Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Patient representation from clinical narratives for early diagnosis of rare diseases: a comparative study on Activated PI3K-δ Syndrome

Safa Alsaidi

safa.alsaidi@institutimagine.org

Inria, Inserm, Université Paris Cité, HeKA U1346

Marc Vincent

Data Science Platform and/or Clinical Bioinformatics Laboratory, INSERM UMR1163, Imagine Institute, Université Paris Cité

Maxime Douillet

Data Science Platform and/or Clinical Bioinformatics Laboratory, INSERM UMR1163, Imagine Institute, Université Paris Cité

Tomás Brogueira

IT and INESC-ID, Instituto Superior Técnico, Universidade de Lisboa

Nizar Mahlaoui

Pediatric Immuno-Hematology and Rheumatology Unit, Necker Enfants Malades University Hospital, Assistance Publique–Hôpitaux de Paris

Felipe Suarez

Department of Adult Hematology, Necker-Enfants Malades University Hospital and Centre de Référence des déficits Immunitaires Héréditaires, AP-HP, INSERM U1163, Imagine Institute, UPC

Sven Kracker

Laboratory of Lymphocyte Activation and Susceptibility to EBV Infection, INSERM UMR1163, Imagine Institute, Université Paris Cité

Bénédicte Neven

Pediatric Immuno-Hematology and Rheumatology Unit, Necker Enfants Malades University Hospital, Assistance Publique–Hôpitaux de Paris

Miguel Couceiro

IT and INESC-ID, Instituto Superior Técnico, Universidade de Lisboa

Adrien Coulet

Inria, Inserm, Université Paris Cité, HeKA U1346

Nicolas Garcelon

Data Science Platform and/or Clinical Bioinformatics Laboratory, INSERM UMR1163, Imagine Institute, Université Paris Cité

Research Article

Keywords: Activated phosphoinositide 3-kinase delta syndrome , Primary immunodeficiency , Machine learning , Patient representation learning , Clinical text embeddings , Phenotype-based classification , Early diagnosis , Clinical decision support

Posted Date: November 2nd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10115343/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Title Page.

Title of Manuscript: Patient representation from clinical narratives for early diagnosis of rare diseases: a comparative study on Activated PI3K-δ Syndrome

Authors:

Safa Alsaidi[a,b] (safa.alsaidi@institutimagine.org) **(corresponding author)**

Marc Vincent[b] (marc.vincent@institutimagine.org),

Maxime Douillet[b] (maxime.douillet@institutimagine.org),

Tomás Brogueira[c] (tomas.brogueira@tecnico.ulisboa.pt),

Nizar Mahlaoui[d,e] (nizar.mahlaoui@aphp.fr),

Felipe Suarez[g] (felipe.suarez@aphp.fr),

Sven Kracker[f] (sven.kracker@inserm.fr),

Bénédicte Neven[d] (benedicte.neven@aphp.fr),

Miguel Couceiro[c] (miguel.couceiro@tecnico.ulisboa.pt),

Adrien Coulet[a] (adrien.coulet@inria.fr),

Nicolas Garcelon[b] (nicolas.garcelon@institutimagine.org)

Affiliations:

[a] Inria, Inserm, Université Paris Cité, HeKA U1346, 75015 Paris, France

[b] Data Science Platform and/or Clinical Bioinformatics Laboratory, INSERM UMR1163, Imagine Institute, Université Paris Cité, 75015 Paris, France

[c] IT and INESC-ID, Instituto Superior Técnico, Universidade de Lisboa, 1000-029 Lisbon, Portugal

[d] Pediatric Immuno-Hematology and Rheumatology Unit, Necker Enfants Malades University Hospital, Assistance Publique–Hôpitaux de Paris, 75015 Paris, France

[e] French National Reference Center for Primary Immune Deficiencies (CEREDIH), Necker Enfants Malades University Hospital, Assistance Publique–Hôpitaux de Paris, 75015 Paris, France

[f] Laboratory of Lymphocyte Activation and Susceptibility to EBV Infection, INSERM UMR1163, Imagine Institute, Université Paris Cité, 75015 Paris, France

[g] Department of Adult Hematology, Necker-Enfants Malades University Hospital and Centre de Référence des déficits Immunitaires Héréditaires, Assistance Publique-Hôpitaux de Paris, INSERM U1163, Imagine Institute, Université Paris Cité, 75015 Paris, France

Corresponding Author:

Name: Safa Alsaidi

Mailing Address: 24 Bd du Montparnasse, 75015 Paris

Telephone number: +33637510469

Email: safa.alsaidi@institutimagine.org

Abstract.

Background:

Activated phosphoinositide 3-kinase delta syndrome (APDS), first described in 2013, is a rare primary immunodeficiency with an estimated prevalence of 1–2 cases per million. Caused by mutations in *PIK3CD* or *PIK3R1*, it leads to immune dysregulation, lymphoproliferation, recurrent infections, and malignancy. Early identification is critical to enable timely initiation of personalized targeted therapies. This study aimed to evaluate whether machine learning applied to clinical narratives could support earlier identification of APDS patients using information extracted from electronic clinical notes.

Results:

Two approaches derived from clinical notes were evaluated: transformer-based embeddings and phenotype vectors based on UMLS concepts. Machine learning models were trained on 222 patients, including 29 APDS patients. Both approaches achieved strong performance on the full dataset (F1 = 0.82 ± 0.13 and 0.80 ± 0.12, respectively). Phenotype-based analyses included statistical testing, feature importance, ablation studies, and expert review. These analyses identified clinically relevant features including immunodeficiency, adenopathy, lymphopenia, and splenomegaly. Performance remained robust when evaluated on a temporally constrained subset using only pre-diagnosis data (F1 = 0.79 ± 0.14 and 0.74 ± 0.19), supporting the feasibility of earlier detection. Post hoc validation on an independent cohort of 100 patients further supported real-world applicability.

Conclusions:

Machine learning approaches based on clinical narratives can reliably identify APDS patients, including prior to diagnosis. These findings suggest that such methods could support

earlier detection of undiagnosed APDS patients, reducing diagnostic odyssey and improving patient care.

Keywords.

Activated phosphoinositide 3-kinase delta syndrome

Primary immunodeficiency

Machine learning

Patient representation learning

Clinical text embeddings

Phenotype-based classification

Early diagnosis

Clinical decision support

Background

Rare diseases collectively affect an estimated 6% of the global population, corresponding to approximately 300 million people worldwide [1,2]. Their clinical manifestations are often highly variable, even among patients with the same disorder, and may overlap with other conditions [3]. This variability, coupled with clinicians' limited experience, frequently leads to prolonged diagnostic journeys, often referred to as a "diagnostic odyssey", that can span years or decades[3-6].

Accurately representing patient health using diverse clinical data is key to improving diagnosis, assessing risk, and providing personalized care [7-18]. Traditional clinical research mainly relies on structured variables like laboratory values, demographics, and predefined risk factors [7,9,13]. In contrast, modern healthcare systems produce large amounts of unstructured data, including hospitalization reports and discharge summaries, which contain detailed descriptions of symptoms, disease progression, and clinician reasoning [11]. However, unstructured data is difficult to analyze due to its heterogeneity and free-text format, requiring advanced natural language processing (NLP) techniques to extract meaningful information.

