

A Machine Learning Approach for Identifying People with Neuroinfectious Diseases in Electronic Health Records: Algorithm Development and Validation

Arjun Singh, Shadi Sartipi, Haoqi Sun, Rebecca Milde, Niels Turley, Carson Quinn, G Kyle Harrold, Rebecca L. Gillani, Sarah E. Turbett, Sudeshna Das, Sahar Zafar, Marta Fernandes, M Brandon Westover, Shibani S Mukerji

Submitted to: JMIR Medical Informatics
on: June 12, 2024

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on its website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressly prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Original Manuscript.....	5
Supplementary Files.....	35
Figures	36
Figure 1.....	37
Figure 2.....	38
Figure 3.....	39
Multimedia Appendixes	40
Multimedia Appendix 1.....	41
Multimedia Appendix 2.....	41
Multimedia Appendix 3.....	41
Multimedia Appendix 4.....	41
Multimedia Appendix 5.....	41

A Machine Learning Approach for Identifying People with Neuroinfectious Diseases in Electronic Health Records: Algorithm Development and Validation

Arjun Singh^{1, 2*} MBBS; Shadi Sartipi^{1, 3*} PhD; Haoqi Sun² PhD; Rebecca Milde⁴; Niels Turley⁴; Carson Quinn^{1, 5} MD; G Kyle Harrold^{1, 2, 5} MD; Rebecca L. Gillani¹ MD, PhD; Sarah E. Turbett^{6, 7} MD; Sudeshna Das^{1, 2} PhD; Sahar Zafar^{1, 2} MD; Marta Fernandes^{1, 2*} PhD; M Brandon Westover^{1, 2, 4} MD, PhD; Shibani S Mukerji^{1, 2*} MD, PhD

¹ Department of Neurology Massachusetts General Hospital Boston US

² Harvard Medical School Boston US

³ Harvard Medical School Boston-a US

⁴ Department of Neurology Beth Israel Deaconess Medical Center Boston US

⁵ Department of Neurology Brigham and Women's Hospital Boston US

⁶ Department of Medicine Massachusetts General Hospital Boston US

⁷ Department of Pathology Massachusetts General Hospital Boston US

*these authors contributed equally

Corresponding Author:

Shibani S Mukerji MD, PhD

Department of Neurology
Massachusetts General Hospital
Boston
US

Abstract

Background: Identifying NID cases using International Classification of Diseases (ICD) billing codes is often imprecise, while manual chart reviews are labor-intensive. ML models can leverage unstructured electronic health records (EHR) to detect subtle NID indicators, process large data volumes efficiently, and reduce misclassification. While accurate NID classification is needed for research and clinical decision support, using unstructured notes for this purpose remains underexplored.

Objective: To develop and validate a machine learning (ML) model to identify neuroinfectious diseases (NID) from unstructured patient notes.

Methods: Clinical notes from patients who had undergone lumbar puncture were obtained using the EHR of an academic hospital network (Mass General Brigham, MGB), with half associated with NID-related diagnostic codes. Ground-truth was established by chart review with six NID-expert physicians. NID keywords were generated with regular expressions, and extracted texts were converted into bag-of-words representations using n-grams (n=1, 2, 3). Notes were randomly split into training (80%) and hold-out testing (20%) sets. Feature selection was performed using logistic regression with L1 regularization. An extreme gradient boosting (XGBoost) model classified NID cases, and performance was evaluated using the Area Under the Receiver Operating Curve (AUROC) and the Precision-Recall Curve (AUPRC). Performance of the natural language processing (NLP) model was contrasted with LLaMA 3.2 auto-regressive model on the MGB test set. The NLP model was additionally validated on external data from an independent hospital (Beth Israel Deaconess Medical Center, BIDMC).

Results: This study included 3,000 patient notes from MGB from January 22, 2010, to September 21, 2023. Of 1,284 initial n-gram features, 342 were selected, with the most significant features being 'meningitis,' 'ventriculitis,' and 'meningoencephalitis.' The XGBoost model achieved an AUROC of 0.98 (95% CI: 0.96 - 0.99) and AUPRC of 0.89 (95% CI: 0.83 - 0.94) on MGB test data. In comparison, NID identification using ICD-billing codes showed high sensitivity (0.97) but poor specificity (0.59), overestimating NID cases. LLaMA 3.2 improved specificity (0.94) but had low sensitivity (0.64) and an AUROC of 0.80. In contrast, our NLP model balanced specificity (0.96) and sensitivity (0.84), outperforming both methods in accuracy and reliability on MGB data. When tested on external data from BIDMC, the NLP model maintained an AUROC of 0.98 (95% CI: 0.96 - 0.99), with an AUPRC of 0.78 (95% CI: 0.66 - 0.89).

Conclusions: The NLP model accurately identifies NID cases from clinical notes. Validated across two independent hospital datasets, the model demonstrates feasibility for large-scale NID research and cohort generation. With further external validation, our results could be more generalizable to other institutions.

(JMIR Preprints 12/06/2024:63157)

DOI: <https://doi.org/10.2196/preprints.63157>

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**

Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.

Only make the preprint title and abstract visible.

No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain visible.

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <http://www.jmir.org>

No. Please do not make my accepted manuscript PDF available to anyone.

Original Manuscript

A Machine Learning Approach for Identifying People with Neuroinfectious Diseases in Electronic Health Records: Algorithm Development and Validation

Arjun Singh, MD^{1,2*}, Shadi Sartipi, PhD^{1,2*}, Haoqi Sun, PhD^{2,3}, Rebecca Milde^{3,4}, Niels Turley³, Carson Quinn, MD^{1,2,5}, G Kyle Harrold, MD^{1,2,5}, Rebecca L. Gillani, MD, PhD^{1,2}, Sarah E. Turbett, MD^{2,6,7}, Sudeshna Das, PhD^{1,2}, Sahar Zafar, MD^{1,2}, Marta Fernandes, PhD^{1,2,**}, M Brandon Westover, MD, PhD^{1,2,3,**}, Shibani S Mukerji, MD, PhD^{1,2,**}

*These authors contributed equally as co-first authors.

**These authors contributed equally as co-senior authors.

¹ Department of Neurology, Massachusetts General Hospital, Boston, MA 02114, USA

² Harvard Medical School, Boston, MA 02115, USA

³ Department of Neurology, Beth Israel Deaconess Medical Center, Boston, MA 02215, USA

⁴ Northeastern University, Boston, MA 02115, USA

⁵ Department of Neurology, Brigham and Women's Hospital, Boston, MA 02115, USA

⁶ Department of Medicine, Massachusetts General Hospital, Boston, MA 02114, USA

⁷ Department of Pathology, Massachusetts General Hospital, Boston, MA 02114, USA

Corresponding Author:

Shibani S. Mukerji, MD, PhD

Address: 114 13th St, Mukerji Lab for Neuroinfectious Disease, 2nd Floor JJ

Charlestown, MA 02129-2020

Phone: 857-282-9950

Email: smukerji@mgb.org

ABSTRACT

Objective: To develop and validate a machine learning (ML) model to identify neuroinfectious diseases (NID) from unstructured patient notes.

Background: Identifying NID cases using International Classification of Diseases (ICD) billing codes is often imprecise, while manual chart reviews are labor-intensive. ML models can leverage unstructured electronic health records (EHR) to detect subtle NID indicators, process large data volumes efficiently, and reduce misclassification. While accurate NID classification is needed for research and clinical decision support, using unstructured notes for this purpose remains underexplored.

