Large-scale deep learning analysis for the early diagnosis of primary immunodeficiencies

Giorgos Papanastasiou (georgios.papanastasiou@pfizer.com)  
Pfizer Inc, New York, NY, USA

Guang Yang  
Imperial College London, UK

Dimitris Fotiadis  
Department of Biomedical Research, Institute of Molecular Biology and Biotechnology, FORTH, Ioannina, Greece

Nikolaos Dikaios  
Mathematics Research Center, Academy of Athens, Athens, Greece

Chengjia Wang  
Heriot Watt University, UK

Ahsan Huda  
Pfizer Inc, New York, NY, USA

Luba Sobolevsky  
Immunoglobulin National Society, Woodland Hills, CA, USA

Gurinder Sidhu  
Pfizer Inc, New York, NY, USA

Donna Palumbo  
Pfizer Inc, New York, NY, USA

Research Article

Keywords: primary immunodeficiency, combined immunodeficiency, common variable immunodeficiency, early diagnosis, deep learning

Posted Date: December 29th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-2422924/v1

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

Primary immunodeficiency (PID) is a group of heterogeneous disorders resulting from immune system defects. The early PID diagnosis is compromised by the heterogeneous manifestations along with low clinical awareness. Most PID cases are significantly underdiagnosed leading to increased mortality, comorbidities and healthcare visits and costs. Among PID, combined immunodeficiencies (CID) are characterized by complex immune defects. Common variable immunodeficiency (CVID) is among the most common types of PID. In light of available treatments for CID and CVID, it is critical to systematize their early diagnosis. Our study objectives were two-fold. First, we developed and evaluated an accurate deep learning model to analyze administrative medical claims data from EHRs towards systematizing screening and early identification of CID and CVID. Second, we revealed the most important CID- and CVID-associated clinical phenotypes and their combinations, demonstrating a systematic methodology to improve early identification of these PID. All data were composed of medical claims derived from the Optum® de-identified electronic health record (EHR) database. Four large cohorts were generated: 797, 797, 2,312, and 19,924 CID/CVID cases and equal control sizes in Cohorts 1–4, respectively (a total of 47,660 cases and controls). Two deep learning models were developed (TabMLPNet and TabResNet) and compared against baseline models. Univariate logistic regression was used to calculate odds ratios across all clinical phenotypes and their combinations. The TabMLPNet model showed the highest diagnostic performance across cohorts with sensitivity, specificity, and overall accuracy ranging from 0.82–0.88, 0.82–0.85, and 0.80–0.87, respectively. For the first time, we identified distinctive combinations of antecedent phenotypes associated with CID/CVID per cohort, being consisted of respiratory infections/conditions, genetic anomalies, cardiac defects, autoimmune diseases, blood disorders and malignancies. Most phenotypes emerged were well described in the literature, which validated our findings. Moreover, several less well documented individual phenotypes (i.e., asthma, coagulation defects complicating pregnancy, cancer of lymphoid histiocytic tissue, lymphoid leukemia chronic) were also identified, which can lead to better clinical surveillance of PID. We demonstrated a generalized and accurate method evaluated on a large EHR-derived cohort of CID/CVID cases and controls. Our methodology can lead to the development of new clinical guidelines and pathways for earlier identification of the most important antecedent phenotypes and their combinations, enhance clinical awareness and be used to improve PID diagnosis and outcomes on a population level.

Introduction

Primary immunodeficiency (PID) is an heterogeneous group of disorders resulting from defects of one or more components of the immune system\(^1\). PID patients are susceptible to serious, life-threatening infections, organ damage, secondary malignancies and autoimmune diseases\(^2\). As of 2020, more than 450 PID subtypes have been discovered which were linked to 485 genetic defects\(^3,4\). This is an increase from over 350 PIDs in 2017\(^5\). As further PID research is conducted, it is anticipated that this number will continue to increase. Early PID diagnosis is critical to improving health outcomes, and reducing morbidity and mortality\(^6\). Improvements in genetic, immunologic, imaging and medical assessments allow the
characterization and therapeutic intervention of many PID disorders\textsuperscript{4,6}. A significant challenge to the early PID diagnosis is the highly heterogeneous clinical presentation, across and within PID subtypes\textsuperscript{3–6}. Another critical barrier is the low awareness of PID among primary care practitioners and hence, a lack of referral to clinical immunologists, leading to suboptimal diagnostic evaluation\textsuperscript{1,6}.