Recent advancements in artificial intelligence, especially NLP and large language models (LLMs), have enabled the automatic extraction of meaningful patterns from unstructured text [12-18]. Transformer-based models such as BERT enable *text representation learning*, transforming unstructured clinical narratives into dense numerical embeddings [12-16]. If one considers that a patient's healthcare trajectory can be summarized as a collection of clinical documents describing their care journey, then text representations derived from these documents can be aggregated to form patient-level representations through a process, commonly referred to as *patient representation learning* [13-18]. These representations can

95 then be applied in various tasks, such as disease classification and outcome prediction, and
96 are particularly promising for rare diseases characterized by clinical heterogeneity [23,24].

97 Activated PI3Kδ Syndrome (APDS) is a rare genetic immune disorder caused by pathogenic
98 variants in either *PIK3CD* or *PIK3R1,* with an estimated prevalence of 1-2 affected people
99 per million [25-27]. Affected patients typically experience recurrent infections, immune
100 dysregulation, and increased malignancy risk, though clinical expression varies widely
101 between individuals [26-33]. Many of these features overlap with other primary immune
102 disorders, contributing to substantial diagnostic uncertainty [23,24,27]. Registry data
103 indicates that only a small fraction of APDS cases are currently identified, highlighting
104 substantial underdiagnosis [31-33]. Importantly, the genetic defects causing APDS were only
105 described in 2013 [34,35], meaning that many patients accumulated extensive clinical
106 histories before a definitive diagnosis became possible [31-36].

107 In this context, systematically identifying APDS patients from clinical documentation could
108 substantially accelerate diagnosis [36]. Clinical notes often capture early symptoms,
109 treatment responses, and physician impressions prior to genetic diagnosis. Learning patient
110 representations from such text enables models to capture latent disease signatures [37]. Such
111 approaches are particularly useful in rare diseases, where subtle phenotypic patterns may
112 otherwise remain hidden. In parallel, leveraging phenotype codes extracted via the Unified
113 Medical Language System (UMLS) from clinical notes allow clinical features to be
114 standardized into machine-readable concepts, facilitating patient stratification and analysis.

115 Previous work has proposed pipelines combining phenotype extraction, semantic similarity,
116 and imbalance-aware learning strategies, such as cost-sensitive learning or resampling
117 techniques, to address the challenges of rare disease classification. Together, these
118 approaches aim to amplify informative clinical signals while reducing noise introduced by

large control populations [38]. More recently, studies have also explored the use of LLMs for APDS classification, offering the ability to capture complex contextual and linguistic patterns beyond predefined phenotype ontologies. However, their application to APDS remains limited and challenging due to the scarcity of labeled cases, high class imbalance, and the risk of overfitting in small datasets [39]. Recent work, such as DeepRare [40], has begun to address these challenges by combining deep learning with phenotype-driven representations and rare disease–specific modeling strategies, demonstrating improved performance in low-sample settings. This has led to growing interest in interpretable models that produce human-readable explanations, align with domain knowledge, and can be evaluated in conjunction with expert clinical review [41,42].

In this study, we evaluate the feasibility of accurately classifying APDS patients using unstructured clinical narratives, with a focus on documentation collected prior to diagnosis to APDS diagnosis. Specifically, we leverage French-language clinical models to derive patient embeddings and compare their performance with representations based on UMLS-extracted phenotypes from clinical documents. To assess early diagnostic potential, we restrict part of the analysis to clinical records predating diagnosis, including cases documented before APDS was first described in 2013. By comparing transformer-based and phenotype-based representations, we demonstrate that phenotype representations capture much of the same signal while offering enhanced interpretability. Overall, this work contributes to the growing field of rare disease diagnosis and demonstrates the potential of patient representation learning to support earlier identification of underdiagnosed disorders such as APDS.

Methods

1. Dataset

We used the Necker Hospital data warehouse (Dr. Warehouse), which comprises approximately 11 million pseudonymized clinical documents covering a wide range of rare diseases [43]. From this resource, we identified 49 confirmed APDS patients and 196 controls from the Immunology department, all selected with at least one hospital admission between 2018 and 2023.

For APDS patients, we determined each patient's date of diagnosis by identifying the first occurrence of relevant terms in their clinical notes, including: "APDS," "APDS1," "APDS2," "PIK3CD," "PIK3R1," "PI3K," and "PI3KCD." We also excluded the phenotypic term *primary immunodeficiency*, which typically indicates that the patient had already been diagnosed. All clinical notes dated from the diagnosis date onward were excluded. Controls were selected from the same department, using a structured query to match APDS patients on key clinical and demographic characteristics, i.e., sex, age, follow-up duration (±10%), time between first and last visit, number of clinical events, and care pathways. A ratio of two control patients per APDS case was used.

Following dataset construction, we applied a series of preprocessing steps to standardize the clinical note data and ensure comparability across patients. We first applied interquartile range (IQR)–based filtering on the full dataset. Patients with fewer than three notes (below the lower bound) were excluded and patients with more than 20 notes were truncated by retaining their earliest 20 notes in chronological order. This threshold was selected to limit the overrepresentation of control (non-APDS) patients as APDS patients had a maximum number of notes of 20. The final dataset included 222 patients (29 APDS patients and 193 controls), with a total of 2,627 clinical notes (mean number 11.83 ± 6.02 notes per patient).

In addition, each patient is also associated with a set of UMLS phenotypes by applying the automated phenotyping pipeline described in Vincent et al. [44], which combines a dictionary-based search for phenotypic terms with context-aware detection using deep learning. The dictionary relies on a subset of terms from the UMLS 2023AB vocabulary, focusing on semantic types related to phenotypes. The deep learning modules further enhance the extraction by performing disambiguation (filtering mentions unrelated to phenotypes), subject attribution (assigning mentions to the patient versus family members), and negation detection (identifying whether a phenotype is affirmed or negated). This resulted in 2,563 unique phenotypes, with a mean number of 40.25 ± 48.36 unique phenotypes per patient.

To rule out temporal leakage, we verified the time between the last recorded clinical event and the APDS diagnosis date for all APDS patients. The shortest interval was 16 days, confirming that no post-diagnosis data were present during model training in the full dataset. With the objective of testing if early diagnosis would be possible, we designed a restricted dataset to test the robustness of our approach in a more challenging setting. To this aim, we measured the temporal distance between the last clinical event and the diagnosis date and computed the median of this gap across APDS patients (154 days) and retained only those whose last event occurred ≥154 days before diagnosis. This produced our constrained subset, which is composed of 14 APDS patients, including 7 patients whose last available document predates 2013, i.e., before APDS genetic diagnosis was first established. This subset contains a total of 2,485 clinical notes (mean: 12.00 ± 5.91 notes per patient). In

addition, 2,516 unique phenotypes were included, with a mean number of 40.36 ± 49.6 unique phenotypes per patient.

2. Experimental Design

We defined the APDS diagnosis task as a binary classification problem, where a patient representation X_i maps to an output label Y_i. Depending on the experiment, X_i represents the patient i with an embedding that is either:

a. A vector representation learned by a transformer model from the patient's clinical notes, or
b. A multi-hot encoding vector representing UMLS phenotypes extracted from the same notes.

The output label Y_i equals 1 if the patient i has APDS and 0 otherwise.

2.1. Patient Representation Learning

2.1.1. Transformer-based Embeddings

Based on prior work on French clinical document embeddings, we selected CamemBERT-bio [45], CamemBERT-base [46], and DrBERT [47] to generate transformer-based representations from clinical documents. Due to the variability in clinical note length, document encoding was performed using a sliding-window and hierarchical pooling strategy adapted from prior work [48]. The overall pipeline is illustrated in Figure 1. Briefly, token-level embeddings were aggregated into sentence-, document-, and patient-level representations using mean–max pooling. Full implementation details are provided in Appendix A in the Additional File 1.

