology consultation*) and contextual matchers for attribute assignment (e.g., labeling a maculopathy mention as *clinic* or *preclinic* based on surrounding words). Modules were also implemented to detect negation ("*no signs of maculopathy*"), historical references ("*ophthalmology consultation in the past*"), family history ("*father had lupus*"), and speculative language ("*possible maculopathy*"). These qualifier modules had been previously developed by data scientists at the EDS and were integrated into the pipeline to enrich contextual understanding of the extracted entities. All pattern definitions were constructed using regular expressions informed by the manually annotated corpus and domain knowledge.

Table 1. Entities and Associated Attributes Extracted by the NLP Pipeline

Entity	Definition	Associated Attribute
HCQ	Mentions of hydroxychloroquine or Plaquenil use	**Contraindication**: HCQ cannot be used due to a medical reason **No Contraindication**: HCQ is not contraindicated or explicitly safe to continue **Treatment Discontinuation:** HCQ treatment is stopped
Ophthalmology Visit	Mentions of eye exams, ophthalmology visits, or retinal imaging related to HCQ monitoring.	**Follow-up Done**: Follow-up already completed **Follow-up Scheduled**: Appointment planned **Follow-up Prescribed**: Follow-up recommended but not yet scheduled
HCQ-induced Maculopathy	References to maculopathy, retinal toxicity, vision loss, or related diagnoses associated with HCQ	**Preclinic**: Early or suspected signs of toxicity **Clinic**: Confirmed maculopathy diagnosis **Unknown**: Maculopathy mentioned without clear stage
Global Qualifiers	Modules used to refine interpretation of mentions across entities	**Negation**: clinical concept was explicitly ruled out **History**: mention refers to a past event rather than a current one **Hypothesis**: uncertainty or a hypothetical mention of the concept **Family**: condition applies to a family member rather than the patient

Pipeline components were implemented in modular steps. Initial development focused on flat matchers to identify key terms, followed by the addition of contextual windows to extract finer-grained attribute information (e.g., maculopathy status: clinic, preclinic, unknown). Pipeline outputs were repeatedly compared to the manually annotated training set to identify false positives and false negatives. This comparison guided refinements to entity boundaries, attribute labeling rules, and context window sizes.

Pipeline outputs were repeatedly compared to the manually annotated training set to identify false positives (incorrect extractions) and false negatives (missed mentions). Each comparison round involved reviewing discrepancies between the pipeline predictions and the annotations to determine whether errors were due to overly broad patterns, missing edge cases, or inconsistent phrasing in the clinical text. These reviews directly informed refinements to the pipeline, including narrowing or expanding regular expressions, redefining entity boundaries, adjusting the structure and size of contextual windows, and improving the logic used to assign attributes. Particular attention was paid to cases where overlapping qualifiers or ambiguous phrasing caused misclassification, as these often pointed to areas where the rule logic or prioritization hierarchy required adjustment.

When discrepancies emerged, such as missed cases due to unexpected phrasing or incorrectly labeled mentions caused by overlapping qualifiers, the regular expressions and logic for attribute prioritization were updated. For instance, ophthalmology follow-up attributes were assigned a hierarchy of importance (scheduled > done > prescribed), and only the highest-priority value was retained per instance. Negated mentions of maculopathy were explicitly labeled as “no maculopathy” to prevent misclassification. A sensitivity analysis was also conducted to assess the impact of context window size on prediction quality, particularly in cases where the difference between predicted and annotated spans reflected boundary mismatches rather than content disagreement.

The iterative refinement process improved the pipeline’s precision and also highlighted minor inconsistencies in the training annotations, which were corrected as needed. All regular expression patterns and matcher configurations were version-controlled and organized in modular scripts within a GitLab repository to ensure reproducibility and scalability.

2.5 NLP algorithm evaluation

To assess the performance of the NLP pipeline, a second set of manual annotations was created to serve as a gold-standard reference. For this validation set, a new sample of clinical notes was drawn from the broader cohort of patients identified as possible lupus cases. Unlike the training set, the validation sample was not restricted to patients with documented HCQ mentions. This approach was taken to reduce selection bias and to better reflect the heterogeneity of real-world clinical documentation. To avoid overlap between the training and validation sets, all patients included in the initial annotation phase were excluded from the sampling pool for the gold-standard set.

As with the training data, only note types likely to contain relevant clinical information were retained. Irrelevant documents—such as anesthesiology reports, administrative notes, radiology reports, and scanned documents—were excluded based on metadata. After this filtering step, a sample of 600 clinical notes was manually reviewed and annotated using the annotation schema.

Of the 600 notes reviewed, only 167 contained at least one annotated mention of a target entity. This is expected in the context of real-world EHR data, where individual patients often have a large volume of clinical documentation, much of which is unrelated to the specific concepts being targeted. Consequently, a substantial portion of the sampled notes lacked relevant content and did not contribute to the performance evaluation.

The rule-based pipeline was applied to this same set of 600 clinical notes, and its predicted outputs were directly compared to the gold-standard reference. Predicted outputs from the pipeline were compared directly to these annotations to assess extraction performance. Evaluation metrics included precision, recall, and F1 score, which are standard in information extraction tasks. In this context, recall (analogous to sensitivity) represents the proportion of relevant mentions correctly identified by the pipeline, while precision reflects the proportion of identified mentions that were actually correct. The F1 score, calculated as the harmonic mean of precision and recall, provides a balanced measure of overall performance. Specificity was not calculated, as doing so would require defining a comprehensive set of true negative spans, phrases that could have been extracted but were not, which is ill-defined in entity recognition tasks and could produce misleading performance estimates.