Methods: Clinical notes from patients who had undergone lumbar puncture were obtained using the EHR of an academic hospital network (Mass General Brigham, MGB), with half associated with NID-related diagnostic codes. Ground-truth was established by chart review with six NID-expert physicians. NID keywords were generated with regular expressions, and extracted texts were converted into bag-of-words representations using n-grams (n=1, 2, 3). Notes were randomly split into training (80%), 2400 notes out of 3000, and hold-out testing (20%), 600 notes out of 3000, sets. Feature selection was performed using logistic regression with L1 regularization. An extreme gradient boosting (XGBoost) model classified NID cases, and performance was evaluated using the Area Under the Receiver Operating Curve (AUROC) and the Precision-Recall Curve (AUPRC). Performance of the natural language processing (NLP) model was contrasted with Llama 3.2 autoregressive model on the MGB test set. The NLP model was additionally validated on external data from an independent hospital (Beth Israel Deaconess Medical Center, BIDMC).

Results: This study included 3,000 patient notes from MGB from January 22, 2010, to September 21, 2023. Of 1,284 initial n-gram features, 342 were selected, with the most significant features being 'meningitis,' 'ventriculitis,' and 'meningoencephalitis.' The XGBoost model achieved an

AUROC of 0.98 (95% CI: 0.96 - 0.99) and AUPRC of 0.89 (95% CI: 0.83 - 0.94) on MGB test data. In comparison, NID identification using ICD-billing codes showed high sensitivity (0.97) but poor specificity (0.59), overestimating NID cases. Llama 3.2 improved specificity (0.94) but had low sensitivity (0.64) and an AUROC of 0.80. In contrast, our NLP model balanced specificity (0.96) and sensitivity (0.84), outperforming both methods in accuracy and reliability on MGB data. When tested on external data from BIDMC, the NLP model maintained an AUROC of 0.98 (95% CI: 0.96 - 0.99), with an AUPRC of 0.78 (95% CI: 0.66 - 0.89).

Conclusions: The NLP model accurately identifies NID cases from clinical notes. Validated across two independent hospital datasets, the model demonstrates feasibility for large-scale NID research and cohort generation. With further external validation, our results could be more generalizable to other institutions.

Keywords: Neurology; Neuroinfectious Disease; Electronic Health Records (EHR); Machine Learning

INTRODUCTION

Meningitis and encephalitis pose serious threats to the health of an individual, potentially leading to severe neurological compromise or even death [1–7]. Many neuroinvasive pathogens, including viruses (e.g., enterovirus, herpes simplex virus, West Nile virus, HIV), bacteria (e.g., *S. pneumoniae*, *Mycobacterium tuberculosis*), and fungi (e.g., *Cryptococcus*), are linked to long-term cognitive sequelae [2,8]. Recent population-based studies suggest associations between pathogen exposures, especially neuroinvasive viruses causing encephalitis or meningitis, and the subsequent risk of developing dementia such as Alzheimer's disease (AD) [9–12]. The risk of AD appears to increase with longer time since the first infection, peaking after 12-30 years [9]. Prior viral encephalitis and bacterial or viral meningitis, were found to be associated with significantly increased hazard ratios for future AD, with rates of 30.72-fold and 2.81-fold respectively, based on the Finnish Biobank, suggesting that exposure to neuroinvasive pathogens may contribute to lower cognitive reserve, likely in part due to brain inflammation [10].

These findings underscore the need for comprehensive, longitudinal studies to elucidate the mechanisms by which neuroinvasive pathogens might contribute to neurodegenerative diseases. However, such research is severely hampered by the scarcity of large, well-annotated hospital datasets. This data gap not only impedes rigorous epidemiological analyses but also hinders mechanistic investigations that could reveal how pathogens interact with neural tissues to potentially trigger or accelerate neurodegenerative processes. Therefore, to support evidence-based clinical decision-making and enable large-scale, high-powered research into the long-term consequences of NIDs, accurately identifying these conditions in hospital patient records remains a significant challenge.

NIDs stem from a diverse array of pathogens and often present with clinically indistinguishable symptoms such as fever, headache, and confusion. Physicians often grapple with challenges in

promptly identifying these pathogens, resorting to less precise shotgun testing methods, leading to delayed or unconfirmed diagnoses, unnecessary antimicrobial treatments, and preventable morbidity [13–18]. Most epidemiological studies linking infectious disease burden to subsequent dementia rely on healthcare databases that use International Classification of Diseases (ICD) codes 9th (ICD-9) and 10th (ICD-10) revisions (combined as ICD codes) and do not validate diagnoses with microbial assays [19]. The limited reliability of ICD codes in accurately identifying infectious diseases has been previously reported [20–23], more recently underscored by the 56-57% positive predictive value observed in the Danish National Patient Registry investigating herpes simplex virus encephalitis [19]. The combination of ICD miscoding and the potential misclassification of encephalitis etiology poses significant obstacles to generating actionable recommendations for preventing AD and other dementias. This challenge further hinders clinical guidance, particularly in cases of viral encephalitis with unknown origins—a subgroup identified in epidemiological studies as having the highest risk of future dementia [10,11].

Recent efforts to move beyond ICD-based coding have aimed at improving the timely and accurate diagnosis of NID etiologies, particularly in differentiating between viral and bacterial meningitis. These efforts predominantly employ approaches such as decision trees though ensemble methods. For example, a study utilizing a Brazilian dataset of 12,774 patients described by 19 clinical attributes aimed at distinguishing bacterial from aseptic meningitis found that a hybrid bagging approach with Naive Bayes Trees outperformed 27 other tested models. This model achieved a sensitivity of 96% and a specificity of 82% [24]. While promising, most other classification studies on NIDs, employing techniques like logistic regression [25–27] or hybrid models [24,27–32] are primarily in pediatric populations, lack external validation, or rely on small datasets with limited cross-validation [33]. To our knowledge, no study has utilized unstructured data such as clinical notes, which limits the insights that could be derived.

The 2015 global research priorities for infections that affect the nervous system underscored the need

for accurate disease burden estimates and improved tools for neurological and cognitive impairment assessment [2]. Natural language processing (NLP) tools can leverage the rich information present in large-scale comprehensive electronic health records (EHRs), including unstructured clinical assessments and notes [34–36]. Therefore, NLP holds the potential to surpass the accuracy of traditional ICD-based phenotyping while reducing reliance on semi-structured data formats [30,35,37,38].

When structured lab data is unavailable, NLP techniques leveraging unstructured clinical notes can potentially generate more accurate NID cohorts than ICD codes alone. Many NLP algorithms are computationally lighter and require less computing infrastructure compared to more recent large language models (LLMs), making them more feasible for widespread research applications. Building on this advantage, our goal is to develop a more accurate and efficient method for classifying NID cases in EHRs. Previous studies have successfully used NLP to identify patients with epilepsy [38] or classify neurological outcomes from medical notes [39].

Here, we aim to tackle NID patients by building upon these studies and detail our NLP algorithm leveraging discharge summaries and progress notes to identify NID patients and investigate the model's accuracy across two independent hospitals. Our NLP algorithm outperforms ICD codes in identifying NID patients and achieves competitive performance compared to Llama 3.2 autoregressive model (an LLM with 3B parameters) in zero-shot learning tasks, making it a valuable tool for large-scale EHR-based research to investigate the relationship between NID exposure and short- and long-term neurological outcomes.

METHODS

Figure 1 provides a summary overview of the key steps involved in this study, including data extraction, note selection, regular expression development, ground-truth labeling, text preprocessing,

feature extraction, model training, and evaluation.