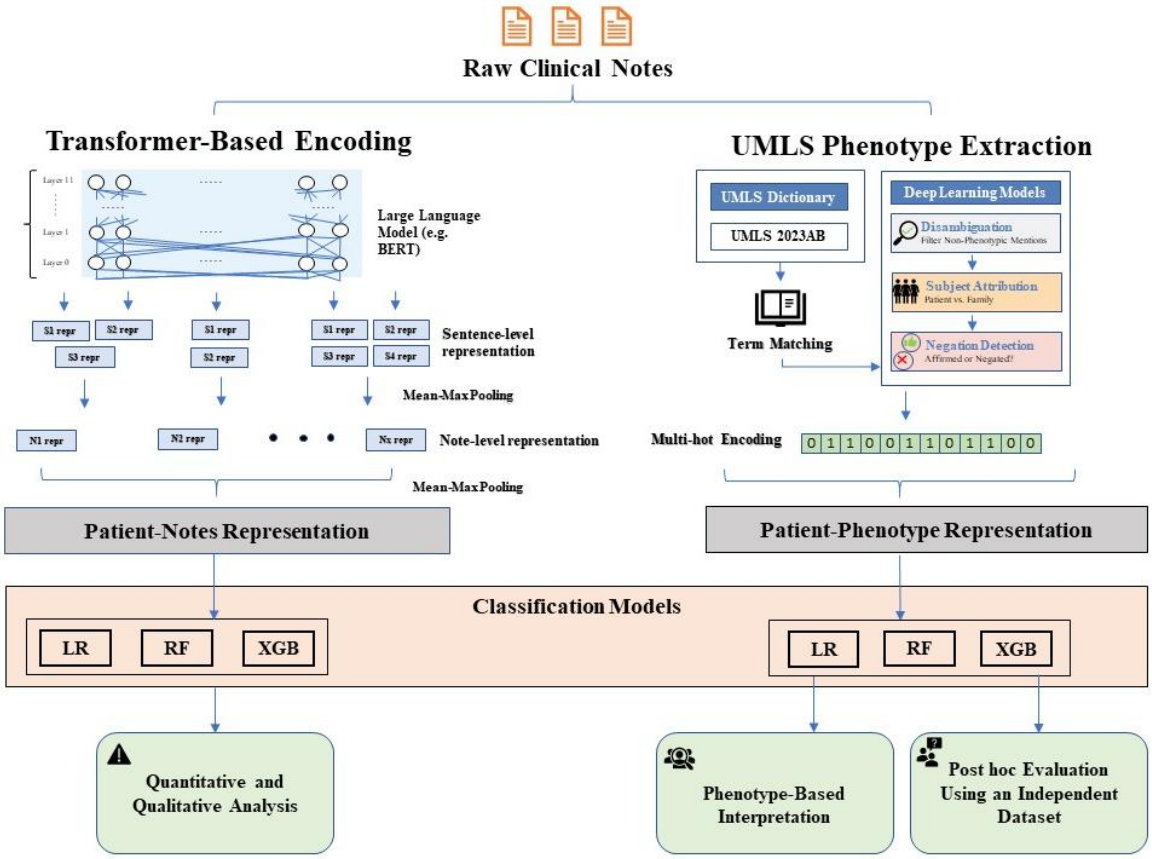


Figure 1. APDS Patient Representation Pipeline Overview. Legend: Starting from raw clinical documents, two parallel processing strategies are applied: (1) extraction of phenotypes using a UMLS-based pipeline, and (2) direct encoding of unstructured text using transformer models. These representations are then used to train classification models. Quantitative and Qualitative analysis are conducted on both phenotypes and transformer-based encodings. Interpretability analyses (ANOVA, feature importance, and ablation) are conducted exclusively on phenotype-based representations, followed by expert clinical validation. A post hoc evaluation on an independent dataset is performed to assess the clinical validity of phenotype-based predictions.

2.1.2. UMLS Phenotype-Based Representations

To encode patient representations in a consistent way, each patient is represented as a multi-hot vector over the full set of extracted phenotypes, indicating the presence (1) or absence (0)

of each concept. No dimensionality reduction was applied to enable a thorough investigation of potentially relevant phenotypic concepts.

3. Model Training and Evaluation Protocol

Each representation is used as input to train three different classifiers: Logistic Regression (LR), Random Forest (RF), and XGBoost (XGB). Performance was evaluated using 5-fold cross-validation, with three folds used for hyperparameter tuning. The evaluation metrics included balanced accuracy, precision, sensitivity, specificity, F1-score, the area under the receiver operating characteristic curve (AU-ROC), and the area under the precision–recall curve (AU-PRC). Balanced accuracy was used to account for class imbalance and is defined as the average of sensitivity and specificity.

To address class imbalance, we applied SMOTE [49] oversampling with a ratio of 0.5 for transformer-based representations. For the phenotype-based representation, class weights were used. The optimal strategy for each representation was selected based on preliminary experiments conducted on the full dataset.

4. Phenotype Importance Analysis

To enhance clinical interpretability, particularly given the limited transparency of transformer-based models, phenotype-based models were analyzed using three complementary methods: univariate statistical testing (ANOVA), model-based feature importance (LR, RF, and XGB), and ablation studies to assess the contribution of individual phenotypes. Details on each method in terms of statistical association, predictive contribution, and discriminative necessity is presented in Table C1 in the Additional File 3. Univariate ANOVA was used to evaluate the statistical association between each phenotype and APDS status, producing an F-value that reflects the strength of this association. Model-

based feature importance was computed for each classifier across cross-validation folds and aggregated at the phenotype level. For LR, feature importance was computed using the model coefficients. Coefficients were extracted from each fold and averaged across folds to obtain a mean coefficient per phenotype. For the RF, feature importance was assessed using SHAP (SHapley Additive exPlanations) values computed with a TreeExplainer. SHAP values were calculated for the positive class, and the mean absolute SHAP value for each phenotype was averaged across samples and folds. For XGB, SHAP values were also computed using a TreeExplainer, with explanations generated on the raw prediction margin (log-odds), as recommended for gradient-boosted tree models [50]. Mean absolute SHAP values were then aggregated across samples and folds to rank features by global importance. We conducted ablation studies using our best performing classifier, where each phenotype was removed from the feature set, and the resulting decrease in weighted AU-ROC was measured to assess its impact on model performance. These analyses were complemented by expert clinical review.

5. Constrained Subset Evaluation

To assess model robustness in early clinical settings, the full evaluation pipeline was repeated on the constrained subset. Phenotype-based analyses (ANOVA and model-based feature importance) and expert validation were also conducted to determine whether key signals were also present in earlier clinical records.

6. Post hoc Experiment

To further assess the clinical validity of phenotype-based predictions, we performed an additional post hoc evaluation using an independent dataset of 100 randomly selected patients (12 APDS, 88 controls) from the immunology department. We conducted a blinded clinician review focusing solely on the misclassified cases. Each patient's phenotype profile,

containing all documented clinical features used by the model, was compiled into a spreadsheet with anonymized patient identifiers for expert validation.