In addition to evaluating all ophthalmology follow-up mentions extracted by the pipeline (referred to as OPH), a more targeted subset called OPH2 was created to improve clinical relevance. OPH2

included only ophthalmology follow-up mentions that appeared in the same note as a reference to either HCQ or maculopathy. This filtering step was introduced because ophthalmology can be mentioned in a wide range of clinical contexts unrelated to HCQ use—for example, in follow-up for cataracts, glaucoma, or routine vision exams. Including these unrelated mentions would introduce noise into the evaluation and inflate the number of false positives for the use case of interest. By focusing only on ophthalmology follow-up mentions that co-occur with HCQ or maculopathy, the OPH2 subset allows for a more accurate assessment of the pipeline's performance in detecting clinically meaningful monitoring related to HCQ-induced maculopathy.

2.6 Analysis methods

After validation, the NLP pipeline was applied to the full set of clinical notes from the three predicted lupus cohorts, lupus erythematosus (LE), systemic lupus erythematosus (SLE), and cutaneous lupus erythematosus (CLE), as defined by the previously validated identification algorithms. These subcohorts were selected for descriptive analysis because they represent more clinically specific populations.

Descriptive statistics were first calculated across the three subcohorts (LE, SLE, CLE) to quantify the frequency of clinically relevant events extracted via NLP. Among HCQ-exposed patients, we reported the proportion with at least one mention of ophthalmology follow-up, maculopathy, or HCQ treatment discontinuation. A “positive maculopathy” case was defined as a patient with at least one non-negated maculopathy mention, excluding those explicitly labeled as “no maculopathy” by the NLP pipeline. This definition was used as the primary operationalization of prevalence throughout the analysis. To assess how prevalence estimates varied under different thresholds for documentation frequency, we also reported the number of maculopathy-positive patients with only 1, 2–4, 5–9, and ≥10 mentions. This served to examine the robustness of prevalence estimates under alternate definitions based on documentation intensity. In addition, we summarized the proportion of patients with both positive and negated maculopathy mentions, which may reflect ambiguity or evolving diagnostic impressions over time.

We further examined maculopathy prevalence by subtype, as determined by contextual qualifiers extracted by the NLP pipeline. Patients whose maculopathy mentions lacked any assigned subtype were grouped into an additional “unspecified” category to preserve these positive cases in the analysis. For each subtype, we calculated the proportion of patients with at least one ophthalmology follow-up mention and the proportion with a treatment discontinuation mention.

This allowed us to explore whether documentation of specific toxicity profiles was associated with clinical management.

Because multiple clinical notes may exist for each patient, all NLP-derived annotations were aggregated at the patient level. Binary outcomes (e.g., presence of ≥1 maculopathy mention) and count outcomes (e.g., number of maculopathy mentions per patient) were computed accordingly. All analyses were restricted to HCQ-exposed patients to maintain a consistent denominator population focused on individuals at risk for HCQ-induced maculopathy.

2.7 Comparative Analysis of SLE and CLE Cohorts

The objective of the comparative analysis was to assess whether clinical documentation patterns, particularly around safety monitoring and treatment changes, differed between the SLE and CLE cohorts, in a way that could reflect differences in clinical complexity, care practices, or monitoring intensity.

To ensure independence between groups, patients who appeared in both cohorts were assigned to the SLE group. These overlapping individuals were excluded from the CLE cohort, resulting in a filtered CLE population used for all inferential analyses. This decision was based on the assumption that systemic lupus erythematosus represents a more clinically comprehensive condition; retaining overlapping patients in both groups would have violated the assumption of independence and potentially biased estimates.

Two types of statistical comparisons were conducted between the SLE and CLE cohorts:

- **Prevalence comparisons:** Chi-squared tests were used to compare the prevalence of each clinical concept (maculopathy, ophthalmology follow-up, and HCQ treatment discontinuation), defined as the proportion of HCQ-exposed patients with at least one corresponding mention in their clinical notes.

- **Mention frequency comparisons:** Mann-Whitney U tests were used to assess differences in the distribution of documentation frequencies per patient between the SLE and CLE cohorts for each clinical concept. Rank-biserial correlation coefficients were calculated to estimate effect sizes, with conventional thresholds applied for interpretation ((0.01 = small, 0.03 = medium, 0.5 = large). Statistical significance was set at $p < 0.05$.

2.8 Hypothetical Logistic Regression Analysis

To illustrate how structured and unstructured EHR data could support risk modeling, we designed a hypothetical logistic regression to estimate the likelihood of documented maculopathy among hydroxychloroquine (HCQ)-exposed patients. The dependent variable was a binary indicator of maculopathy documentation, defined as having at least one non-negated mention of maculopathy in the clinical notes.

The model includes several independent variables selected based on known risk factors for HCQ-induced retinal toxicity. While some of these variables are conceptually relevant, some are not currently extractable from the AP-HP EDS in its present form. As such, this model remains theoretical and is intended to outline a potential future direction for HER-based surveillance research:

- Length of HCQ treatment (years): a key predictor of toxicity, but not currently available due to lack of reliable treatment start dates.
- Average daily HCQ dose (mg/kg/day): a major risk factor, but cannot be computed without structured dosage data.
- Age at first HCQ mention: could be calculated and included as a continuous variable.
- Cohort assignment (SLE vs CLE): derivable from existing identification algorithms.
- Sex (male vs female): available as a structured demographic variable.

2.8.1 Logistical Regression Equation

Let Y1 if a patient has at least one positive (non-negated) maculopathy mention, and Y=0 otherwise. The model would take the form:

$$\log\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta_0 + \beta_1 \cdot \text{TreatmentLength} + \beta_2 \cdot \text{Dosage} + \beta_3 \cdot \text{Age} + \beta_4 \cdot \text{SLE} + \beta_5 \cdot \text{Sex}$$

This conceptual model is not intended to estimate biological risk of disease, but rather the likelihood of documentation of maculopathy within the EDS. It highlights how future improvements in data completeness could support more nuanced analyses of care processes, treatment exposure, and clinical documentation patterns.