Awareness campaigns by advocacy groups have identified the warning signs to help identify PID patients (Supplementary Table 1)\textsuperscript{7}. Although clinically relevant, this specific set of manifestations does not provide a comprehensive list of clinical phenotypes for systematizing PID screening\textsuperscript{8}. Apart from severe combined immunodeficiency (SCID) for which newborn screening is established in the United States, population-based screening for PID does not exist\textsuperscript{1–5,9}. Therefore, underdiagnosis, misdiagnosis, or diagnosis delay is common in PID\textsuperscript{9,10}. Undiagnosed patients are subject to increased mortality and morbidity\textsuperscript{3,4,6}, and are associated with increased healthcare costs\textsuperscript{9}. Of note, the National Institute of Health estimates that PID may be affecting 1–2\% of the population, with recent meta-analyses suggesting that 70–90\% of PID patients remain undiagnosed even in countries with well-established diagnostic facilities\textsuperscript{10–12}.

Among primary immunodeficiencies, combined immunodeficiencies (CID) are a group of genetic disorders characterized by T-cell impairments, leading to concurrent B-cell and in some cases NK-cell defects\textsuperscript{1,9}. The most severe CID subtype, SCID, is characterized by a profound T-cell deficiency\textsuperscript{1,3–5,9}. If not treated at early infancy, SCID is fatal. Newborn screening and hematopoietic stem cell (HSCT) or bone marrow transplantation (BMT) have been established as the gold-standard in treating SCID\textsuperscript{9,13}. Other CID subtypes not marked by an absence of T-cells, are associated with variable co-morbidities, and disease progression, and are among the least investigated immune deficiencies\textsuperscript{10,13,14}. Pneumonia has been shown to be the most frequent severe infection in CID patients\textsuperscript{1,13}. CVID is characterized by B-cell defects and is the second most frequent PID (after selective IgA deficiency)\textsuperscript{1,2,10}. Currently, CID and CVID have well-established, available treatment options. HSCT and BMT are the clinical standard definitive treatments for CID\textsuperscript{1,10}. Immunoglobulin (Ig) replacement therapy is a critical therapeutic intervention in CID and CVID that reduces severe infections, end-organ damage, hospitalizations and overall morbidity and mortality. Concomitant antimicrobial therapies are frequently employed to reduce the severity of infections\textsuperscript{13,14}. Considering the underdiagnosis and the available treatments for CID/CVID, it is important to establish methodologies for screening their clinical phenotypes for which there are no other means of systematic identification.

Numerous recent studies demonstrated the merits of machine learning for the accurate analysis of electronic health records (EHRs)\textsuperscript{15–18}. To learn non-linear representations due to large heterogeneities in the CID/CVID data, we developed a wide (linear component, designed to perform cross-product transformations over sparse features) jointly trained with a deep learning (non-linear component, to learn dense embeddings in the low-dimensional space) component\textsuperscript{19,20}. Our proposed model was capable to learn both sparse features and dense embeddings via the wide and deep components respectively, and can be transferred to other diseases characterized by non-linearities due to large heterogeneities.
Our study objectives were two-fold. First, we developed and evaluated an accurate deep learning model to analyze administrative medical claims data from EHRs towards systematizing screening and early identification of CID and CVID. Second, we revealed the most important CID- and CVID-associated clinical phenotypes and their combinations, demonstrating a systematic methodology to improve early identification of these PID disorders.

Methods

Data extraction and curation

All data used for training (80%) and testing (20%) the machine learning models, were extracted from the Optum® de-identified EHR data (Optum, Inc., Eden Prairie, MN). The data were composed of medical claims containing clinical history and demographics across all participants. Four large cohorts were generated: 797, 797, 2,312, and 19,924 PID cases and equal control sizes in Cohorts 1–4, respectively (Fig. 1). This makes a total of 47,660 cases and controls. International Classification of Diseases (ICD) codes for PID identification in Cohorts 1–4 are presented in Supplementary Table 2.

Across all cohorts, the PID cases and controls were 1:1 matched for age, gender, race, ethnicity, duration of medical history (in months), and the number of healthcare visits, using propensity score (PS) matching. This led to an equal number of PID patients and PS-matched controls across cohorts. When ICD-10 or ICD-9 codes were present in the medical claims for a patient, the corresponding disease description was added in the list of clinical history and considered as co-morbidity.

Our observation time frame was from 1 January 2008 to 31 December 2021, which consisted of ~100 million USA patients in total. Data extraction, pre-processing, model training and testing of the Optum data were performed in accordance with the Declaration of Helsinki. Before model training, all the ICD-9 codes present in our data were converted to ICD-10 codes using the updated general equivalence mappings (2018 GEMS) from the https://www.cms.gov/ website. The presence or absence of ICD codes was used as features for training the machine learning models without considering the ICD temporal sequence per patient..

Cohort Generation

Figure 1 shows the study workflow, for each cohort. Cohort 1 was initially examined (discovery cohort). As we progressively moved from Cohort 2 up to 4, we aimed to evaluate model diagnostic performance by gradually increasing data heterogeneity and diversifying PID and PS-matched control settings within each cohort increment.