Results

1. Model Performance on the Full Dataset and the Constrained Subset

Table 1 reports performance of the best-performing model for each representation, selected based on the F1 score. Across all experiments, LR consistently emerged as the best-performing classifier. Among transformer-based models, CamemBERT-bio achieved the best overall performance and was therefore selected for the main comparison. Results on the full and constrained dataset are provided in Appendix B in the Additional File 2.

On the full dataset, transformer-based representations achieved strong performance, a balanced accuracy of 0.86 ± 0.09, with high precision (0.96 ± 0.09) and an F1 score of 0.82 ± 0.13. Sensitivity was more moderate (0.73 ± 0.19), indicating that while most APDS cases were correctly identified, some cases were missed. Nevertheless, discriminative performance remained high, as reflected by high AU-ROC (0.96 ± 0.03) and AU-PRC (0.89 ± 0.07) values. For comparison, the best-performing phenotype-based model, reported in our previous study [51], achieved a comparable F1 score of 0.80 ± 0.12. Notably, sensitivity was higher (0.86 ± 0.13), while precision was lower (0.79 ± 0.21), suggesting that phenotype-based models are more effective but at the cost of increased false positives. This configuration achieved the highest discriminative performance, with AU-ROC of 0.98 ± 0.01 and AU-PRC of 0.94 ± 0.04.

On the constrained subset, performance remained high. Transformer-based representations achieved an F1-score of 0.79 ± 0.14, with balanced accuracy of 0.85 ± 0.09. Although sensitivity remained moderate 0.70 ± 0.18, AU-ROC (0.91 ± 0.12) and AU-PRC (0.80 ±

0.17) indicated that discriminative signals persist in earlier clinical records. For phenotype-based representations, sensitivity reached 0.93 ± 0.13, indicating strong detection of APDS cases even in the constrained setting. However, precision decreased (0.64 ± 0.23), reflecting increased uncertainty and a higher number of false positives. Balanced accuracy remained high (0.94 ± 0.08), and AU-ROC (0.99 ± 0.01) and AU-PRC (0.97 ± 0.07) values confirmed strong discriminative power.

Table 1. Performance of LR Across Representations and Datasets.

Dataset	Repr	B. Acc (Mean ± Std)	Precision (Mean ± Std)	Sens (Mean ± Std)	Spec (Mean ± Std)	F1 (Mean ± Std)	AU-ROC (Mean ± Std)	AU-PRC (Mean ± Std)
Full	Text	0.86 ± 0.09	0.96 ± 0.09	0.73 ± 0.19	0.99 ± 0.01	0.82 ± 0.13	0.96 ± 0.03	0.89 ± 0.07
	Phen - 2563 features	0.91 ± 0.06	0.79 ± 0.21	0.86 ± 0.13	0.95 ± 0.05	0.80 ± 0.12	0.98 ± 0.01	0.94 ± 0.04
Cons. subset	Text	0.85 ± 0.09	0.93 ± 0.15	0.70 ± 0.18	0.99 ± 0.01	0.79 ± 0.14	0.91 ± 0.12	0.80 ± 0.17
	Phen - 2516 features	0.94 ± 0.08	0.64 ± 0.23	0.93 ± 0.13	0.95 ± 0.05	0.74 ± 0.19	0.99 ± 0.01	0.97 ± 0.07

Legend: Performance of the best-performing classifier (LR) for each data representation and experimental settings. Metrics are reported as mean ± standard deviation across cross-validation folds. ***Abbreviations:*** *Cons.: Constrained; Repr: Representation; B. Acc = Balanced Accuracy; Sens = Sensitivity; Spec = Specificity; F1 = F1 score; AU-ROC = Area Under the Receiver Operating Characteristic Curve; AU-PRC = Area Under the Precision-Recall Curve.*

2. Phenotype Analysis on the Full Dataset

To gain clearer insight into the key manifestations driving classification, we performed a detailed analysis of our phenotype-based representations.

2.1. Univariate Analysis (ANOVA)

ANOVA revealed strong associations between several phenotypes and APDS status. *Immunodeficiency* emerged as the dominant phenotype, followed by *combined immunodeficiency, CD4 T-cell lymphopenia, bronchiectasis, lymphopenia, hyper IgM syndrome,* and *adenopathy* (Figure 2). Additional features such as *bronchitis, splenomegaly, and otitis* also ranked highly were also among the top-ranked phenotypes.

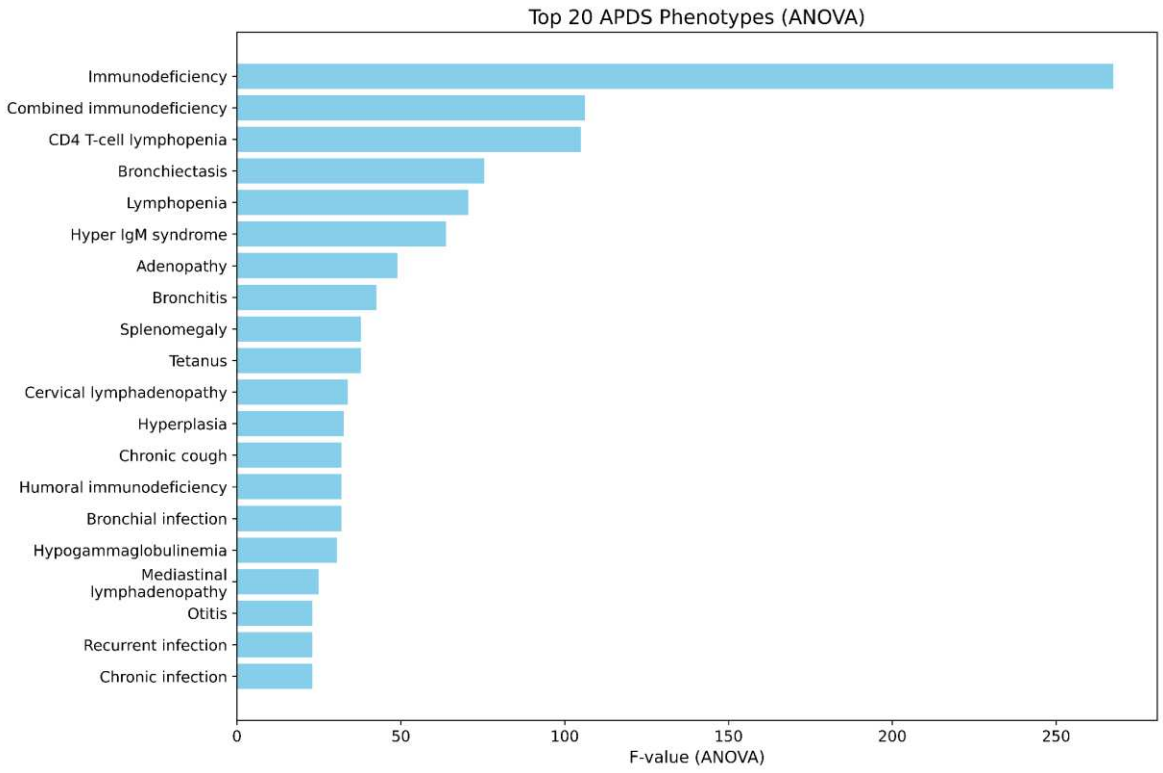


Figure 2. Top 20 APDS-Associated Phenotypes (ANOVA) on the Full Dataset. Legend: Phenotypes are ranked by F-value from ANOVA, indicating the strength of their statistical association with APDS. Higher F-values correspond to stronger associations.