3. RESULTS

3.1 NLP Pipeline Performance

The NLP pipeline was evaluated on the gold-standard validation set consisting of 600 clinical notes, of which 167 contained at least one annotated entity. Performance was assessed for both entity recognition and attribute extraction tasks.

Table 2. Entity and Attribute-Level Performance of the Rule-Based NLP Algorithm

Entity Label	Predicted Mentions	Gold-Standard Mentions	True Positives	False Positives	False Negatives	Precision	Recall	F1 score
HCQ	470	473	470	0	3	1	0.994	0.997
OPH	95	88	76	19	12	0.8	0.864	0.831
OPH2	77	88	76	1	12	0.987	0.864	0.921
MACULOPATHY	11	12	11	0	1	1	0.917	0.957
Attribute								
HCQ								
Contraindication (CI)	2	2	2	0	0	1	1	1
No Contraindication (NO CI)	4	6	4	0	2	1	0.667	0.8
Treatment Discontinuation	33	39	30	3	9	0.909	0.769	0.833
Ophthalmology								
Follow-up Done	18	41	15	3	26	0.833	0.3366	0.508
Follow-up Scheduled	9	17	9	0	8	1	0.529	0.692
Follow-up Prescribed	2	15	2	0	13	1	0.133	0.235
Maculopathy								
Clinic	6	7	5	1	2	0.833	0.714	0.769
Pre Clinic	1	2	1	0	1	1	0.5	0.667
Unknown	0	1	0	0	1	0	0	0
No Maculopathy	2	2	2	0	0	1	1	1

Core entities were reliably identified, particularly hydroxychloroquine (HCQ) and maculopathy. HCQ mentions were extracted with perfect precision (1.000) and near-perfect recall (0.994), resulting in an F1 score of 0.997. Maculopathy detection also performed well, with an F1 score of 0.957 (precision = 1.000, recall = 0.917). These results reflect the pipeline's ability to identify not only the presence of documented maculopathy but also its explicit absence. The attribute "no maculopathy" was detected with perfect precision and recall (F1 = 1.000), reinforcing the system's capacity to distinguish between confirmed cases and statements that rule out maculopathy.

Ophthalmology-related entities showed more variable performance. For the unfiltered entity (OPH), which included all mentions of ophthalmology regardless of context, the pipeline achieved a precision of 0.800 and recall of 0.864 (F1 = 0.831). However, a relatively high number of false positives (n = 19) likely reflects incidental mentions unrelated to HCQ monitoring. As described in the Methods, a filtered version of this entity (OPH2) was created by restricting evaluation to ophthalmology mentions that co-occurred with either HCQ or maculopathy. This contextual filtering significantly improved performance, with precision increasing to 0.987 and the F1 score

rising to 0.921, while recall remained stable. These results support the value of context-based constraints for improving entity specificity in EHR text.

Attribute-level performance varied by concept. Ophthalmology follow-up attributes were generally extracted with high precision but more modest recall, reflecting the diverse ways such follow-up is documented in clinical notes. “Follow-up scheduled” achieved strong performance (F1 = 0.692), followed by “follow-up done” (F1 = 0.508). “Follow-up prescribed” was the most challenging, with an F1 score of only 0.235 due to very low recall. These results suggest that while the rule-based approach can reliably extract well-formatted mentions, it may miss more varied or ambiguous phrasings.

Attributes based on more formulaic or commonly documented concepts performed better. The “stop treatment” label showed high reliability (F1 = 0.833), and both “CI” (contraindication to HCQ) and “NO CI” were extracted with perfect or near-perfect accuracy (F1 = 1.000 and 0.800, respectively). Similarly, maculopathy subtype labels such as “clinic” and “preclinic” were identified with reasonably high accuracy, achieving F1 scores of 0.769 and 0.667, respectively. These subtypes were based on contextual patterns surrounding maculopathy mentions, such as clinical confirmation or early indicators of toxicity. Performance was lower or unreported for less frequent categories. For example, the “unknown” subtype had no correct predictions, resulting in an F1 score of 0. This was due to the extremely limited number of annotated instances in the gold-standard set (n = 1), which limited the pipeline's ability to validate matching patterns. Conversely, “no maculopathy,” which reflects explicit negation, was identified with perfect precision and recall, consistent with the use of standardized negation terms in clinical text.

Overall, the pipeline demonstrated high accuracy for the core clinical entities relevant to this study, with acceptable performance for most associated attributes. Lower-performing categories indicate areas for potential improvement in future iterations, particularly around less consistently documented follow-up actions.

3.2 NLP output results

3.2.1 Cohort characteristics and NLP coverage

The NLP pipeline was applied to clinical notes from patients in three predicted lupus cohorts: lupus erythematosus (LE), systemic lupus erythematosus (SLE), and cutaneous lupus erythematosus (CLE). The total number of patients in each cohort was as follows: LE = 10,352; SLE = 7,279; CLE = 4,643.

All ophthalmology-related results reported in this section and throughout the remainder of the analysis are based on the filtered entity set (OPH2), which includes only ophthalmology mentions from clinical notes that also reference either hydroxychloroquine (HCQ) or maculopathy, as defined in the Methods. Using this filtered approach, the proportion of patients with at least one ophthalmology mention was 71.97% in the LE cohort, 76.22% in SLE, and 62.87% in CLE. All subsequent ophthalmology analyses are restricted to this OPH2-defined cohort.

All downstream analyses were restricted to patients with documented hydroxychloroquine (HCQ) exposure, as identified by the NLP pipeline. This ensured a consistent denominator population focused on individuals at risk for retinal toxicity and relevant to the clinical outcomes assessed.