Since pneumonia is the most frequent severe infection in CID, we first aimed to investigate whether we can identify CID patients with pneumonia against PS-matched controls with no diagnosis of PID and pneumonia (Cohort 1). To define pneumonia in PID groups and controls, all ICD-10/-9 codes referring to pneumonia subtypes were used from the "Influenza and pneumonia" category in
We then replicated the model training by examining if we can identify CID patients with pneumonia against PS-matched random controls with no diagnosis of PID with and without pneumonia (Cohort 2). Model training was subsequently reproduced to detect CID patients against PS-matched random controls, both with and without pneumonia (Cohort 3). Finally, to increase further data heterogeneity and re-investigate whether models can accurately identify PID in diverse patient settings, we aimed to detect CID and CVID patients against PS-matched random controls, both with and without pneumonia (Cohort 4). Across all cohorts, PS-matched controls had no diagnosis of any type of CID, CVID and PID. In Cohorts 2–4, the term “with and without pneumonia” in PID patients and random controls was reflecting the pneumonia incidence in the general population.

Data Preparation And Feature Selection

For both PID patients and controls in Cohorts 1–4, patient demographics and ICD-10 / ICD-9 codes were extracted from the Optum® patient and diagnosis tables and used as input features for model training.

To create model features, an hierarchical ICD code mapping was performed using the “regexp_replace” SQL function, by sequentially combining information from the Sub Chapter, Major and Short Description levels. These levels correspond to the diagnosis category, name and description respectively, as obtained from the most updated ICD Data R package (http://cran.nexr.com/web/packages/icd/icd.pdf). In our implementation, we used the 2020 ICD-10-Clinical Modification release to account for new ICD-10 codes.

As in clinical practice a PID patient may be assigned with multiple ICD codes corresponding to general or subtypes of immunodeficiency terms; all other (than the PID ICD codes used for defining Cohorts 1–4) immunodeficiency-related features identified were removed as confounding variables, to avoid biasing model training. These confounding variables were (with ICD-10 code in parentheses): “other specified immunodeficiencies” (D84.8), “nonfamilial hypogammaglobulinemia” (D80.1), “immunodeficiency with predominantly antibody defects” (D80.9), “other immunodeficiencies” (D84.89), “immunodeficiency unspecified” (D84.9), “selective deficiency of immunoglobulin G [IgG] subclasses” (D80.3), “selective deficiency of immunoglobulin A [IgA]” (D80.2), “selective deficiency of immunoglobulin M [IgM]” (D80.4), “immunodeficiency with predominantly antibody defects unspecified” (D80.9), “antibody deficiency with near-normal immunoglobulins or with hyperimmunoglobulinemia” (D80.6), and “other immunodeficiencies with predominantly antibody defects” (D80.8).

Pre-processing

All pre-processing and machine learning model development were developed in Python 3.7 using pandas, numpy, scipy, matplotlib, GridSearchCV, scikit-learn (for classic machine learning baseline models), and PyTorch widedeep (for deep learning models). Following data preparation, the number of features (ICD codes) identified were Cohort 1: 2,188; Cohort 2: 2,154; Cohort 3: 3,522; and Cohort 4: 10,445 features. For each patient within Cohort 1–4, one-hot encoded categorical values were generated based on whether a patient was positive or negative across each diagnosis ICD code (defined as 1 and 0, respectively).
Therefore, the feature dimension “d” for each machine learning model in Cohorts 1–4 was: 2,188 × 2; 2,154 × 2; 3,522 × 2; and 10,445 × 2, respectively. For logistic regression (LR) and support vector machine (SVM), the one-hot encoded categorical values across each ICD code were used as inputs. In deep learning models, the one-hot encoded values were converted into binary value embeddings across each ICD code, using the “tab_preprocessor.fit_transform” (PyTorch widedefine library) function.

To develop an end-to-end method for future deployment in clinical settings, both deep learning models were customised to also pre-process and analyze numerical data (e.g. blood biomarkers, medical imaging biomarkers; see Supplementary Fig. 1), next to categorical data that were used for this work.

**Machine Learning Models**

We developed 2 deep learning models (both with wide and deep components): a multi-layer perceptron (MLP)-based with dense layers and an MLP model with dense layers in which we incorporated a series of ResNet blocks, named TabMLPNet and TabResNet, respectively (Supplementary Fig. 1). The wide and deep components were jointly trained. The wide (linear model) component was used to learn sparse features via cross-product transformations, whilst the deep component (deep neural network) was implemented to generalize by transforming all dense embeddings in the low-dimensional space into non-zero predictions. The pyramidal architecture of the TabMLPNet model involved 3 dense layers of 64 neurons, followed by 3 dense layers of 32 neurons. The architecture of the TabResNet model consisted of 5 dense ResNet blocks followed by an MLP structure with 4 dense layers of 100, 100, 50 and 50 neurons, respectively. In both models, each dense layer was followed by a ReLU activation. The input dimension (size) for each deep learning model was d x b, where b is the batch size. Both models were trained using a batch size of 128 across 200 epochs, with a dropout of 0.1 per dense layer. The learning rate used to perform joint training of the wide and deep components was 0.001. Joint training of the wide and deep components was performed using the Adam optimizer.