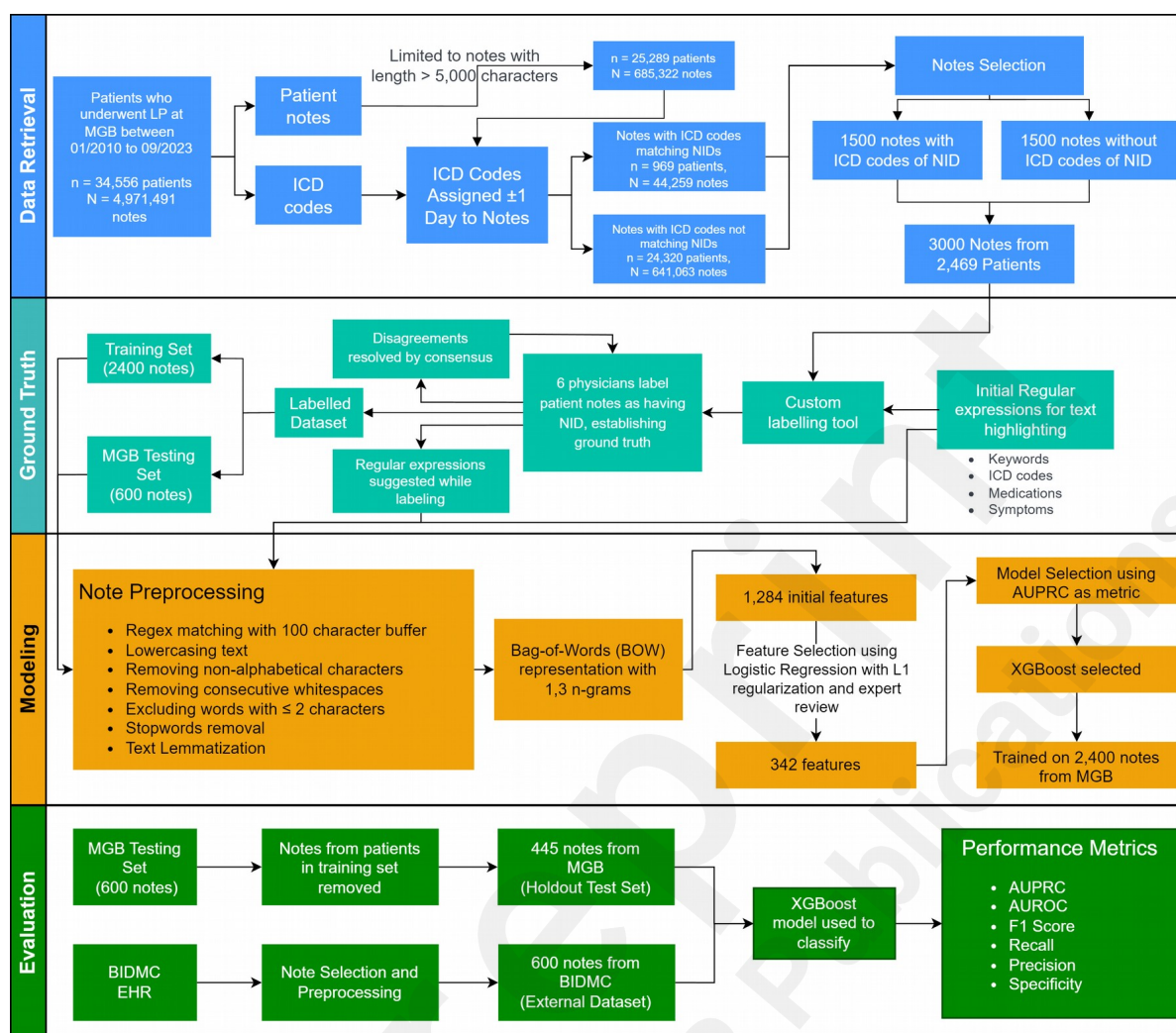


Figure 1. Methods Overview: Data Retrieval and Machine Learning Pipeline for NID Classification.

Abbreviations: AUPRC (Area Under the Precision-Recall Curve), AUROC (Area Under the Receiver Operating Characteristic Curve), BIDMC (Beth Israel Deaconess Medical Center), BOW (Bag-of-Words), EHR (Electronic Health Records), ICD (International Classification of Diseases), LP (Lumbar Puncture), n (number of unique patients), N (number of unique notes), MGB (Mass General Brigham), NID (NeuroInfectious Diseases).

Data Extraction and Note Selection

We included 34,556 patients who underwent lumbar punctures (LP), sourcing medical notes from the

EHR database of the Mass General Brigham (MGB) network of hospitals. This approach enriched individuals likely to be evaluated for CNS infections, with a total of 4,971,491 notes extracted between 01/22/2010 and 09/21/2023. ICD diagnostic codes billed one calendar day before and after the notes were extracted and labeled as likely to be associated with NID. Multimedia Appendix 1 lists the NIDX ICD 9 and 10 codes.

We excluded notes with fewer than 5,000 characters, as they likely lacked the necessary detail for clinicians to accurately assess patient status or provide valid training for models. We categorized notes into two groups: (1) notes with NID-related ICD codes (969 unique patients, 44,259 notes) and (2) those without NID ICD codes (24,320 unique patients, 641,063 notes). From these two groups, we randomly selected 3,000 notes from 2,469 patients at MGB, of which 50%, 1500 out of 3000, were associated with NID-related ICD codes, and 50% were not. For external validation, we also extracted 600 inpatient notes from patients in the Beth Israel Deaconess Medical Center (BIDMC) EHR system located in the Brain Data Science Platform (bdsp.io) maintaining a 50:50 split of cases with and without NID-related ICD codes to match the MGB test set size.

Ethics Approval

EHR data was extracted under protocol (#2013P001024) approved by Institutional Review Board under a waiver of informed consent. No features in manuscript or supplementary material are identifiable.

Ground-Truth Labeling

39,245 regular expressions were formulated based on domain-specific knowledge, categorized as (1) Positive NID (e.g., 'Cerebrospinal fluid (CSF) positive for EBV via PCR', 'ID felt encephalitis and seizures triggered by a viral illness', 'Currently on empiric antibiotics and acyclovir for possible CNS infection'); (2) Negative NID (e.g., 'LPs have been negative', 'no signs of any infectious', 'autoimmune epilepsy/encephalitis'); (3) NID drugs (e.g., ceftriaxone, acyclovir, trimethoprim-

sulfamethoxazole, etc.); (4) NID-likely keywords (e.g., ‘headache’, ‘fever’, ‘lumbar puncture’ etc.). Regular expressions were developed by listing relevant organisms based on source materials for infectious diagnoses, antimicrobial agents, and symptoms associated with meningitis and encephalitis. Many terms were combined to create comprehensive positive and negative expressions. In addition to domain expert input, source materials included national guidelines related to central nervous system (CNS) diseases in adults and pathogen data from global burden of disease studies [4,40–43]. The positive, negative and antimicrobial regular expressions are provided in Multimedia Appendix 2-4.

Six physicians (AS, CQ, KH, GH, ST, SM), all domain experts in neuroimmunology or infectious diseases, independently classified 500 notes each from the MGB dataset (3,000 notes in total) following a standardized operating procedure. This procedure provided detailed guidelines for reviewing notes including assessing symptoms and evaluate laboratory tests when referenced in the notes. Ambiguous cases flagged by the physicians were reviewed by an independent physician who was not involved in the initial classification. A consensus approach was used to resolve discrepancies in classification. To facilitate labeling, a labeling software tool was created to show the notes while highlighting the keywords (e.g., *meningitis*, *encephalitis*, *ceftriaxone*, etc.) and phrases (e.g., ‘*Meningitis, likely bacterial*’, ‘*CSF pleocytosis noted*’). Domain experts had the option to recommend additional regular expressions during the MGB ground-truth labeling phase. Regular expressions proposed by domain experts were incorporated if proposed on the training dataset only. External model validation on 600 BIDMC notes were classified by two experts (AS, SM) (300 notes each). While the MGB test set was reduced to 445 notes to address overlap with the training set, the BIDMC dataset was solely used for external validation and retained its full size of 600 notes.