3.2.2 Descriptive Results: HCQ-Exposed Patients

Table 3. Clinical event frequencies among HCQ-exposed patients across cohorts

	Cohort					
	LE		SLE		CLE	
Description of variable	Count	%	Count	%	Count	%
Total patients in cohort	10352		7279		4643	
Total patients with at least one clinical note processed by NLP pipeline	10179		7127		4320	
HCQ-exposed patients	10177	98.31%	7125	97.88%	4315	92.94%
ophthalmology follow-up mentions						
1+	7324	71.97%	5466	76.22%	2713	62.87%
2+	6020	59.15%	4681	65.70%	2056	47.65%
HCQ stop-treatment mentions						
1+	4294	42.20%	3097	43.47%	1748	40.51%
2+	2968	29.16%	2200	30.88%	1156	26.79%
HCQ-exposed patients with positive maculopathy mentions	1002	9.85%	860	12.07%	255	5.91%
ophthalmology follow-up mentions						
0	99	9.88%	74	8.60%	40	15.69%
1+	903	90.12%	786	91.40%	215	84.31%
2+	835	83.33%	733	85.23%	194	76.08%
HCQ stop-treatment mentions						
1+	675	67.37%	587	68.26%	158	61.96%
2+	549	54.79%	482	56.05%	127	49.80%

Among HCQ-exposed patients, documentation patterns varied across cohorts in both frequency and intensity of clinical concepts. The proportion of patients with at least one mention of each outcome is summarized in Table 3.

- Positive maculopathy mentions — defined as at least one non-negated maculopathy mention — were identified in 9.85% of HCQ-exposed patients in the LE cohort, 12.07% in SLE, and 5.91% in CLE.

- Ophthalmology follow-up was documented in 71.97% of LE patients, 76.22% of SLE patients, and 62.87% of CLE patients. Notably, a majority of patients had more than one follow-up mention, particularly in LE (59.15%) and SLE (65.70%).

- HCQ treatment discontinuation was identified in 42.20% of LE patients, 43.47% of SLE patients, and 40.51% of CLE patients. Among these, 29–31% had documentation of discontinuation in at least two separate instances, suggesting repeated or reaffirmed treatment decisions over time.

Among patients with positive maculopathy mentions, follow-up and treatment adjustments were also frequent:

- Over 90% of maculopathy patients in LE and SLE had at least one documented ophthalmology follow-up, compared to 84.31% in CLE.

- 83–85% of maculopathy patients in LE and SLE had more than one follow-up mention, compared to 76% in CLE.

- Stop-treatment mentions were documented in 67.37% (LE), 68.26% (SLE), and 61.96% (CLE) of maculopathy-positive patients. Among them, over half had discontinuation documented more than once.

These patterns suggest that while overall documentation intensity was higher in the SLE and LE groups, CLE patients may have received less consistent follow-up or were less frequently flagged for treatment changes in clinical notes. The repeated documentation of follow-up and discontinuation in a significant proportion of cases also suggests these were not isolated mentions but rather part of ongoing care trajectories.

3.2.3 Maculopathy patterns

Table 4. Maculopathy mention patterns among HCQ-exposed patients

	Cohort					
	LE		SLE		CLE	
	Count	%	Count	%	Count	%
Average maculopathy mentions per patient	5.58		6.05		3.78	
Patients with maculopathy mention count of…						
only 1	430	42.91%	362	42.09%	118	46.27%
2-4	302	30.14%	255	29.65%	82	32.16%
5-9	136	13.57%	120	13.95%	30	11.76%
10+	134	13.37%	123	14.30%	25	9.80%
Patients with both "no maculopathy" and maculopathy	248	24.75%	213	24.77%	63	24.71%

Note: All counts and percentages in this table are based on hydroxychloroquine-exposed patients with at least one non-negated mention of maculopathy.

Patterns of maculopathy documentation differed not just in how often the condition was mentioned, but in how extensively it was recorded for individual patients. As shown in Table 4, patients in the SLE and LE cohorts had a higher average number of maculopathy mentions per person (6.05 and 5.58, respectively) compared to those in the CLE cohort (3.78). Frequent documentation was more common in SLE and LE: approximately 27.3% of maculopathy-positive patients in these cohorts had five or more mentions (13.95% + 14.30% in SLE; 13.57% + 13.37% in LE), compared to only 21.56% in CLE (11.76% + 9.80%). Fewer than 10% of CLE patients exceeded ten mentions, suggesting lower intensity of maculopathy documentation in this group.

Ambiguity in documentation was also observed. Approximately one in four patients with a positive maculopathy mention also had a documented negation (“no maculopathy”) elsewhere in their records, a pattern consistent across all cohorts (LE: 24.75%, SLE: 24.77%, CLE: 24.71%). This may reflect evolving clinical trajectories, where earlier notes ruled out toxicity, but later documentation indicated a confirmed or suspected case. Such transitions highlight the importance of considering temporality when interpreting conflicting mentions in longitudinal EHR data.

Table 5. Maculopathy subtypes and clinical follow-up outcomes

	Patient count	% HCQ-exposed patients	% with ≥1 mention of ophtalmology follow-up	% with ≥1 mention of stopping HCQ treatment
LE				
clinic	578	5.68%	88.93%	75.09%
preclinic	144	1.41%	97.22%	48.61%
unknown	9	0.09%	88.89%	88.89%
unspecified	271	2.66%	88.93%	60.15%
total	1002	9.85%	90.12%	67.37%
SLE				
clinic	493	6.92%	90.37%	75.86%
preclinic	119	1.67%	97.48%	51.26%
unknown	9	0.13%	88.89%	88.89%
unspecified	239	3.35%	90.38%	60.25%
total	860	12.07%	91.40%	68.26%
CLE				
clinic	147	3.41%	83.67%	69.39%
preclinic	39	0.90%	94.87%	35.90%
unknown	0	0.00%	0.00%	0.00%
unspecified	69	1.60%	79.71%	60.87%
total	255	5.91%	84.31%	61.96%

Note: All counts and percentages in this table are based on hydroxychloroquine-exposed patients with at least one non-negated mention of maculopathy.