We also developed 2 baseline models: LR and SVM-based. To optimize and fine-tune both models, the GridSearchCV library was used to automatically identify their most optimum parameters. For LR, the multi-parametric space on which grid search performed was: regularization penalty (L1, L2), inverse of regularization strength (0.01, 0.1, 1, 10, 100) and class weight (balanced, none). For SVM, a radial basis function kernel was used for which the grid search was: inverse of regularization strength (0.01, 0.1, 1, 10, 100) on L2 regularization and kernel coefficient (0.001, 0.01, 0.1, 1). A cross-validation of 10 and a train-to-test split ratio of 80:20 was used for all deep and machine (baseline) learning models.

**Mapping ICD Codes Into Phenotypes**

Following machine learning model training and testing, we identified ICD codes that were associated with PID diagnoses in Cohorts 1–4. In the context of interpreting the clinical meaning of these features, we
converted features into clinical phenotypes (diseases), by using the phenome-wide associations studies (PheWAS) Phecode v.1.2 (dedicated ICD to phenotype grouping system)23.

Based on the PheWAS Phecode v.1.2, one or more ICD codes were classified into a distinct phenotype, for each patient. To perform this conversion precisely, the “regexp_replace” SQL function was used to combine information from the Short and Long Description, Major and Sub Chapter levels.

**Statistical analysis**

Statistical analyses were performed in R (R Foundation for statistical computing, Vienna, Austria). PS-matching was performed using the MatchIt library (the “glm” distance measure was used). All machine learning models were evaluated by calculating the area under the receiver-operating-characteristic (ROC) curve (AUC). We report the sensitivity, specificity, positive predictive value, negative predictive value, overall accuracy, and ROC AUC (Table 2, Supplementary Table 3).

Odds ratios (ORs) and significance levels for features and phenotypes were calculated using the glm library. Statistical significance was defined as a two-sided P value < 0.05. Temporal analysis of the top clinical phenotypes was performed using Box and Whisker plots. Tableau (Tableau 2021.4, Seattle, USA) was used for temporal analysis visualizations.

**Results**

**Participants**

The study involved 3 parts as follows: (1) Deep and machine learning models were trained and tested for the diagnosis of CID patients with pneumonia in a large medical claims dataset (Optum; discovery cohort). (2) All models were validated using 2 more CID cohorts from the same dataset, in which the pneumonia filter was removed from the controls and CID cases/controls, respectively. (3) Models were further validated in the largest and most diversified cohort generated, for the diagnosis of CID and CVID patients. All relevant diagnostic codes are listed in Supplementary Table 2.

Patient demographics are shown in Table 1. In brief, age was similar across cohorts (mean age ranging from 44–48 years). Most patients were female (53.1–62.2%) and Caucasian (82.4–88.1%). The number of healthcare visits was consistently higher in PID cases against controls.
Table 1
Baseline demographics and clinical characteristics of the 4 cohorts included in the study.

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Cohort 1</th>
<th>Cohort 2</th>
<th>Cohort 3</th>
<th>Cohort 4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PID Cases (n = 797)</td>
<td>Controls (n = 797)</td>
<td>PID Cases (n = 2,312)</td>
<td>Controls (n = 19,924)</td>
</tr>
<tr>
<td>Age and Gender</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age (years)</td>
<td>46 ± 26</td>
<td>46 ± 26</td>
<td>46 ± 26</td>
<td>48 ± 26</td>
</tr>
<tr>
<td>Male (%)</td>
<td>46.9</td>
<td>46.3</td>
<td>46.6</td>
<td>45.4</td>
</tr>
<tr>
<td>Ethnicity (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Caucasian</td>
<td>82.4</td>
<td>85.1</td>
<td>83.4</td>
<td>88.1</td>
</tr>
<tr>
<td>African American</td>
<td>8.7</td>
<td>7.3</td>
<td>8.4</td>
<td>6.7</td>
</tr>
<tr>
<td>Asian</td>
<td>1.4</td>
<td>1.5</td>
<td>1.3</td>
<td>0.5</td>
</tr>
<tr>
<td>Other/Unknown</td>
<td>7.5</td>
<td>6.1</td>
<td>7</td>
<td>4.7</td>
</tr>
<tr>
<td>Patient History</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Diagnosis History duration (years)</td>
<td>10 (8–13)</td>
<td>12 (10–14)</td>
<td>10 (8–13)</td>
<td>12 (9–14)</td>
</tr>
<tr>
<td>Number of visits</td>
<td>201 (103–399)</td>
<td>145 (48–415)</td>
<td>206 (105–399)</td>
<td>182 (45–397)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*aMedian (25th–75th percentile)

PID: primary immunodeficiency.