Preparing Text Data for Feature Extraction

Preprocessing notes to identify model features required: (1) limiting text to regions matching regular

expressions with a 100-character buffer, (2) converting to lowercase letters, (3) removal of non-alphabetical characters, (4) replacing consecutive whitespaces with a single whitespace, (5) using words with more than 2 characters, and (6) removing stop words (*see Multimedia Appendix 5: stopwords.txt*). The remaining text was lemmatized using WordNetLemmatizer [44]. The text was transformed into a bag-of-words representation, considering unigrams (1-gram), bigrams (2-gram), and trigrams (3-gram).

Model Features, Training, and Performance Evaluation

To reduce the number of features and enhance model interpretability, we employed an iterative approach using Logistic Regression with L1 regularization, similar to our previous NLP-based methods on unstructured EHR text in studies on epilepsy or neurological status following COVID-19 hospitalizations [38,45]. The regularization strength (10) and maximum iterations (1000) were chosen to balance between model complexity and computational efficiency. We addressed class imbalances by incorporating class weights inversely proportional to the sample size. We used this process to identify the most common features, resulting in a set of 1,284 features. A manual review by three experts (AS, SM, MBW) eliminated the remaining irrelevant features leading to a reduced set of 342 non-zero features. For example, terms like "Boston," "heart rate," and "primary care" were removed as they were irrelevant to NID classification and had no pathophysiological connection to the condition.

The MGB dataset (3,000 notes) was randomly split 80:20 for training (2,400 notes) and holdout (600 notes) test set; maintaining the 50/50 ICD code distribution. Since there were only 969 unique patients with ICD codes for NIDs, some notes originated from the same patients. To ensure no patient appeared in both the training and test sets, the initial holdout test set of 600 notes was reduced to 445 notes by excluding notes from patients who were also present in the training set. To select the optimal model, we utilized 5-fold cross-validation on the MGB training dataset. Given the

imbalanced nature of the dataset, with only 16%, 480 out of 3000, of the total observations positive for NID, we compared Logistic Regression, Random Forest, and XGBoost using the Area Under the Precision-Recall Curve (AUPRC). AUPRC is more informative for imbalanced datasets as it focuses on the performance of the model in predicting the positive class, unlike metrics such as accuracy or the Area Under the Receiver Operating Characteristic Curve (AUROC)[46–49]. XGBoost was selected for its AUPRC performance.

We then trained an XGBoost model on the 2,400 training notes and tested its performance on both the holdout dataset (445 notes from MGB) and an external dataset consisting of 600 notes from patients at BIDMC. Model performance was based on the AUPRC and the AUROC. We performed bootstrapping with replacement and performed 1,000 iterations to estimate the 95% confidence intervals (CIs) for these metrics.

Evaluation of Llama 3.2 for Zero-Shot NID Classification

We assessed the performance of the Llama 3.2 auto-regressive model (3B parameters) as a zero-shot learning approach for classifying NID from clinical reports in the MGB test set. Llama was neither fine-tuned nor trained on task-specific examples and classified the presence or absence of NID from clinical notes formatted in JavaScript Object Notation.

RESULTS

Cohort Characteristics

The demographic characteristics of the 2,469 patients comprising the 3,000 MGB clinical notes are presented in Table 1 (the linkage between notes and demographic data for the BIDMC dataset was lost and is therefore not shown).

Table 1. Characteristics of the study cohorts.

MGB (N=2469)
Age (years)

Median [IQR]	61.0 [46-73]
Age Group, n (%)	
<18	101 (4.1%)
18-34	254 (10.3%)
35-49	403 (16.3%)
50-64	718 (29.1%)
65+	993 (40.2%)
Sex, n (%)	
Male	1119 (45.3%)
Female	1350 (54.7%)
Race, n (%)	
White	1911 (77.4%)
Black	174 (7.0%)
Asian	86 (3.5%)
Other	298 (12.1%)
Ethnicity, n (%)	
Hispanic	252 (10.2%)
Not Hispanic	2041 (82.7%)
Missing	176 (7.1%)

The median age (IQR) was 61.0 years (46 - 73 years), of which 55% (n=1350) were female sex at birth, 77% (n=1911) reported being of white race, and 83% (n=2041) reported being of non-Hispanic ethnicity. Out of the 3,000 notes, 16% (479 notes) were labeled as NID based on expert review, and 97% of these 479 notes had an NID-related ICD code. Among the 1,500 notes with an NID-related ICD code, 465 notes (31%) were confirmed as NID cases by expert review. Conversely, among the 1,500 notes without a NID-related ICD code, 0.93% (14 notes) were identified as NID cases by expert review. The sensitivity of detecting NID patients using the presence of any NID-related ICD code was 97.1%, and the specificity was 59.1%, suggesting reliance on ICD codes alone for a patient's true clinical diagnosis may be more likely to overestimate, rather than underestimate, the number of cases with NID.

Performance of EHR NID Classifier

The training set consisted of 2,400 notes from MGB where 16.1% (387/2,400) were classified as NID-positive by expert review. The hold-out testing set from MGB consisted of 445 notes where 16% (71/445) were classified as NID-positive.

Performance characteristics using XGBoost on the MGB training set and hold-out test set are shown in Table 2.

Table 2. Average performance for XGBoost model on the training set and two independent testing sets.

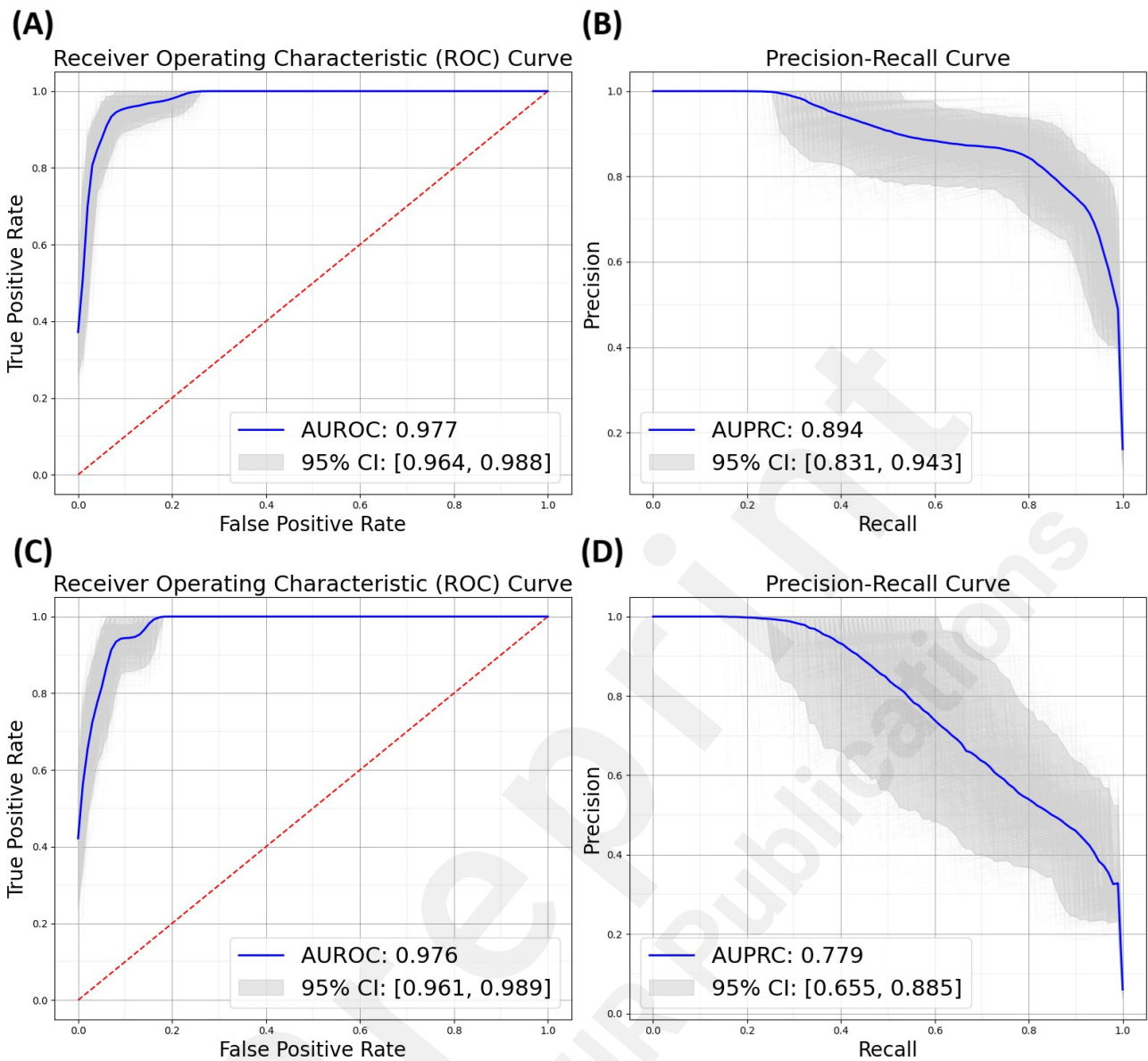
Metric	Training Set (MGB) (2400 notes)	Testing Set (MGB) (445 notes)	Testing Set (BIDMC) (600 notes)
AUROC	1.000 [1.000 - 1.000]	0.977 [0.964 - 0.988]	0.976 [0.961 - 0.989]
AUPRC	1.000 [0.999 - 1.000]	0.894 [0.831 - 0.943]	0.779 [0.655 - 0.885]
F1 Score	0.987 [0.978 - 0.995]	0.822 [0.752 - 0.879]	0.658 [0.528 - 0.778]
Recall	0.974 [0.957 - 0.989]	0.846 [0.753 - 0.923]	0.687 [0.538 - 0.839]
Precision	1.000 [1.000 - 1.000]	0.802 [0.709 - 0.889]	0.637 [0.487 - 0.795]
Specificity	1.000 [1.000 - 1.000]	0.960 [0.939 - 0.978]	0.976 [0.963 - 0.988]