Maculopathy mentions were further stratified by subtype to assess whether different patterns of documented toxicity were associated with clinical follow-up or treatment decisions in Table 5. Across all cohorts, the “clinic” subtype, representing confirmed maculopathy, was the most frequent, accounting for 5.68% of HCQ-exposed patients in LE, 6.92% in SLE, and 3.41% in CLE. The “preclinic” subtype, reflecting early signs or suspicion of toxicity, was less common but still notable in LE (1.41%) and SLE (1.67%), with fewer cases observed in CLE (0.90%). The “unknown” subtype was rarely assigned and absent entirely in CLE.

Follow-up documentation was consistently high across subtypes. Among patients with “clinic” maculopathy, over 88% had ophthalmology follow-up in all cohorts. For “preclinic” cases, follow-up rates exceeded 94% across the board, suggesting that even subtle or suspected toxicity prompted monitoring. However, treatment discontinuation was more variable. While approximately 75% of “clinic” cases in LE and SLE had documented HCQ discontinuation, this dropped to 48–51% among “preclinic” patients, and to 60–61% in the “unspecified” category. In CLE, discontinuation was less frequently documented across all subtypes, particularly in “preclinic” cases (35.90%).

3.3 Comparison between SLE and CLE cohorts

3.3.1 Comparative Prevalence (Chi² tests)

To evaluate differences in outcome prevalence between the SLE and filtered CLE cohorts, Chi-squared tests were conducted for three binary outcomes: maculopathy mentions, ophthalmology follow-up, and HCQ treatment discontinuation.

Table 6. Comparative Prevalence between SLE and CLE cohorts

Outcome	Chi² Statistic	p-value
Maculopathy mentions	142.27	<0.0001
Ophthalmology Follow-up	341.4	<0.0001
HCQ Treatment Discontinuation	23.62	<0.0001

All three comparisons yielded statistically significant differences ($p < 0.0001$). As shown in Table 6, the prevalence of maculopathy mentions was significantly higher in the SLE cohort than in CLE (12.07% vs 5.91%, $p < 0.0001$). Similarly, ophthalmology follow-up was more frequently documented in SLE (76.22%) than in CLE (62.87%, $p < 0.0001$). HCQ treatment discontinuation was also more common in SLE (43.47%) compared to CLE (40.51%), although the difference was less pronounced ($p < 0.0001$). These results confirm that documentation of adverse event monitoring and treatment changes differs significantly between the two cohorts.

3.3.2 Comparative Intensity of Documentation (Mann-Whitney U Tests)

To evaluate differences in the intensity of documentation between the SLE and filtered CLE cohorts, Mann-Whitney U tests were conducted on the per-patient frequency of mentions for maculopathy, ophthalmology follow-up, and HCQ treatment discontinuation. All three comparisons yielded statistically significant results ($p < 0.0001$), indicating that the distribution of mention counts differed between cohorts.

Table 7. Comparative Intensity of Documentation Between SLE and CLE Cohorts (Mann-Whitney U Test Results)

Outcome	U-statistic	p-value	Rank-Biserial Correlation	Effect Size
Maculopathy Mentions	13383315.5	<0.0001	-0.0737	very small
Ophthalmology Follow-up Mentions	16105377.5	<0.0001	-0.292	small
Treatment Discontinuation Mentions	13346896.5	<0.0001	-0.0707	very small

Note: All counts and percentages in this table are based on hydroxychloroquine-exposed patients. Maculopathy mentions are based on non-negated mentions of maculopathy.

Effect sizes, measured using rank-biserial correlation, revealed variation in the magnitude of these differences. Ophthalmology follow-up had the largest effect (r = –0.292), interpreted as small. Maculopathy mentions showed a very small effect (r = –0.0737), as did treatment discontinuation (r = –0.0707). These results suggest that while SLE patients had more frequent documentation across all three outcomes, the differences were modest in size, particularly for maculopathy and treatment discontinuation.

4. DISCUSSION

4.1 Key Results

This study aimed to develop and validate a rule-based NLP pipeline to detect HCQ exposure, maculopathy, ophthalmology follow-up, and treatment discontinuation from unstructured EHR notes, with a specific focus on patients with systemic or cutaneous lupus erythematosus.

The pipeline demonstrated high overall performance, with near-perfect precision and recall for detecting HCQ and maculopathy mentions, and strong performance for clinically meaningful attributes such as treatment discontinuation and confirmed toxicity subtypes. Ophthalmology mentions benefited significantly from contextual filtering, which reduced false positives and improved overall accuracy.

Applying the pipeline at scale to over 10,000 predicted lupus patients revealed that documentation of HCQ monitoring and toxicity was highly variable across cohorts. Patients in the SLE and broader LE groups had higher rates of maculopathy documentation, ophthalmology follow-up, and treatment discontinuation compared to those in the CLE cohort. These differences were evident in both prevalence and intensity of documentation and were statistically significant across all outcomes examined. Stratified analysis further showed that confirmed maculopathy (“clinic” subtype) was more likely to be associated with treatment discontinuation than early or ambiguous findings.

Together, these findings support the feasibility and utility of rule-based NLP approaches for extracting clinically relevant safety data from unstructured text, enabling large-scale evaluation of real-world monitoring practices in chronic disease populations.

4.2 Comparison to Existing Literature

The findings of this study align with and extend existing evidence on the risk and management of hydroxychloroquine (HCQ)-induced retinopathy. Prior large-scale cohort studies have demonstrated that the risk of retinal toxicity increases substantially with both cumulative exposure and higher daily dosing. In a recent cohort study by Melles et al., the prevalence of confirmed retinal toxicity was estimated at 8.6%, with cumulative incidence reaching over 21% after 15 years in patients receiving more than 6 kg/kg/day (12).