Machine Learning Model Performance

Initially, we investigated the diagnostic performance of our deep learning models (TabMLPNet, TabResNet) against baseline models (LR, SVM) developed, in identifying PID against PS-matched controls. All model ROC curves are illustrated in Fig. 2.

The TabMLPNet model outperformed all other models across all 4 cohorts, with ROC AUCs ranging from 0.88–0.94 (Table 2, Supplementary Table 3), showing the highest sensitivity, specificity, positive and
negative predictive value, and overall accuracy across all cohorts, ranging from: 0.82–0.88, 0.82–0.85, 0.80–0.87, 0.81–0.91 and 0.80–0.87, respectively.

Table 2

**Diagnostic performance of the TabMLPNet model.** The TabMLPNet model showed the highest diagnostic performance across all 4 cohorts investigated.

<table>
<thead>
<tr>
<th>Metrics for TabMLPNet</th>
<th>Cohort 1</th>
<th>Cohort 2</th>
<th>Cohort 3</th>
<th>Cohort 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sensitivity</td>
<td>0.88</td>
<td>0.87</td>
<td>0.82</td>
<td>0.87</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.85</td>
<td>0.84</td>
<td>0.82</td>
<td>0.82</td>
</tr>
<tr>
<td>PPV</td>
<td>0.87</td>
<td>0.87</td>
<td>0.80</td>
<td>0.87</td>
</tr>
<tr>
<td>NPV</td>
<td>0.87</td>
<td>0.91</td>
<td>0.81</td>
<td>0.83</td>
</tr>
<tr>
<td>Accuracy</td>
<td>0.87</td>
<td>0.87</td>
<td>0.80</td>
<td>0.85</td>
</tr>
<tr>
<td>ROC AUC</td>
<td>0.94</td>
<td>0.93</td>
<td>0.88</td>
<td>0.91</td>
</tr>
</tbody>
</table>

All other models showed consistently high diagnostic performance in identifying PID against PS-matched controls (ROC AUC range = 0.84–0.93, Supplementary Table 3). The ROC AUC for TabMLPNet and TabResNet were significantly higher compared to LR and SVM in the largest Cohorts 3 and 4 (Supplementary Table 3; P < 0.05). No other significant differences were observed between ROC curves.

**Clinical Phenotype Importance**

The second aim of the study was to identify the most important CID- and CVID-associated clinical phenotypes per cohort. Diagnostic codes were converted into clinical phenotypes and their ORs were calculated.

In Cohorts 1–4, the OR of the top clinical phenotypes ranged from 14.91–1.85, 14.24–1.83, 23.97–1.95, and 11.12–1.66, respectively (Fig. 3; all P ≤ 10^{-3}). For Cohorts 1–3, the top twenty phenotypes are presented (Fig. 3a-c). For the largest Cohort 4, the top 35 phenotypes are shown, reflecting a greater number of phenotypes reaching high statistical significance (Fig. 3d; P ≤ 10^{-3}). A full list of all phenotype ORs, prevalence and statistical significance across cohorts is given in the Supplementary Data.

Several phenotypes were revealed in Cohorts 1–4 (Fig. 3). In Cohort 1, genetic carrier/susceptibility to disease, pneumococcal pneumonia, short stature, valvular heart disease and alveolar/parietoalveolar pneumonopathy were the 5 strongest phenotypes (Fig. 3a). In Cohort 2, autoimmune disease not-elsewhere-classified (NEC), valvular heart disease, chromosomal anomalies, myopathy and non-Hodgkin's
lymphoma were the 5 top phenotypes (Fig. 3b). In Cohort 3, deficiencies of circulating enzymes, autoimmune disease NEC, bone marrow/stem cell transplant, genetic carrier/susceptibility to disease and disorders of purine/pyrimidine metabolism were the top phenotypes (Fig. 3c). In the largest Cohort 4, chromosomal anomalies, disorders of purine/pyrimidine metabolism, lymphoid leukemia chronic, deficiencies of circulating enzymes and bronchiectasis were the strongest phenotypes (Fig. 3d).

Across all cohorts, various other genetic, respiratory, autoimmune, musculoskeletal, blood and blood cancer diseases were revealed (Fig. 3; Supplementary Results). The ORs of the underlying diagnostic codes across all cohorts were also computed and are presented in the Supplementary Fig. 2.

**Temporal Distributions**

We derived the temporal distributions of the 25 most important clinical phenotypes, by tracking the first date of each phenotype diagnosis with reference to PID diagnosis per patient (Fig. 4, Supplementary Results).