AUROC – area under the receiver operating characteristic; AUPRC – area under the precision-recall curve; XGBoost – extreme gradient boosting model. The brackets indicate the 95% confidence interval.

The MGB hold-out test set showed good performance metrics (95% CI), including an AUROC of 0.977 (0.964 - 0.988), AUPRC of 0.894 (0.831 - 0.943), recall of 0.846 (0.753 - 0.923), and F1 score of 0.822 (0.752 - 0.879). The AUROC and AUPRC curves for the hold-out set are illustrated in *Figure 2*.

To assess performance metrics using patient notes extracted outside the MGB network, we ran the NLP-based model on 600 notes from BIDMC of which 5.8% (35/600) were NID-positive by expert review. Although performance declined compared to the MGB test set, the model maintained a high AUROC of 0.976 (95% CI: 0.961–0.989) and an AUPRC of 0.779 (95% CI: 0.655–0.885). Precision was 0.637 (95% CI: 0.487–0.795), and recall at 0.687 (95% CI: 0.538–0.839), resulting in an F1 score of 0.658 (95% CI: 0.528–0.778).

Additionally, we assessed the performance of Llama 3.2, a zero-shot learning model, for classifying NID as present or absent using clinical notes on the MGB test set. The model achieved an AUROC of 0.800 and a balanced F1 score of 0.799, demonstrating high specificity (0.940) but lower recall (0.640). These results highlight notable performance differences between the Llama model and the NLP-based approach, particularly in recall and overall F1 score.



Fig

Figure 2. XGBoost model performance on two sets of unseen notes, as evaluated by AUROC and AUPRC. 95% CIs were estimated using 1000 bootstrapping iterations. **(A)** AUROC and **(B)** AUPRC on notes from MGB and **(C)** AUROC and **(D)** AUPRC on notes from BIDMC. AUROC, area under the receiver operating characteristic (ROC) curve; AUPRC, area under the precision recall curve; CI, confidence interval.

Significant Features for NID Prediction

The feature selection steps reduced the initial training feature sets by 73.36% (342/1284). The top 20

features selected by XGBoost to identify individuals with NIDs are plotted in *Figure 3*.

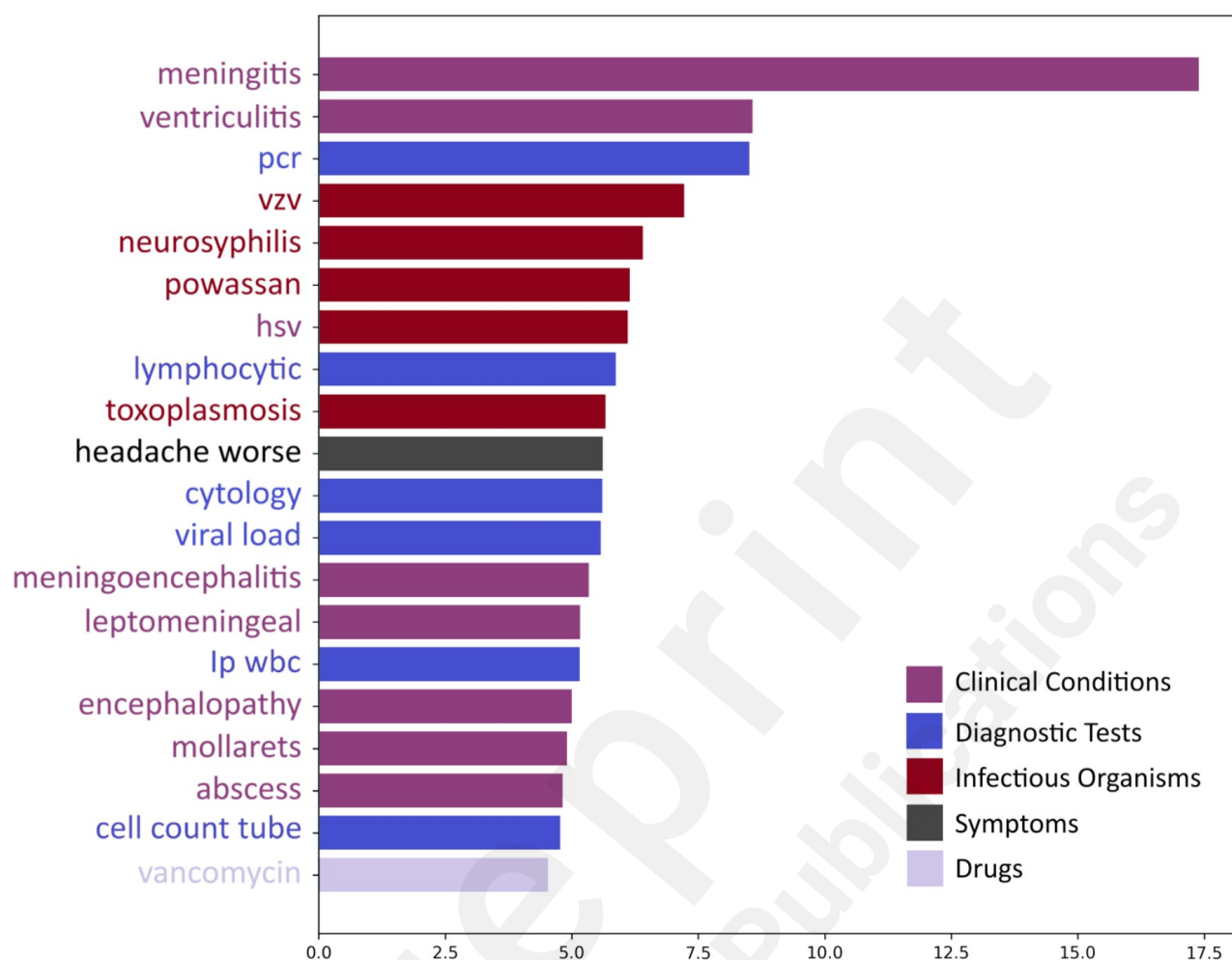


Figure 3. Top 20 features ranked by the feature importance (information gain measured by entropy), arranged in descending order (from top to bottom). The x-axis is the feature importance. It does not differentiate the association directions, i.e., positive, or negative association with the NID label.