While not directly comparable, the prevalence of NLP-detected non-negated maculopathy mentions in this study's SLE cohort (12.1%) falls within this general range, suggesting that real-world clinical documentation may reflect meaningful patterns of suspected toxicity. However, these findings should be interpreted with caution. The prevalence estimates in our study are based on textual mentions, not imaging-confirmed diagnoses, and may be influenced by differences in population characteristics, treatment duration, and screening practices.

Earlier work by Melles and Marmor (2014) similarly emphasized the role of dose and duration, reporting a 7.5% prevalence of retinal toxicity after five years of treatment, with risks increasing significantly after 10 years and in those exceeding 5.0 mg/kg/day (5). In our study, maculopathy documentation was more frequent in the SLE cohort than in CLE. While dosage and treatment duration could not be directly measured, the difference may reflect variations in treatment profiles between the two groups, particularly in terms of duration and intensity of HCQ use. Other contributing factors, such as differences in age or differing comorbidity profiles (such as higher rates of renal impairment in SLE), may also play a role and will be explored in the next phase of this project.

Although our analysis, based on NLP-detected maculopathy mentions, cannot directly measure imaging-confirmed prevalence, the observed proportions (9.9% in LE, 12.1% in SLE, 5.9% in CLE) fall within the expected range reported in the literature. This reinforces that real-world clinical documentation captures clinically meaningful signals of concern, even when structured screening data are unavailable.

Subtypes of maculopathy captured in this study, particularly the "preclinic" category, also mirror findings from the prospective study by Jaumouillé et al. (2015), which reported that 3.8% of patients showed early retinal changes consistent with preclinical toxicity (13). In our data, preclinic cases were frequently followed up by ophthalmology but less often associated with HCQ

discontinuation, suggesting that these findings are often monitored without prompting immediate changes to treatment. This pattern reflects a degree of clinical uncertainty in managing early-stage toxicity, consistent with the broader literature.

Together, these comparisons suggest that real-world EHR documentation, when systematically analyzed using NLP, can reveal meaningful patterns of toxicity monitoring and concern that are concordant with evidence from structured screening studies.

4.4 Interpretation of Clinical Documentation Patterns

The observed differences in maculopathy documentation between SLE and CLE cohorts may reflect underlying variations in disease severity and monitoring intensity. SLE, as a systemic condition with greater clinical complexity, is more likely to prompt regular follow-up and safety surveillance, which may explain the higher prevalence and frequency of maculopathy, ophthalmology, and discontinuation mentions in that group.

In contrast, CLE patients may have less consistent hospital follow-up, possibly due to fewer systemic complications. Some CLE patients may also be monitored in outpatient or private settings, leading to under documentation in hospital EHRs despite ongoing care.

The relatively low rate of HCQ discontinuation among CLE patients with “preclinic” maculopathy further suggests a degree of clinical caution in altering treatment based on early or uncertain signs of toxicity. Still, the fact that most preclinic cases had ophthalmology follow-ups mentioned in their notes indicates that clinicians are responsive to early warnings, even in the absence of confirmed damage. This suggests that clinicians often respond to suspected toxicity with increased monitoring, even when formal confirmation is lacking.

4.5 Implications for Public Health and Clinical Practice

This study demonstrates the potential for rule-based NLP methods to support large-scale pharmacovigilance and adherence monitoring in real-world clinical settings. By systematically extracting information from unstructured clinical notes, NLP pipelines can detect critical safety indicators, such as early signs of toxicity, missed follow-up, or treatment discontinuation, that are often absent from structured EHR fields. Applied across thousands of patients, these tools can help identify care gaps, monitor compliance with clinical guidelines, and surface under-documented risks, all without the need for manual chart review.

As EHR systems continue to expand, integrating NLP outputs into quality improvement initiatives or clinical decision support tools could enhance proactive care. For instance, flagging patients on long-term HCQ therapy without recent ophthalmologic follow-up could prompt timely interventions, reducing the risk of preventable vision loss. Similarly, aggregating data on treatment discontinuation and suspected adverse effects could inform institutional prescribing policies and surveillance protocols. By bridging the gap between narrative documentation and structured analytics, NLP enables scalable, data-driven approaches to chronic disease management and safety monitoring.

This study represents, to our knowledge, the largest real-world cohort of SLE and CLE patients exposed to HCQ for which unstructured clinical data have been analyzed. Among 10,177 HCQ-exposed patients identified through validated algorithms, 1,002 had at least one positive mention of maculopathy. This cohort size exceeds that of any individual study included in a 2023 meta-analysis of HCQ retinopathy, where the largest sample comprised 2,361 HCQ-exposed patients (14). Moreover, large-scale observational studies focusing specifically on CLE remain extremely rare. This is also the first study to compare clinical documentation patterns between CLE and SLE patients at scale using EHR data.

4.6 Strengths

This study represents the first application of a rule-based NLP pipeline for detecting hydroxychloroquine-related maculopathy within a French EHR system. A central strength lies in the development and validation of a high-performing, clinically informed pipeline capable of extracting meaningful indicators of HCQ exposure, toxicity, and monitoring. The pipeline achieved excellent precision and recall for key entities such as HCQ and maculopathy, while contextual filtering (OPH2) substantially improved performance for ophthalmology-related extractions.

In applying the pipeline at scale, the study leveraged over 10,000 lupus patients across distinct subcohorts (LE, SLE, CLE), enabling nuanced comparisons of real-world care practices. The analysis captured not only the presence of maculopathy and related follow-up but also the frequency and recurrence of these mentions, offering insight into the intensity of clinical monitoring. Stratification by maculopathy subtype further allowed differentiation between confirmed and early-stage toxicity, revealing variation in follow-up and treatment patterns. These design choices positioned the study to evaluate both documentation quality and care delivery at scale.