2.2. Model-Based Feature Importance

Feature importance across LR, RF, and XGB models consistently prioritized core immune and infectious phenotypes. An expert-driven review was conducted by two clinical experts and a biologist indicated that 21 of the 27 features (78%) are directly consistent with known APDS manifestations, while the remaining underlined phenotypes were not immediately interpretable as primary APDS features (Table 2).

Table 2. Top APDS-Associated Phenotypes Across Models on the Full Dataset.

Group	Phenotypes
Shared across all three models (LR, RF, XGB)	Immunodeficiency, Splenomegaly, Adenopathy, Lymphopenia, CD4 T-cell lymphopenia, Rhinorrhea, Otitis, Hypogammaglobulinemia, Hyperplasia, Tetanus, Bronchiectasis, Infection
Shared across LR & RF	Combined immunodeficiency, Cervical lymphadenopathy
Shared across LR & XGB	Hypothyroidism, Otorrhea, Bronchitis, Viral load
Shared across RF & XGB	Pain, Anesthesia, Pneumonia, Vomiting, Cough
Unique to LR	Superinfection, Thrombocytopenia
Unique to RF	Hypertrophy
Unique to XGB	Fever

Legend: Top APDS-associated phenotypes identified by LR, RF, and XGB on the full dataset. Underlined phenotypes indicate features not directly associated with core APDS manifestations based on the expert-driven review.

2.3. Ablation Analysis

Ablation studies confirmed the importance of key phenotypes, where the removal of *immunodeficiency* caused the largest performance drop (0.354), followed by *cervical lymphadenopathy, lymphopenia, CD4 T-cell lymphopenia,* and *splenomegaly* (Figure 3).

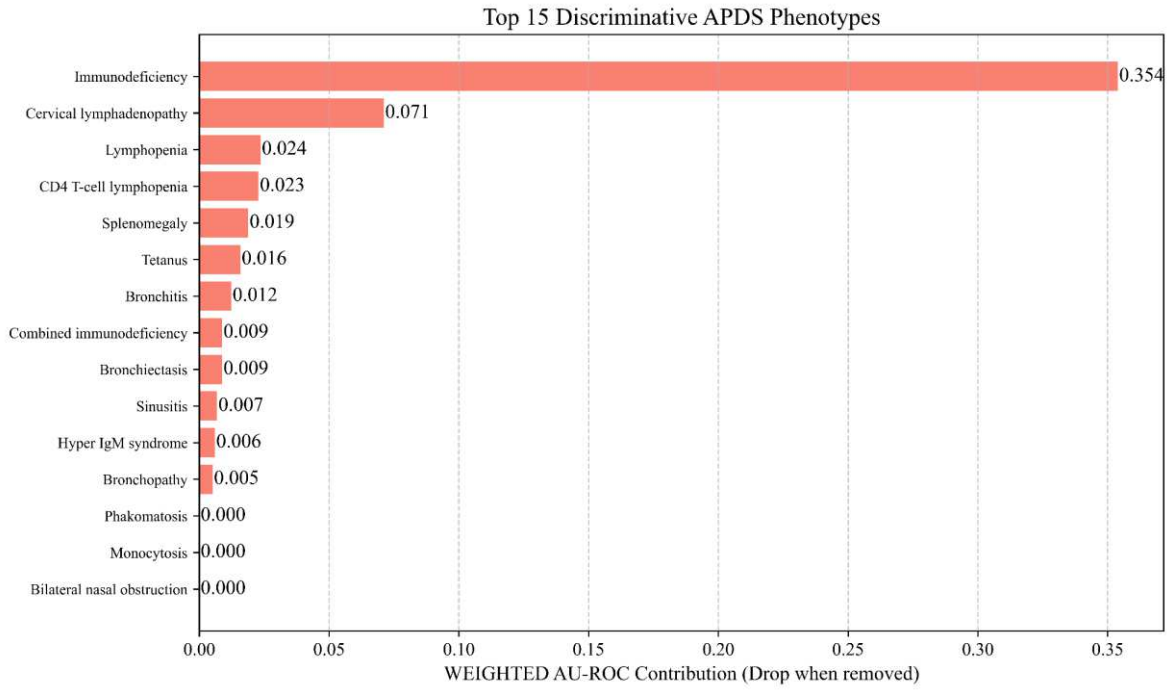


Figure 3. Ablation Study of Phenotype Contributions to APDS Classification using the Full Dataset. Legend: Bars show the decrease in weighted AU-ROC when individual phenotypes are removed from the LR model, highlighting the features with the strongest impact on APDS discrimination.

These findings confirmed that immune-related and lymphoproliferative features are not only statistically associated and predictive but also essential for accurate classification.

3. Phenotype Analysis on the Constrained Subset

Phenotype analysis on the constrained subset yielded similar results. The results from the ANOVA analysis closely mirrored those from the full dataset (Figure 4). Phenotypes such as

immunodeficiency, combined immunodeficiency, CD4 T-cell lymphopenia, bronchiectasis, chronic and recurrent infections, hypogammaglobulinemia, chronic cough, cervical lymphadenopathy, splenomegaly, hyper IgM syndrome, humoral immunodeficiency, and *mediastinal lymphadenopathy* consistently ranked among the top features.

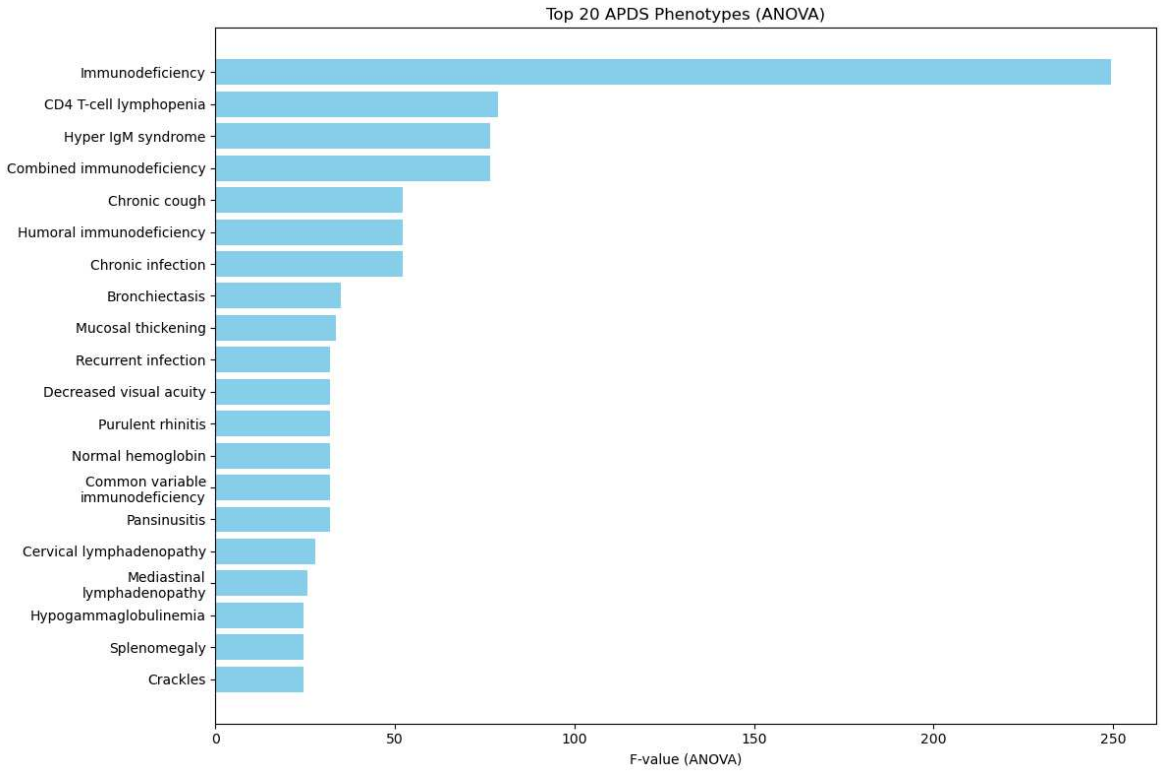


Figure 4. Top 20 APDS-Associated Phenotypes (ANOVA) on the Constrained Dataset.