The importance of these features was assessed based on their average gain across all decision points within the model, providing a measure of their contribution to predictive accuracy. These features primarily consist of clinical indicators and diagnostic markers essential for identifying NIDs. Prominently featured are direct markers of CNS inflammation, such as ‘meningitis’, ‘ventriculitis’, and ‘meningoencephalitis.’ Other features included diagnostic tests like ‘cytology’, ‘viral load’, and ‘pcr’ (polymerase chain reaction), cell types (‘lymphocytic’), specific pathogens, and medical

conditions associated with NID were identified, suggesting identified NID-associated features used in the NLP-classifier were clinically relevant and interpretable.

Error Analysis

In our analysis of cases where the model's predictions differed from the manually assigned labels, we discovered a few common factors that led to incorrect positive and negative classifications across both the internal (MGB) and external (BIDMC) datasets.

A common reason for false positives was the presence of prior NID mentions in patient notes. Another frequent issue was cases where patients were awaiting LP results, leading to a manual negative label, but the notes suggested a possible positive. False positives were also caused by notes indicating symptoms suggestive of NID but lacking confirmatory LP results. Additionally, specific pathogens or conditions, like JC virus neurocysticercosis and neurosyphilis, mentioned in the notes but not confirmed as active NID, added to the error rate. False negatives were often due to notes that did not discuss NID in detail.

The type of clinical note played a significant role; physical therapy and psychiatric notes were less likely to contain detailed clinical information relevant to NID, leading to both false positives and false negatives.

DISCUSSION

In this study, we developed and evaluated an automated NLP-based model in two independent datasets that classifies people with NID using unstructured notes to generate an NID cohort, with the intention that these cohorts could be used for predictions of susceptibility risk and short- and long-term medical outcomes. The NLP model demonstrated high specificity (0.96), outperforming the specificity observed using NID-specific ICD codes (0.59), suggesting that relying on ICD codes

alone may overestimate cases of infectious meningitis or encephalitis. These findings align with recent literature reporting low precision (0.58, 95% CI: 0.53–0.63) for the ICD-10 codes associated with herpesviral (herpes simplex) encephalitis, herpesviral (herpes simplex) meningoencephalitis, and herpesviral encephalitis, common causes of NID, in the well-established Danish National Patient Registry [19]. The lack of specificity underscores the risk of false positives when using ICD coding alone, posing challenges for accurately identifying clinical research cohorts, particularly for rare disorders like NID [50,51].

To contextualize our findings, prior studies in meningitis and encephalitis largely focused on structured data, including lab results, imaging, and other EHR data, to classify meningitis etiology or distinguish subtypes [24,31,32,52–58]. In comparison, alternative machine learning models applied to a Brazilian dataset of 22,602 suspected meningitis cases, including J48, Alternating Decision Tree, and Support Vector Machine, demonstrated varying levels of performance. One model, based solely on observable symptoms, achieved an AUROC of 0.869 for differentiating meningitis from non-meningitis cases [31], while a combined ensemble and decision tree method integrating observable symptoms with rapid CSF analysis attained an AUROC of 0.95 in distinguishing bacterial from viral meningitis in a dataset of known infectious meningitis cases [59]. However, the best model for classifying NID cases (sensitivity 97.0–97.6%) was at the expense of specificity (24.0–43.0%), limiting utility in reducing false positives in clinical decision-making [60]. In the context of HIV and cognitive disorders, machine learning models leveraging clinical data reported AUROCs of 0.83–0.87, though most relied on structured cohort data, not data in the HER, with limited validation in independent datasets [61–63]. These comparisons underscore the value of our NLP-based method as a more precise phenotyping tool, offering an improved balance between sensitivity and specificity for identifying NID cases.

We also evaluated LLMs focusing on the Llama3.2 model applied in the MGB test set. This method yielded an AUROC of 0.800 and a balanced F1 score of 0.799, with high specificity (0.940) but

lower recall (0.640). In contrast, the proposed bag-of-words NLP approach achieved balanced results, with an AUROC of 0.977, AUPRC of 0.894, and an F1 score of 0.822, with high recall (0.846). On the BIDMC validation set, the NLP model achieved an AUROC of 0.98, and while the performance declined (AUPRC reduced from 0.894 [0.831–0.943] to 0.779 [0.655–0.885]) likely due to variations in documentation styles, differences in clinical terminology, and institutional biases in diagnosing rare conditions, the model maintained good discriminative ability. These results demonstrate potential for broader application with further fine-tuning with enhanced precision and interpretability offering significant advantages for NID phenotyping in clinical research.

In our clinical experience, we have observed that unstructured clinical notes serve as a rich source of information crucial for ensuring clinical accuracy and care for NID patients. Here, using only clinical notes, we achieved consistent AUROC exceeding 0.95 across two different EHR systems, with a modest decrease in AUPRC, underscoring the importance of developing classifiers that effectively leverage the information embedded within these notes. Compared to Llama, the bag-of-words approach offers advantages such as reduced computational overhead, faster inference, and ease of deployment in settings with predictable patterns and resource limitations. These attributes make it particularly suitable for time-sensitive applications and adaptable to diverse healthcare settings. Most importantly, the proposed method's balanced performance, particularly in the context of NID classification, highlights its utility in both clinical research and practice. The model's reliance on unstructured notes reduces dependency on specific EHR system designs for structured data, broadening its applicability. By identifying clinically interpretable features such as specific markers of CNS inflammation and diagnostic tests, the model enhances both accuracy and clinical relevance, making it a valuable tool for improving care and outcomes for NID patients.

Our study had several limitations. First, the model was developed and validated using unstructured EHR notes from Massachusetts, impacting its relevance in other regions due to regional variations in medical terminology and disease incidence. For example, region-specific diseases like 'Powassan,'

found predominantly in northeastern states and the Great Lakes region of the United States, was a feature selected in the XGBoost model, and may not generalize to other regions. Second, while notes were enriched for NIDs by considering patients who had an LP procedure, the timing of the procedure in relation to the analyzed notes was not considered when developing the model. This approach aimed to enrich the dataset with relevant LP-related language and details from patient notes while avoiding overfitting the model narrowly to notes from a specific temporal context like the post-procedural period. While this approach prevents overfitting to unique details concentrated in the post-LP notes, it could limit broader applicability. Future work could explore incorporating the temporal relationship between the LP and the notes, which may provide additional context for classification performance. Third, lab results from CSF cell counts or microbial assays were not specifically considered in our analysis. The availability of lab data for each patient is inconsistent due to factors such as its assessment at another institution prior to transfer to a tertiary center, safety concerns regarding CSF retrieval, or delays due to logistical considerations. This NLP-based model, operating independently of lab results, resulted in high AUROC and AUPRC suggesting that in settings where structured lab data is not available, the tool could potentially generate a NID cohort. Our study focused on broad NID classification and did not subtype into encephalitis or meningitis. We anticipate that future model development will involve training on unstructured EHR notes from diverse sources across the United States and may explore incorporating microbial data and antimicrobial therapy details as features.