Although conducted within a single health system (AP-HP), the study leveraged data from 39 university hospitals across the Greater Paris area. This multicentric design increases the linguistic and clinical heterogeneity of the EHR corpus, enhancing the generalizability and robustness of the NLP pipeline across diverse documentation styles and care settings.

4.7 Limitations

This study has several limitations. Although the pipeline benefited from a multicentric dataset within AP-HP, it was developed and tested within a single national health system, which may limit generalizability to other countries or EHR infrastructures. While rule-based NLP methods offer interpretability and strong performance in well-characterized contexts, they are sensitive to linguistic variation and may miss novel phrasings or rare edge cases. This was particularly evident for attributes such as “follow-up prescribed,” which had low recall despite high precision.

The gold-standard evaluation set, while carefully constructed and annotated, included only 600 clinical notes, of which just 12 contained positive maculopathy mentions, limiting the robustness of performance estimates for rare subtypes like “unknown.” Additionally, the absence of double annotation or inter-annotator agreement metrics constrains our ability to fully assess annotation reliability. The limitations were anticipated in our study design, and a second set of gold-standard annotations will be conducted by a clinical supervisor to strengthen the evaluation framework and allow for inter-annotator agreement analysis. In retrospect, some clinical reports included in the validation sample were non-informative, reflecting limitations in the note type filtering strategy used to construct the dataset.

Data limitations specific to the AP-HP Clinical Data Warehouse also impacted the analysis. Due to the structure of the EDS, key contextual details such as treatment start dates, cumulative dosing, and patient-level screening timelines are not available in structured form. Extracting these elements requires targeted efforts within free-text clinical reports, which have not yet been carried out due to time constraints. As a result, incidence rates or duration-adjusted risk estimates could be calculated, and it was not possible to precisely link maculopathy mentions to defined periods of HCQ exposure. This limitation reflects a broader structural challenge in the current state of EDS data accessibility and extraction, which has been identified within the MAXYPLUS project. Dedicated efforts, including targeted funding requests, are currently underway to enhance the granularity of treatment and follow-up data for future analysis.

Finally, while the NLP pipeline was successfully applied across thousands of patients, the project timeline was constrained. With more time, additional annotations and iterative refinements could have further improved the pipeline's generalizability, especially for more challenging entities and rare subtypes.

4.8 Future Directions

Several follow-up analyses are planned to extend this work and address limitations identified during the current internship. One priority is a temporal analysis of patients who had both positive and negated mentions of maculopathy, to examine the ordering of these mentions over time. Specifically, we will assess how often a "no maculopathy" mention precedes a later positive mention, potentially indicating early concern or evolving clinical status, or vise versa, which may reflect clinical improvement, diagnostic, or ambiguity.

A second area of focus involves patients with only a single positive mention of maculopathy. These cases will be manually reviewed to determine whether they reflect confirmed diagnoses or more uncertain references (e.g., differential diagnoses or summaries of prior history). Likewise, "unspecified" subtype mentions will be explored further to assess whether contextual information could retrospectively support subtype assignment.

In parallel, further validation work will target suspected false negatives in the stop-treatment category. Some notes that include phrases such as "treatment stopped" may have been missed by the current rule set due to linguistic variation or contextual complexity. Reviewing these cases could support refinements to the pattern library and improve overall recall for treatment discontinuation events.

In parallel, we plan to re-examine suspected false negatives in the stop-treatment category, where mentions of HCQ discontinuation may have been missed due to varied phrasing or because the contextual matcher window was too narrow to capture surrounding cues. These cases will inform refinements to the rule-based patterns and improve recall.

Finally, future work will also explore alternative methods for extracting ophthalmology follow-up attributes. Given the linguistic variability and lower recall observed for certain attributes (e.g., "follow-up prescribed"), we plan to evaluate the use of transformer-based models fine-tuned on annotated data. This approach may enhance performance for more challenging extraction tasks and complement the current rule-based pipeline.

5. CONCLUSION

This study demonstrates the effective development and validation of a rule-based NLP pipeline capable of extracting HCQ exposure, maculopathy, ophthalmology follow-up, and treatment discontinuation from unstructured clinical notes. The algorithm achieved high performance for key clinical entities, including near-perfect scores for HCQ and maculopathy detection, and effectively handled contextual attributes such as negation and follow-up documentation.

Applied at scale across thousands of lupus patients in a large, multicentric French hospital system, the pipeline generated meaningful insights into real-world care patterns. While direct clinical validation and granular treatment timelines remain ongoing challenges, this work illustrates how structured and unstructured EHR data can be leveraged to support pharmacovigilance and care quality evaluation in chronic disease populations. Importantly, it also highlights the role of interpretable, rule-based NLP approaches in surfacing clinically relevant information that is often absent from structured fields.