Legend: Phenotypes are ranked by F-value from ANOVA, indicating the strength of their statistical association with APDS. Higher F-values correspond to stronger associations.

Expert analysis on the model-based feature importance revealed that 22 of the 30 features (73%) were directly consistent with known APDS manifestations (Table 3). Several additional phenotypes not observed in the full dataset emerged, including *crackles, hyper IgM syndrome, rhonchi, pulmonary infection, asthenia, lymphoma, allergy, nausea, and xerosis.* Five features – *tetanus, pain, vomiting, fever, and anesthesia* – were also previously identified as non-specific in the full dataset.

Table 3. Top APDS-Associated Phenotypes Across Models on the Constrained Dataset.

Group	Phenotypes
Shared across all three models (LR, RF, XGB)	Immunodeficiency, Infection, CD4 T-cell lymphopenia, Adenopathy, Rhinorrhea, Otitis, Pneumonia, Hypogammaglobulinemia
Shared across LR & RF	Splenomegaly, Cervical lymphadenopathy, Crackles, Rhonchi, Hyper IgM syndrome
Shared across LR & XGB	Superinfection, Pulmonary infection, Tetanus
Shared across RF & XGB	Pain, Anesthesia, Cough, Vomiting, Asthenia
Unique to LR	Hypothyroidism, Combined immunodeficiency, Chronic cough, Lymphoma
Unique to RF	Nausea
Unique to XGB	Fever, Allergy, Xerosis, Bronchiectasis

Legend: Top APDS-associated phenotypes identified by LR, RF, and XGB on the constrained dataset. Underlined phenotypes indicate features not directly associated with core APDS manifestations based on the expert-driven review.

4. Post hoc Experiment Results

On the independent dataset, the phenotype-based model correctly identified all 12 APDS patients, achieving perfect sensitivity. Among the 88 control patients, 15 were misclassified (false positives).

The blinded clinician review showed:

- For 9 cases, all reviewers unanimously agreed that the patient did not have APDS.
- In 2 cases, 2 of 3 reviewers suggested possible APDS or recommended genetic testing.
- In 4 cases, 1 reviewer suggested APDS.

Overall, all 15 cases achieved majority agreement among reviewers (*i.e.,* at least two of three reviewers agreed on the classification). These results suggest that the model prioritizes sensitivity and may flag patients who could still warrant further clinical evaluation.

Discussion

1. Model Performance and Robustness

A clear distinction was observed between representations: phenotype-based models consistently achieved higher sensitivity, whereas transformer-based models favored precision. These differences likely reflect how each representation interacts with linear classifiers and decision thresholds, resulting in distinct precision–sensitivity trade-offs [52,53]. While F1 ranking slightly favored text-based models, phenotype-based representations achieved higher AU-ROC values, suggesting stronger overall discriminative ability and indicating that threshold optimization could yield comparatively improved F1 performance for phenotype-based approaches.

In rare diseases such as APDS, where positive cases are inherently scarce, sensitivity is particularly important. Clinically, identifying as many true cases as possible is essential, as missed cases may delay diagnosis and treatment. In this light, the higher sensitivity of phenotype-based representations suggests greater suitability for pre-screening applications, despite lower precision.

Despite these differences, both representations showed strong performance using LR, which is consistent with rare-disease settings where simpler, regularized models often generalize more effectively from limited data [54]. These findings were further supported by the high performance observed on the constrained subset, despite the reduced sample size and stricter temporal constraints. Overall, these results demonstrate that both narrative clinical text and phenotypic data contain early, informative signals enabling retrospective identification of APDS.

2. Error Analysis

Qualitative review through detailed case-by-case examination of misclassified cases in collaboration with an immunologist revealed several sources of error. For transformer-based models, false negatives were primarily attributable to the lack of sufficiently distinctive APDS-related information in clinical documentation prior to diagnosis. False positives frequently arose in patients with a family history of APDS, introducing misleading contextual signals. Across both representations, some misclassifications reflected the clinical heterogeneity of APDS and its overlap with other primary immunodeficiencies. These findings highlight the inherent ambiguity of free-text narratives in rare and phenotypically diverse diseases.

3. Phenotypic Insights and Clinical Relevance

Our phenotype-based analysis consistently identified immune dysregulation and lymphoproliferative features as core indicators of APDS across both datasets. In contrast, underlined features in Tables 2 and 3 were interpreted as secondary or contextual variables, reflecting vaccination records, diagnostic procedures, or nonspecific infection-related episodes rather than disease-defining characteristics.

Comparison with authoritative rare disease resources, specifically Orphanet (ORPHA:397596) [55], further supported these findings as seen in Table C2 in the Additional File 3. This concordance indicates that the features identified by the model are both statistically robust and biologically meaningful rather than spurious associations.

Beyond classification, these findings have important clinical implications. Phenotype-driven models may support earlier detection of APDS and similar rare diseases, particularly when genetic testing is unavailable or delayed. The identified phenotypes could also help define clinical archetypes, supporting patient stratification, personalized care, and the development of realistic synthetic data for training and validating clinical decision support systems.

4. Post hoc Findings and Interpretation

The post hoc evaluation provided additional evidence of clinical relevance. Notably, the model correctly identified all APDS patients in the independent cohort, ensuring that no affected patients were overlooked.

The blinded clinician review revealed that many of the 15 misclassified patients were not arbitrary errors, where 10 out of the 15 patients exhibited **s**ubstantial overlap with APDS-associated phenotypes. Several reviewers suggested genetic testing for a subset of these patients, highlighting that the model's false positives may still correspond to clinically

relevant cases that warrant additional attention. This behavior aligns well with the intended use of the model as a clinical triage tool and highlights the importance of combining computational tools with human expertise. In practice, this may facilitate earlier identification of atypical or complex cases and support more timely referral for genetic testing, improving early detection and reducing underdiagnosis.

Taken together, these findings reinforce the potential of phenotype-based models to support rare disease diagnosis by prioritizing sensitivity, highlighting clinically meaningful patterns, and assisting clinicians in identifying patients who may benefit from further evaluation. Future work should focus on improving specificity while preserving this high sensitivity and validating these approaches in larger and more diverse populations.

Conclusions

This study demonstrates that machine learning approaches applied to clinical narratives can reliably identify patients with Activated phosphoinositide 3-kinase delta syndrome (APDS), including before formal diagnosis. Both transformer-based and phenotype-based representations achieved strong predictive performance, while phenotype-driven approaches provided greater interpretability and sensitivity, making them particularly suitable for early screening applications in rare disease settings.

Importantly, the consistent identification of clinically relevant immune and lymphoproliferative features across datasets supports the biological validity of the learned signals and highlights the value of phenotype-based analyses beyond prediction alone. The robust performance observed in temporally constrained settings further suggests that meaningful disease-related patterns are present in early clinical records and can be leveraged for retrospective identification.