FUNDING: Grants supporting investigators include funding from the National Institute of Health (R01MH131194, R01MH134823 to S.S.M., and RF1AG064312, RF1NS120947, R01AG073410, R01HL161253, R01NS126282, R01AG073598, R01NS131347, R01NS130119 to M.B.W.), and the Claflin Distinguished Scholar award from the Massachusetts General Hospital to S.S.M.

DISCLOSURES: Dr. Westover is a co-founder, scientific advisor, consultant to, and has personal equity interest in Beacon Biosignals. Dr. Turbett received research funding from SeLux Diagnostics and has been a consultant for CARBX.

Abbreviation

AD: Alzheimer's Disease AUPRC: Area Under the Precision-Recall Curve

AUROC: Area Under the Receiver Operating Curve

BIDMC: Beth Israel Deaconess Medical Center

BOW: Bag-of-Words

CI: Confidence Interval

CNS: Central Nervous System

CSF: Cerebrospinal fluid

EHR: Electronic Health Record

ICD: International Classification of Diseases

LLM: Large Language Model

LP: Lumbar Punctures

MGB: Mass General Brigham

ML: Machine Learning

NID: NeuroInfectious Diseases

REFERENCES

1. Hasbun R, Rosenthal N, Balada-Llasat JM, Chung J, Duff S, Bozzette S, Zimmer L, Ginocchio CC. Epidemiology of Meningitis and Encephalitis in the United States, 2011-2014. *Clin Infect Dis Off Publ Infect Dis Soc Am* 2017 Aug 1;65(3):359–363. PMID:28419350
2. John CC, Carabin H, Montano SM, Bangirana P, Zunt JR, Peterson PK. Global research priorities for infections that affect the nervous system. *Nature* Nature Publishing Group; 2015 Nov;527(7578):S178–S186. doi: 10.1038/nature16033
3. Kiyani M, Hodges SE, Adil SM, Charalambous LT, Liu B, Lee H-J, Parente B, Perfect JR, Lad SP. Outcomes and Health Care Resource Utilization of Adult Bacterial Meningitis in the United States. *Neurol Clin Pract Wolters Kluwer*; 2021 Apr;11(2):117–126. doi: 10.1212/CPJ.0000000000000868
4. Wunrow HY, Bender RG, Vongpradith A, Sirota SB, Swetschinski LR, Novotney A, Gray AP, Ikuta KS, Sharara F, Wool EE, Aali A, Abd-Elsalam S, Abdollahi A, Abdul Aziz JM, Abidi H, Aboagye RG, Abolhassani H, Abu-Gharbieh E, Adamu LH, Adane TD, Addo IY, Adegboye OA, Adekiya TA, Adnan M, Adnani QES, Afzal S, Aghamiri S, Aghdam ZB, Agodi A, Ahinkorah BO, Ahmad A, Ahmad S, Ahmadzade M, Ahmed A, Ahmed A, Ahmed JQ, Ahmed MS, Akinosoglou K, Akililu A, Akonde M, Alahdab F, AL-Ahdal TMA, Alanezi FM, Albelbeisi AH, Alemayehu TBB, Alene KA, Al-Eyadhy A, Al-Gheethi AAS, Ali A, Ali BA, Ali L, Ali SS, Alimohamadi Y, Alipour V, Aljunid SM, Almoustanyir S, Al-Raddadi RM, Alvis-Guzman N, Al-Worafi YM, Aly H, Ameyaw EK, Ancuceanu R, Ansar A, Ansari G, Anyasodor AE, Arabloo J, Aravkin AY, Areda D, Artamonov AA, Arulappan J, Aruleba RT, Asaduzzaman M, Atalell KA, Athari SS, Atlaw D, Atout MMW, Attia S, Awoke T, Ayalew MK, Ayana TM, Ayele AD, Azadnajafabad S, Azizian K, Badar M, Badiye AD, Baghcheghi N, Bagheri M, Bagherieh S, Bahadory S, Baig AA, Barac A, Barati S, Bardhan M, Basharat Z, Bashiri A, Basnyat B, Bassat Q, Basu S, Bayileyegn NS, Bedi N, Behnouth AH, Bekel AA, Belete MA, Bello OO, Bhagavathula AS, Bhandari D, Bhardwaj P, Bhaskar S, Bhat AN, Bijani A, Bineshfar N, Boloor A, Bouaoud S, Buonsenso D, Burkart K, Cámara LA, Castañeda-Orjuela CA, Cernigliaro A, Charan J, Chattu VK, Ching PR, Chopra H, Choudhari SG, Christopher DJ, Chu D-T, Couto RAS, Cruz-Martins N, Dadras O, Dai X, Dandona L, Dandona R, Das S, Dash NR, Dashti M, De la Hoz FP, Debela SA, Dejen D, Dejene H, Demeke D, Demeke FM, Demessa BH, Demetriades AK, Demissie S, Dereje D, Dervišević E, Desai HD, Dessie AM, Desta F, Dhama K, Djalalinia S, Do TC, Dodangeh M, Dodangeh M, Dominguez R-MV, Dongarwar D, Dsouza HL, Durojaiye OC, Dziedzic AM, Ekat MH, Ekholuenetale M, Ekundayo TC, El Sayed Zaki M, El-Abid H, Elhadi M, El-Hajj VG, El-Huneidi W, El-Sakka AA, Esayas HL, Fagbamigbe AF, Falahi S, Fares J, Fatehizadeh A, Fatima SAF, Feasey NA, Fekadu G, Fetensa G, Feyissa D, Fischer F, Foroutan B, Gaal PA, Gadanya MA, Gaipov A, Ganesan B, Gebrehiwot M, Gebrekidan KG, Gebremeskel TG, Gedef GM, Gela YY, Gerema U, Gessner BD, Getachew ME, Ghadiri K, Ghaffari K, Ghamari S-H, Ghanbari R, Ghazy RMM, Ghozali G, Gizaw ABA, Glushkova EV, Goldust M, Golechha M, Guadie HA, Guled RA, Gupta M, Gupta S, Gupta VB, Gupta VK, Gupta VK, Hadi NR, Haj-Mirzaian A, Haller S, Hamidi S, Haque S, Harapan H, Hasaballah AI, Hasan I, Hasani H, Hasanian M, Hassankhani H, Hassen MB, Hayat K, Heidari M, Heidari-Foroozan M, Heidari-Soureshjani R, Hezam K, Holla R, Horita N, Hossain MM, Hosseini M-S, Hosseinzadeh M, Hostiuc S, Hussain S, Hussein NR, Ibitoye SE, Ilesanmi OS, Ilic IM, Ilic MD, Imam MT, Iregbu KC, Ismail NE, Iwu CCD, Jaja C, Jakovljevic M, Jamshidi E, Javadi Mamaghani A, Javidnia J, Jokar M, Jomehzadeh N, Joseph N, Joshua CE, Jozwiak JJ,