Most phenotypes preceded PID diagnoses across all cohorts. The number of phenotypes that had a median of first diagnosis greater than 3 months before PID diagnosis in Cohorts 1–4 were: 16, 21, 15 and 15, respectively. Their median value range in months before PID diagnosis were: 34.1–3.4, 36.5-4.0, 32.4–3.1, and 29.3–3.2, respectively. At a threshold of 6 months before PID diagnosis, the phenotype numbers were 12, 17, 9, and 9 in Cohorts 1–4, respectively.

**Time Frame Of Diagnoses Prior To Pneumonia**

In Cohort 1, 20-year time frames of ICD codes and clinical phenotypes in CID cases against controls were computed, prior to (-10 years) and after (+10 years) the diagnosis of pneumonia, used as a common feature between cases and controls (Supplementary Figs. 3–4). The 20-year time frames depict the cumulative proportion of patients with each phenotype, which equals the sum of the proportions from each of the years preceding or following pneumonia diagnosis. It is obvious that most ICD codes and phenotypes started being diagnosed before the first pneumonia diagnosis in CID cases against controls.

**Combinations**

Further, we estimated ORs of combined phenotypes associated with PID (Table 3). Various heterogeneous combinations were revealed, mostly consisting of antecedent phenotypes with a median of first diagnosis at least 6 months before PID diagnosis (Table 3 available in the Supplemental Files, Fig. 4).

In Cohorts 1–4, the top combinations consisted of antecedent phenotypes with a median of first diagnosis at least 6 months before PID diagnosis were: disorders involving the IM + decreased white BCC + asthma (OR = 6.53, 95% CI: 2.22–8.75); non-Hodgkin's lymphoma + pneumonia + fever of unknown origin (OR = 6.96, 95% CI: 3.76–10.20); bone marrow/stem cell transplant + disorders involving the IM + asthma (OR =
6.83, 95% CI: 4.22–9.44); psoriatic arthropathy + autoimmune disease NEC + asthma (OR = 6.25, 95% CI: 4.73–7.77), respectively.

**Discussion**

We have performed a large-scale analysis of medical claims data derived from a nationally representative EHR database, by devising a deep learning-based methodology. Our method showed consistently high diagnostic performance in identifying CID/CVID across 4 cohorts with clinically diverse patient profiles. Furthermore, we identified the top antecedent phenotypes associated with these PIDs. We also revealed the top phenotype combinations for each cohort and showed that they consist mostly of antecedent heterogeneous diseases.

To the best of our knowledge, we were the first to interrogate large EHR data for the early diagnosis of CID/CVID and the identification of antecedent phenotype combinations through deep learning and OR calculations. Moreover, none of the previous studies focused on CID\(^24\–27\). The most recent study developed an LR model using 6,422 patients, of whom 247 were diagnosed with PID\(^24\). By modeling co-morbidities, they showed moderate ROC AUC (0.62) which was improved (0.70) when laboratory results and radiology procedures were considered. Rider et al developed a Bayesian network and showed 87% and 91% sensitivity and specificity in discriminating PID against controls, using 3,460 pediatric patients (~50% with PIDs)\(^25\). Abyazi et al, identified different proteomic profiles in patients with noninfectious complications against uncomplicated CVID, implementing unsupervised learning in 72 participants\(^26\). Emmaneel et al, developed a computational pipeline to discriminate CVID from other PIDs and healthy controls, using flow cytometry data from 179 participants\(^27\). Unlike our work, the last 2 studies focused to study differences in the molecular profiles of PID patients and did not aim to improve screening and early PID diagnosis in the frontline of clinical practice. By improving early PID diagnosis via population-based screening, it is possible to significantly reduce co-morbidities, mortalities, healthcare visits and costs, through timely patient access to available definitive and supportive treatments for both CID and CVID. Our method reached high diagnostic performance using large EHR data and revealed phenotypes and combinations that can have merit for the early diagnosis of CID and CVID.