These findings are particularly relevant for the rare disease field, where diagnostic delays remain a major challenge. By enabling earlier identification of patients who may benefit from targeted genetic testing or specialized clinical evaluation, such approaches could help reduce diagnostic odysseys, improve patient management, and facilitate timely access to personalized therapies. More broadly, this work supports the potential of artificial intelligence–driven analysis of routine clinical data as a scalable strategy for improving rare disease detection and clinical decision support.

List of Abbreviations.

NLP – Natural Language Processing

LLM – Large Language Model

APDS – Activated PI3Kδ Syndrome

UMLS – Unified Medical Language System

LR – Logistic Regression

RF – Random Forest

XGB – XGBoost

AU-ROC – Area under the Receiver-Operating-Characteristic Curve

AU-PRC – Area under the Precision–Recall Curve

ANOVA – Analysis of Variance

SHAP – SHapley Additive exPlanations

Declarations

Ethics approval and consent to participate

This retrospective study was conducted on the clinical reports of patients followed at Necker-Enfants Malades Hospital (AP-HP). The study was carried out in compliance with the MR-

004 reference methodology of the French Data Protection Authority (CNIL). In accordance with this framework, patients were informed of the potential reuse of their data for research purposes and were given the opportunity to exercise their right to object; none of the included patients objected. This data processing was registered with the Data Protection Officer of AP-HP under reference no. 2026 04171640. The study was performed in accordance with the principles of the Declaration of Helsinki.

Consent for publication

Not applicable

Availability of data and materials

The clinical datasets from this study are not publicly available, as institutional officials expressed concern about the inability to guarantee anonymity. Aggregated data and datasets containing coarse-grained phenotypes are available upon request to the corresponding author.

Competing interests

NM, MC, TB, MD, MV, BN, FS, and SA declare that they have no competing interests. SK reports collaboration agreements, as well as payments and travel support, from Pharming Group N.V. AC has received funding from Sanofi and Doctolib. NG is a co-founder of CoDoc, a company specializing in health data warehouses.

Funding

This work was supported by the French National Research Agency (ANR) through the project “Analogies: from theory to tools and applications” (AT2TA) (ANR-22-CE23-0023), and partially funded under the “Investissements d’Avenir” program (ANR-10-IAHU-01). SK is a staff researcher at the Centre National de la Recherche Scientifique.

Authors' contributions

SA conceived the study, designed and conducted the experiments, performed data analysis and interpretation, and was the major contributor in writing the manuscript. MC, AC, and NG supervised the work, contributed to the study design and interpretation of the results, and critically reviewed and revised the manuscript. MD contributed to dataset construction and cohort development. MV contributed to the development and implementation of the phenotype extraction pipeline and provided methodological expertise. NM, BN, FS, and SK contributed clinical and biological expertise related to APDS, participated in the interpretation and validation of the findings from a clinical perspective, and critically reviewed the manuscript. TB contributed to manuscript revision and review. All authors read and approved the final manuscript.

Acknowledgements

Not applicable

References

1. Schieppati A, Henter JI, Daina E, Aperia A. Why rare diseases are an important medical and social issue. The Lancet. 2008;371(9629):2039-41.
2. Rehm HL. Time to make rare disease diagnosis accessible to all. Nat Med. 2022;28(2):241-2.
3. Bauskis A, Strange C, Molster C, Fisher C. The diagnostic odyssey: insights from parents of children living with an undiagnosed condition. Orphanet J Rare Dis. 2022;17(1):233.
4. Chaudhary A, Kumar V. Rare diseases: a comprehensive literature review and future directions. J Rare Dis. 2025;4(1):33.
5. Al-Worafi YM. Patient Care Related Issues in the Developing Countries: Diagnosis. InHandbook of Medical and Health Sciences in Developing Countries: Education, Practice, and Research. Cham: Springer; 2023. p. 1-21.
6. Gahl WA, Markello TC, Toro C, Fajardo KF, Sincan M, Gill F, et al. The national institutes of health undiagnosed diseases program: insights into rare diseases. Genet Med. 2012;14(1):51-9.
7. Liu L, Li H, Hu Z, Shi H, Wang Z, Tang J, Zhang M. Learning hierarchical representations of electronic health records for clinical outcome prediction. AMIA Annu Symp Proc. 2019;2019:597.
8. Fridgeirsson EA, Sontag D, Rijnbeek P. Attention-based neural networks for clinical prediction modelling on electronic health records. BMC Med Res Methodol. 2023;23(1):285.

9. Liu F, Li L, Wang X, Luo F, Liu C, Su J, et al. MHGRL: An Effective Representation Learning Model for Electronic Health Records. In: Proc LREC-COLING 2024; 2024. p. 11272-11282.
10. Amirahmadi A, Ohlsson M, Etminani K. Deep learning prediction models based on EHR trajectories: A systematic review. J Biomed Inform. 2023;144:104430.
11. Martin-Sanchez F, Verspoor K. Big data in medicine is driving big changes. Yearb Med Inform. 2014;23(1):14-20.
12. Spasic I, Nenadic G. Clinical text data in machine learning: systematic review. JMIR Med Inform. 2020;8(3):e17984.
13. Si Y, Du J, Li Z, Jiang X, Miller T, Wang F, et al. Deep representation learning of patient data from Electronic Health Records (EHR): A systematic review. J Biomed Inform. 2021;115:103671
14. Steinberg E, Jung K, Fries JA, Corbin CK, Pfohl SR, Shah NH. Language models are an effective representation learning technique for electronic health record data. J Biomed Inform. 2021;113:103637.
15. Yang X, Chen A, PourNejatian N, Shin HC, Smith KE, Parisien C, et al. A large language model for electronic health records. NPJ Digit Med. 2022;5(1):1-9. https://doi.org/10.1038/s41746-022-00742-2
16. Kauffman J, Miotto R, Klang E, Costa A, Norgeot B, Zitnik M, et al. Embedding Methods for Electronic Health Record Research. Annu Rev Biomed Data Sci. 2025;8.
17. Xian S, Grabowska ME, Kullo IJ, Luo Y, Smoller JW, Walunas TL, et al. Transformer patient embedding using electronic health records enables patient stratification and progression analysis. NPJ Digit Med. 2025;8(1):521.
18. Gao Y, Myers S, Chen S, Dligach D, Miller TA, Bitterman D, et al. When raw data prevails: are large language model embeddings effective in numerical data

554 representation for medical machine learning applications? In: Findings ACL EMNLP;
555 2024. p. 5414-28.
556 19. Devlin J, Chang MW, Lee K, Bert KT. Pre-training of deep bidirectional transformers
557 for language understanding. arXiv. 2018;1810.04805.
558 20. Alsentzer E, Murphy J, Boag W, Weng WH, Jindi D, Naumann T, et al. Publicly
559 available clinical BERT embeddings. In: Proc ClinNLP; 2019. p. 72-78.
560 21. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained
561 biomedical language representation model for biomedical text mining.
562 Bioinformatics. 2020;36(4):1234-40.
563 22. Modell V. The impact of physician education and public awareness on early diagnosis
564 of primary immunodeficiencies: Robert A. Good Immunology Symposium. Immunol
565 Res. 2007;38(1):43-7.
566 23. Tavakol M, Jamee M, Azizi G, Sadri H, Bagheri Y, Zaki-Dizaji M, et al. Diagnostic
567 approach to the patients with suspected primary immunodeficiency. Endocr Metab
568 Immune Disord Drug Targets. 2020;20(2):157-71.
569 24. Singh A, Joshi V, Jindal AK, Mathew B, Rawat A. An updated review on activated
570 PI3 kinase delta syndrome (APDS). Genes Dis. 2020;7(1):67-74.
571 25. Lougaris V, Piane FL, Cancrini C, Conti F, Tommasini A, Badolato R, et al.
572 Activated phosphoinositide 3-kinase (PI3Kδ) syndrome: an Italian point of view on
573 diagnosis and new advances in treatment. Ital J Pediatr. 2024;50(1):103.
574 26. Vanselow S, Wahn V, Schuetz C. Activated PI3Kδ syndrome–reviewing challenges
575 in diagnosis and treatment. Front Immunol. 2023;14:1208567.
576 27. Thouenon R, Moreno-Corona NC, Poggi L, Durandy A, Kracker S. Activated
577 PI3Kinase Delta Syndrome—A multifaceted disease. Front Pediatr. 2021;9:652405.