REFERENCES

1. Pons-Estel GJ, Alarcón GS, Scofield L, Reinlib L, Cooper GS. Understanding the epidemiology and pathogenesis of systemic lupus erythematosus. Nat Rev Rheumatol. 2010;6(12):693–704.
2. Lood C, Alarcón-Riquelme ME. The pathogenesis of cutaneous lupus erythematosus. Curr Opin Rheumatol. 2018;30(5):550–7.
3. Tsakonas E, Joseph L, Esdaile JM. Hydroxychloroquine treatment in lupus. Lupus. 1998;7(1):10–15.
4. Yusuf IH, Charbel Issa P, Ahn SJ. Hydroxychloroquine-induced retinal toxicity. Front Pharmacol. 2023;14:1196783.
5. Melles RB, Marmor MF. The risk of toxic retinopathy in patients on long-term hydroxychloroquine therapy. JAMA Ophthalmol. 2014;132(12):1453–60.
6. Lin WC, Chen JS, Chiang MF, Hribar MR. Applications of artificial intelligence to electronic health record data in ophthalmology. Transl Vis Sci Technol. 2020;9(2):13.
7. Lundberg IE, et al. Using EHR and NLP to identify lupus nephritis phenotypes. BMC Med Res Methodol. 2022;22:251.
8. Petri M, et al. Electronic health record algorithms for identifying patients with systemic lupus erythematosus. Arthritis Rheumatol. 2022;74(4):652–60.
9. Aringer M, Costenbader K, Daikh D, Brinks R, Mosca M, Ramsey-Goldman R, et al. 2019 European League Against Rheumatism/American College of Rheumatology classification criteria for systemic lupus erythematosus. Arthritis Rheumatol. 2019;71(9):1400–12.
10. Wajsbürt P. Metanno: a modular annotator building framework [computer software]. GitHub; 2022. Available from: https://doi.org/10.5281/zenodo.10689827
11. Wajsburt P, Petit-Jean T, Dura B, Cohen A, Jean C, Bey R. EDS-NLP: efficient information extraction from French clinical notes [computer software]. Zenodo; 2022. Available from: https://doi.org/10.5281/zenodo.6424993
12. Melles RB, Jorge AM, Marmor MF, Zhou B, Conell C, Niu J, et al. Hydroxychloroquine dose and risk for incident retinopathy: a cohort study. Ann Intern Med. 2023;176(4):492–500.
13. Jaumouillé S, Espargillière D, Mouriaux F, Mortemousque B. Évaluation en pratique clinique des nouvelles stratégies de dépistage des maculopathies toxiques au plaquénil, selon les nouvelles recommandations de l'American Academy of Ophthalmology. Étude prospective monocentrique à propos de 184 patients. J Fr Ophtalmol. 2015;38(6):520–6.

14. Worme A, Tamariz L, Palacio A, Nemeth Z, Farbman M. A meta-analysis of the prevalence and risk factors for retinal toxicity in rheumatologic patients on hydroxychloroquine therapy [abstract]. Arthritis Rheumatol. 2018;70(Suppl 9).

APPENDIX

List of International Classification of Diseases (ICD)-10 discharge diagnosis codes

ICD-10 code M32 (Systemic Lupus Erythematosus)

ICD-10 code L93 (Cutaneous Lupus Erythematosus)

ICD-10 code M33 (Dermatomyositis)

ICD-10 code M34 (Systemic sclerosis)

ICD-10 code L94 (Scleroderma)

ICD-10 code N085 (Glomerular disease in connective tissue disorders)

ICD-10 code M35 (Sjögren syndrome)

ICD-10 code M069 (Rheumatoid arthritis)

Table A1: Annotation Guide

Entity	Label	Definition	Associated Attribute (Definition)
Medication	HCQ	Mentions of hydroxychloroquine or Plaquenil use	**Contraindication**: HCQ cannot be used due to a medical reason **No Contraindication**: HCQ is not contraindicated or explicitly safe to continue
Ophthalmology Visit	OPHTALMO	Mentions of eye exams, ophthalmology visits, or retinal imaging related to HCQ monitoring.	**Follow-up Done**: Follow-up already completed **Follow-up Scheduled**: Appointment planned **Follow-up Prescribed**: Follow-up recommended but not yet scheduled
HCQ-induced Maculopathy	MACULOPATHY	References to maculopathy, retinal toxicity, vision loss, or related diagnoses associated with HCQ	**Preclinic**: Early or suspected signs of toxicity **Clinic**: Confirmed maculopathy diagnosis **Unknown**: Maculopathy mentioned without clear stage
Treatment Adjustment	STOP_TREATMENT	Mentions of discontinuation of HCQ treatment	*(No sub-attributes)*

Abstract - French

L'hydroxychloroquine (HCQ) est un traitement essentiel du lupus, mais son usage prolongé comporte un risque de toxicité rétinienne. La surveillance de la maculopathie induite par l'HCQ est donc cruciale. Toutefois, les informations cliniques pertinentes (suivi ophtalmologique, arrêt du traitement) sont souvent consignées dans les textes libres des dossiers médicaux électroniques (DME), ce qui complique leur exploitation systématique. Ce projet visait à développer et valider une pipeline de traitement automatique de langage naturel basée sur des règles, capable d'identifier les mentions d'exposition à l'HCQ, de maculopathie, de suivi ophtalmologique et d'arrêt du traitement dans des DME français.

La pipeline a été construite à partir de notes cliniques annotées manuellement et évaluées sur un jeu de validation de 600 comptes-rendus. Elle a ensuite été appliquée à grande échelle à trois cohortes de patients atteints de lupus (systémique, cutané, et élargi), identifiées dans l'Entrepôt de Données de Santé (EDS) de l'AP-HP.

Le système a montré d'excellentes performances pour les entités principales, avec des scores F1 de 0,997 pour l'HCQ et de 0,957 pour la maculopathie. Le filtrage contextuel a amélioré la précision des mentions ophtalmologiques (F1 = 0,921). À l'échelle de plus de 10 000 patients, la pipeline a révélé que la documentation des signes de toxicité et du suivi associé variait selon les sous-types de lupus, avec une fréquence plus élevée chez les patients atteints de lupus systémique.

Ces résultats montrent qu'une pipeline NLP basée sur des règles permet d'exploiter efficacement les textes non structurés pour la surveillance de la toxicité médicamenteuse. Cette approche ouvre la voie à des analyses plus fines dès que des données sur les durées de traitement et les posologies seront accessibles.

Mots-clés : lupus érythémateux, hydroxychloroquine, maculopathie, dossiers médicaux électroniques, traitement automatique du langage naturel