Our findings are clinically important because our predictive scheme identified disease combinations, which are the first time to be reported for the early diagnosis of PID\(^1,2,10,13\). In particular, there were distinctive combinations of antecedent (>6 months) phenotypes such as respiratory infections or conditions (asthma, pneumonia, bronchiectasis), genetic anomalies (genetic carrier/susceptibility to disease, lack of normal PD, chromosomal anomalies), cardiac defects, autoimmune diseases (psoriatic arthropathy, autoimmune disease NEC, celiac disease, disorders involving the immune mechanism), blood disorders and malignancies (non-Hodgkin's lymphoma), associated with both CID and CVID (Table 3, Figs. 3–4). Since Cohort 4 involves both CID and CVID, the phenotype combinations revealed can increase early suspicion of potential PID before further categorization to CID or CVID. Deploying our proposed method for clinical EHRs, these respiratory and non-respiratory combinations can potentially expand the existing clinical warning signs and systematize the early diagnosis of CID/CVID.
CID is characterized by complex immune defects and are among the least investigated PIDs\textsuperscript{10,13,14}. Since pneumonia is their most frequent severe infection\textsuperscript{1,28}, we first aimed to investigate pneumonia phenotype patterns when discriminating CID against controls, both with pneumonia (Cohort 1). Among the top phenotypes, we identified pneumococcal and the broader bacterial pneumonia subtypes (Fig. 3). In contrast to the general pneumonia phenotype, the above pneumonia subtypes did not significantly precede PID diagnosis (Fig. 4). This might reflect lack of pneumonia categorization early in the CID spectrum. In Cohorts 2–4, the pneumonia filter was gradually removed from controls and patients, to simulate the pneumonia incidence in the general population. In Cohort 2, there were similar pneumonia findings to Cohort 1, with non-pneumonia diseases dominating in Cohorts 3–4.

Most individual respiratory, genetic, autoimmune, blood and malignancy phenotypes revealed across cohorts are reported in the literature and recent PID surveys (Fig. 3)\textsuperscript{1–3,10–14}. Numerous congruent phenotypes were identified across Cohorts 1–4. Of note, there were also unknown individual phenotypes emerged, such as asthma (in Cohorts 1–4), coagulation defects complicating pregnancy or postpartum (Cohorts 3–4) and cancer of lymphoid histiocytic tissue (Cohort 4)\textsuperscript{2,13}. Among the most severe early manifestations, lymphoid leukemia chronic was the third most important antecedent phenotype in Cohort 4 (Figs. 3–4). Despite hematologic malignancies being known to be associated with PIDs, there is low awareness of lymphoid leukemia chronic in PID patients\textsuperscript{29}. Moreover, hypoparathyroidism (Cohorts 1–2), autoimmune diseases NEC (Cohorts 2–4), systemic lupus erythematosus (Cohort 3), psoriatic arthropathy (Cohorts 3–4), rheumatoid arthritis (Cohort 2) and celiac disease (Cohorts 3–4) were the top antecedent autoimmune conditions associated with CID/CVID. Our findings can therefore raise awareness and support treatment optimization strategies for co-treating early both the underlying disorder (CID/CVID) and each of these respiratory, oncological and endocrinological diseases.

To summarize, this study has several strengths making it clinically relevant to the medical community. First, we evaluated different perspectives of CID and CVID, by developing a generalized and accurate method and a comprehensive evaluation procedure across all 4 (CID/CVID) cohorts. Second, we developed a deep learning model that can learn non-linearities due to large heterogeneities by modeling both sparse (rare) features and dense (non-rare) embeddings present in the data. Hence, our deep learning method can be transferred to other diseases characterized by clinical heterogeneities, including other PID. Third, we evaluated our model on a large cohort of US patients (N = 47,660 participants). Fourth, our method is versatile, as it is designed to co-analyze blood sample, genetic and imaging biomarkers as continuous variables when available (Supplementary Fig. 1). Fifth, since our analytical approach is based on the conversion of ICD codes to explicit clinical phenotypes, it has broad applicability for systematizing early diagnosis of other underdiagnosed heterogeneous diseases.

Several limitations should be considered when interpreting our findings. The main limitation is reliance on ICD codes from EHR hence, cases and controls could sometimes be miscoded. However, model-derived PID identification was consistent across cohorts and the combinations revealed consisted of antecedent phenotypes that are widely reported in the literature\textsuperscript{1–3,10–14}. Clinical history might be misinformed because of differences across regional, institutional, or individual ICD coding processes. Converting into
phenotypes may have minimized any biases from miscoded disease subtypes. Although we aimed to investigate diverse PID profiles across cohorts, we consistently identified congruent phenotypes between Cohorts 1–3 (CID) and Cohort 4 (CVID), which reflects the identification of consistent patterns in the presence of PID. Our dataset did not involve laboratory, genetic and imaging data, which could further enhance the diagnostic performance, clinical information, and result interpretations. Our goal is to deploy our method on external clinical EHR data thus, investigating multi-modal data analyses is our main future endeavor.

In conclusion, the proposed scheme achieved accurate and generalizable performance for the early diagnosis of CID and CVID. Our methodology can lead to new clinical insights and expanded guidelines for the early identification of phenotype combinations, increase clinical awareness and be used to improve PID diagnosis and outcomes on a population level.

**Declarations**

**Data availability statement**

The datasets used for this study could not be made publicly available due to a data use commercial agreement between Pfizer and Optum. However, the authors encourage collaborations and would like to declare that the data can be made available to qualified investigators upon request with evidence of institutional review board approval.