28. Hanson J, Bonnen PE. Systematic review of mortality and survival rates for APDS. Clin Exp Med. 2024;24(1):17.

29. Zhu K, Li Q, Han L, Sun J. Activated phosphoinositide 3-kinase delta syndrome: Pathogenesis, clinical manifestations, and treatment. Pediatr Discov. 2024;2(3):e2504.

30. Alberto GL, Emmanuel AC, Alberto YD, Daniel AB, Fernando LP, Eugenia VC, et al. Approach to Primary Immunodeficiencies for the Non-Immunologist Physicians. Open Access Libr J. 2019;6:e5419.

31. Mahlaoui N, Jais JP, Brosselin P, Mignot C, Beaurain B, Brito C, et al. Prevalence of primary immunodeficiencies in France is underestimated. J Allergy Clin Immunol. 2017;140(6):1731-3.

32. Ryan P, Redenbaugh V, McGucken J, Kindle G, Devlin LA, Coulter T, et al. Inborn errors of immunity on the island of Ireland—a cross-jurisdictional UKPID/ESID registry report. J Clin Immunol. 2022;42(6):1293-9.

33. Maccari ME, Wolkewitz M, Schwab C, Lorenzini T, Leiding JW, Aladjdi N, et al. Activated phosphoinositide 3-kinase δ syndrome: update from the ESID Registry and comparison with other autoimmune-lymphoproliferative inborn errors of immunity. J Allergy Clin Immunol. 2023;152(4):984-96.

34. Angulo I, Vadas O, Garçon F, Banham-Hall E, Plagnol V, Leahy TR, et al. Phosphoinositide 3-kinase δ gene mutation predisposes to respiratory infection and airway damage. Science. 2013;342(6160):866-71.

35. Deau MC, Heurtier L, Frange P, Suarez F, Bole-Feysot C, Nitschke P, et al. A human immunodeficiency caused by mutations in the *PIK3R1* gene. *J Clin Invest*. 2014;124(9):3923–3928.

36. FitzPatrick AM, Chin AT, Nirenberg S, Cunningham-Rundles C, Sacco K, Perlmutter J, et al. Piloting an automated query and scoring system to facilitate APDS patient identification from health systems. Front Immunol. 2025;15:1508780.
37. Yang J, Liu C, Deng W, Wu D, Weng C, Zhou Y, et al. Enhancing phenotype recognition in clinical notes using large language models: PhenoBCBERT and PhenoGPT. Patterns. 2024;5(1).
38. Faviez C, Vincent M, Garcelon N, Boyer O, Knebelmann B, Heidet L, et al. Performance and clinical utility of a new supervised machine-learning pipeline in detecting rare ciliopathy patients based on deep phenotyping from electronic health records and semantic similarity. Orphanet J Rare Dis. 2024;19(1):55.
39. Rider NL, Li Y, Chin AT, DiGiacomo DV, Dutmer C, Farmer JR, et al. Evaluating large language model performance to support the diagnosis and management of patients with primary immune disorders. J Allergy Clin Immunol. 2025.
40. Zhao W, Wu C, Fan Y, Qiu P, Zhang X, Sun Y, et al. An agentic system for rare disease diagnosis with traceable reasoning. Nature. 2026:1-0.
41. Raposo VL. The fifty shades of black: about black box AI and explainability in healthcare. Med Law Rev. 2025;33(1):fwaf005.
42. Sciaccotta R, Barone P, Murdaca G, Fazio M, Stagno F, Gangemi S, et al. Decoding immunodeficiencies with artificial intelligence: A new era of precision medicine. Biomedicines. 2025;13(8):1836.
43. Garcelon N, Neuraz A, Salomon R, Faour H, Benoit V, Delapalme A, et al. A clinician friendly data warehouse oriented toward narrative reports: Dr. Warehouse. J Biomed Inform. 2018;80:52-63.

624 44. Vincent M, Douillet M, Lerner I, Neuraz A, Burgun A, Garcelon N. Using deep
625 learning to improve phenotyping from clinical reports. In: MEDINFO 2021; 2022. p.
626 282-286.
627 45. Martin L, Muller B, Suarez PO, Dupont Y, Romary L, de La Clergerie ÉV, et al.
628 CamemBERT: a tasty French language model. In: Proc ACL; 2020. p. 7203-7219.
629 46. Touchent R, de la Clergerie ÉV. CamemBERT-bio: leveraging continual pre-training
630 for cost-effective models on French biomedical data. In: Proc LREC-COLING; 2024.
631 p. 2692-2701.
632 47. Labrak Y, Bazoge A, Dufour R, Rouvier M, Morin E, Daille B, et al. DrBERT: A
633 robust pre-trained model in French for biomedical and clinical domains. In: Proc
634 ACL; 2023. p. 16207-16221.
635 48. Alsaidi S, Vincent M, Boyer O, Garcelon N, Couceiro M, Coulet A. Comparing
636 representations of long clinical texts for the task of patient-note identification. In:
637 Proc CL4Health; 2025. p. 109-123.
638 49. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority
639 over-sampling technique. J Artif Intell Res. 2002;16:321-57.
640 50. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. From local
641 explanations to global understanding with explainable AI for trees. Nat Mach Intell.
642 2020;2(1):56-67.
643 51. Alsaidi S, Brogueira T, Mahlaoui N, Vincent M, Pelegrina G, Garcelon N, et al.
644 Shapley regression for rare disease diagnosis support: a case study on APDS. In:
645 Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI
646 2026); 2026 Aug; Bremen, Germany. 2026. doi:10.48550/arXiv.2605.08897.

52. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One. 2015;10(3):e0118432.

53. Martinčić-Ipšić S, Miličić T, Todorovski L. The influence of feature representation of text on the performance of document classification. Appl Sci. 2019;9(4):743.

54. Hidayat R, Mahdiana D, Fergina A. Comparative analysis of logistic regression, SVM, Xgboost, and random forest algorithms for diabetes classification. J Teknol Sist Inf Apl. 2024;7(1):281-91.

55. Pavan S, Rommel K, Mateo Marquina ME, Höhn S, Lanneau V, Rath A. Clinical practice guidelines for rare diseases: the orphanet database. PLoS One. 2017;12(1):e0170365.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- OrphanetALSAIDIAdditonalfile1.docx
- OrphanetALSAIDIAdditonalfile2.docx
- OrphanetALSAIDIAdditonalfile3.docx