**Code availability**

The underlying code for this study [and training/validation datasets] is not publicly available for proprietary reasons but can be made available to qualified investigators upon request.

**Acknowledgements**

This study was sponsored by Pfizer. The sponsor played no role in study design, data collection, analysis and interpretation of data, or the writing of this manuscript. Medical writing support was provided by Alex Frings, PharmD, of Engage Scientific Solutions (Fairfield, CT) and funded by Pfizer.

**Author contributions**

GP: Author of the manuscript. Conceptualized the objectives, devised the machine learning models, performed statistical analysis, prepared the data analysis software and interpreted the findings of the manuscript. GY, DF, ND, CW and AH contributed in the methodology and validation of the machine learning models and statistical modeling, and edited the manuscript. LS, MS and GS contributed in the study design and approved the validity of the research questions. DP contributed in the study design, project administration, approved the validity of the research questions and acquired funding. All authors contributed in the interpretation of the data, read and critically revised the manuscript and approved the final version of the manuscript.
Competing Interests statement

This study was supported by Pfizer. The authors would like to declare that the methods and outcomes of this work are based on scientific research principles and there is no conflict-of-interest due to Pfizer's commercial interests. GP, AH, GS, and DP are full-time employees of Pfizer and hold stock/stock options. LS is a full-time employee of Immunoglobulin National Society. GY, DF, ND, and CW do not have any financial competing interests to declare. The authors do not have any non-financial competing interests to declare.

Funding

This study was sponsored by Pfizer.

References


**Figures**
**Figure 1**

**Study workflow.** CID: Combined immunodeficiency; CVID: Common variable immunodeficiency.
Figure 2

Receiver operating characteristic curves for all machine learning models developed and evaluated in the four cohorts. a) PID patients with pneumonia against pneumonia patients without PID. b) PID patients with pneumonia against randomly selected patients without PID. c) PID patients with and without pneumonia against randomly selected patients without PID. d) All PID patients (combined and common variable PID patients) with and without pneumonia against randomly selected patients without PID. PID: Primary immunodeficiency.
Figure 3

Odds ratio (OR, blue) and prevalence (% red) for the top clinical phenotypes associated with PID, identified in Cohorts 1-4 (a-d, respectively). All clinical phenotypes associated with the diagnosis of PID at a significance level of $p \leq 10^{-3}$ which had an OR > 1.5 were included in the illustrations. Univariate logistic regression was used to calculate the odds ratios. PD: Physiological development, UNS: Unspecified, NEC: Not elsewhere classified, F: Family, P: Personal, CC: Certain conditions, LH: Lymphoid hematopoietic, C1-C4: Declare congruent phenotypes in Cohorts 1-4.
Figure 4

Temporal analysis of the top 25 clinical phenotypes identified in Cohorts 1-4 (by odds ratio). Temporal analysis was estimated in terms of Box and Whisker plots. The red line within Box and Whisker plots represents the median. The blue line illustrates the median of the pneumonia diagnosis when present in the data. The Box and Whisker plots are calculated in reference to the PID diagnosis showed as black solid line in each Cohort illustration. C1-C4: Declare congruent phenotypes in Cohorts 1-4. The top 25 clinical phenotypes associated with the diagnosis of PID at a significance level of $p \leq 10^{-3}$ were included in the illustrations.
Development and evaluation of a machine learning model pipeline for the early and systematic identification of PID (CID and CVID). US nationally representative medical claims were used to develop a cohort of combined immunodeficiency (CID) patients and non-CID controls (both with pneumonia, Cohort 1). Diagnosis (ICD) codes were extracted and used as variables to train the TabMLPNet model. Subsequently, the same methodology and model were internally tested in 3 replication cohorts (Cohorts 2-
4). To derive clinical insights for the early diagnosis of PID across all cohorts, the ICD codes were then converted to clinical phenotypes and odds ratios were calculated to estimate hierarchical clinical phenotype importance (Cohorts 1-4). Further, phenotype temporal profiles and combinations were extracted and assessed in terms of their associations with PID. Clinical phenotypes will be used to enrich the diagnostic criteria for the early PID detection, including expanding the existing clinical warning signs and improving patient outcomes on population level. In a future clinical setting, additional clinical validation with blood and genetic tests, as well as medical imaging diagnosis (lung CT, cardiovascular CT or MRI, anatomical or functional CT, MRI or ultrasound) can be implemented to automate the PID (CID and CVID) diagnosis, aiming to ultimately establish a robust clinical method for the early and accurate stratification of patients with PID. EHR: Electronic Health Records, CID: Combined immunodeficiency, CVID: Common variable immunodeficiency, PID: Primary immunodeficiency, CT: Computed tomography, MRI: Magnetic resonance imaging.

**Supplementary Files**

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

- [SupplementaryMaterials.docx](#)
- [Table3.pdf](#)