Kabir Z, Kalankesh LR, Kalhor R, Kamal VK, Kandel H, Karaye IM, Karch A, Karimi H, Kaur H, Kaur N, Keykhaei M, Khajuria H, Khalaji A, Khan A, Khan IA, Khan M, Khan T, Khatab K, Khatatbeh MM, Khayat Kashani HR, Khubchandani J, Kim MS, Kisa A, Kisa S, Kompani F, Koohestani HR, Kothari N, Krishan K, Krishnamoorthy Y, Kulimbet M, Kumar M, Kumaran SD, Kuttikkattu A, Kwarteng A, Laksono T, Landires I, Laryea DO, Lawal BK, Le TTT, Ledda C, Lee S, Lee S, Lema GK, Levi M, Lim SS, Liu X, Lopes G, Lutzky Saute R, Machado Teixeira PH, Mahmoodpoor A, Mahmoud MA, Malakan Rad E, Malhotra K, Malik AA, Martinez-Guerra BA, Martorell M, Mathur V, Mayeli M, Medina JRC, Melese A, Memish ZA, Mentis A-FA, Merza MA, Mestrovic T, Michalek IM, Minh LHN, Mirahmadi A, Mirmosayyeb O, Misganaw A, Misra AK, Moghadasi J, Mohamed NS, Mohammad Y, Mohammadi E, Mohammed S, Mojarrad Sani M, Mojiri-forushani H, Mokdad AH, Momtazmanesh S, Monasta L, Moni MA, Mossialos E, Mostafavi E, Motaghinejad M, Mousavi Khaneghah A, Mubarik S, Muccioli L, Muhammad JS, Mulita F, Mulugeta T, Murillo-Zamora E, Mustafa G, Muthupandian S, Nagarajan AJ, Nainu F, Nair TS, Nargus S, Nassereldine H, Natto ZS, Nayak BP, Negoï I, Negoï RI, Nejadghaderi SA, Nguyen HQ, Nguyen PT, Nguyen VT, Niazi RK, Noroozi N, Nouraei H, Nuñez-Samudio V, Nuruzzaman KM, Nwatah VE, Nzoputam CI, Nzoputam OJ, Oancea B, Obaidur RM, Odetokun IA, Ogunsakin RE, Okonji OC, Olagunju AT, Olana LT, Olufadewa II, Oluwafemi YD, Oumer KS, Ouyahia A, P a M, Pakshir K, Palange PN, Pardhan S, Parikh RR, Patel J, Patel UK, Patil S, Paudel U, Pawar S, Pensato U, Perdigão J, Pereira M, Peres MFP, Petcu I-R, Pinheiro M, Piracha ZZ, Pokhrel N, Postma MJ, Prates EJS, Qattea I, Raghav PR, Rahbarnia L, Rahimi-Movaghar V, Rahman M, Rahman MA, Rahmanian V, Rahnavard N, Ramadan H, Ramasubramani P, Rani U, Rao IR, Rapaka D, Ratan ZA, Rawaf S, Redwan EMM, Reiner Jr RC, Rezaei N, Riad A, Ribeiro da Silva TM, Roberts T, Robles Aguilar G, Rodriguez JAB, Rosenthal VD, Saddik B, Sadeghian S, Saeed U, Safary A, Saheb Sharif-Askari F, Saheb Sharif-Askari N, Sahebkar A, Sahu M, Sajedi SA, Saki M, Salahi S, Salahi S, Saleh MA, Sallam M, Samadzadeh S, Samy AM, Sanjeev RK, Satpathy M, Seylani A, Sha'aban A, Shafie M, Shah PA, Shahrokhi S, Shahzamani K, Shaikh MA, Sham S, Shannawaz M, Sheikh A, Shenoy SM, Shetty PH, Shin JI, Shokri F, Shorofi SA, Shrestha S, Sibhat MM, Siddig EE, Silva LMLR, Singh H, Singh JA, Singh P, Singh S, Sinto R, Skryabina AA, Socea B, Sokhan A, Solanki R, Solomon Y, Sood P, Soshnikov S, Stergachis A, Sufiyan MB, Suliankatchi Abdulkader R, Sultana A, T y SS, Taheri E, Taki E, Tamuzi JJJ, Tan K-K, Tat NY, Temsah M-H, Terefa DR, Thangaraju P, Tibebu NS, Ticoalu JHV, Tillawi T, Tincho MB, Tleyjeh II, Togholi R, Tovani-Palone MR, Tufa DG, Turner P, Ullah I, Umeokonkwo CD, Unnikrishnan B, Vahabi SM, Vaithinathan AG, Valizadeh R, Varthya SB, Vos T, Waheed Y, Walde MT, Wang C, Weerakoon KG, Wickramasinghe ND, Winkler AS, Woldemariam M, Worku NA, Wright C, Yada DY, Yaghoubi S, Yahya GATY, Yenew CY, Yesiltepe M, Yi S, Yiğit V, You Y, Yusuf H, Zakham F, Zaman M, Zaman SB, Zare I, Zareshahrabadi Z, Zarrintan A, Zastrozhin MS, Zhang H, Zhang J, Zhang Z-J, Zheng P, Zoladl M, Zumla A, Hay SI, Murray CJL, Naghavi M, Kyu HH. Global, regional, and national burden of meningitis and its aetiologies, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. *Lancet Neurol* 2023 Aug 1;22(8):685–711. doi: 10.1016/S1474-4422(23)00195-3

5. Edmond K, Clark A, Korczak VS, Sanderson C, Griffiths UK, Rudan I. Global and regional risk of disabling sequelae from bacterial meningitis: a systematic review and meta-analysis. *Lancet Infect Dis* 2010 May;10(5):317–328. PMID:20417414
6. van de Beek D, Schmand B, de Gans J, Weisfelt M, Vaessen H, Dankert J, Vermeulen M. Cognitive impairment in adults with good recovery after bacterial meningitis. *J Infect Dis* 2002 Oct 1;186(7):1047–1052. PMID:12232850

7. Davis AG, Nightingale S, Springer PE, Solomons R, Arenivas A, Wilkinson RJ, Anderson ST, Chow FC, Tuberculous Meningitis International Research Consortium. Neurocognitive and functional impairment in adult and paediatric tuberculous meningitis. *Wellcome Open Res* 2019;4:178. PMID:31984243
8. Hansen A-BE, Vestergaard HT, Dessau RB, Bodilsen J, Andersen NS, Omland LH, Christiansen CB, Ellermann-Eriksen S, Nielsen L, Benfield T, Sørensen HT, Andersen CØ, Lebech A-M, Obel N. Long-Term Survival, Morbidity, Social Functioning and Risk of Disability in Patients with a Herpes Simplex Virus Type 1 or Type 2 Central Nervous System Infection, Denmark, 2000-2016. *Clin Epidemiol* 2020;12:745–755. PMID:32765109
9. Douros A, Santella C, Dell’Aniello S, Azoulay L, Renoux C, Suissa S, Brassard P. Infectious Disease Burden and the Risk of Alzheimer’s Disease: A Population-Based Study. *J Alzheimers Dis JAD* 2021;81(1):329–338. PMID:33780369
10. Levine KS, Leonard HL, Blauwendraat C, Iwaki H, Johnson N, Bandres-Ciga S, Ferrucci L, Faghri F, Singleton AB, Nalls MA. Virus exposure and neurodegenerative disease risk across national biobanks. *Neuron* 2023 Apr 5;111(7):1086-1093.e2. PMID:36669485
11. Sipilä PN, Heikkilä N, Lindbohm JV, Hakulinen C, Vahtera J, Elovainio M, Suominen S, Väänänen A, Koskinen A, Nyberg ST, Pentti J, Strandberg TE, Kivimäki M. Hospital-treated infectious diseases and the risk of dementia: a large, multicohort, observational study with a replication cohort. *Lancet Infect Dis* 2021 Nov;21(11):1557–1567. PMID:34166620
12. Whitson HE, Colton C, El Khoury J, Gate D, Goate A, Heneka MT, Kaddurah-Daouk R, Klein RS, Shinohara ML, Sisodia S, Spudich SS, Stevens B, Tanzi R, Ting JP, Garden G. Infection and inflammation: New perspectives on Alzheimer’s disease. *Brain Behav Immun - Health* 2022 Apr 22;22:100462. PMID:36118272
13. Gomez CA, Pinsky BA, Liu A, Banaei N. Delayed Diagnosis of Tuberculous Meningitis Misdiagnosed as Herpes Simplex Virus-1 Encephalitis With the FilmArray Syndromic Polymerase Chain Reaction Panel. *Open Forum Infect Dis* 2016 Dec 7;4(1):ofw245. PMID:28540320
14. Miller AC, Koeneman SH, Arakkal AT, Cavanaugh JE, Polgreen PM. Incidence, Duration, and Risk Factors Associated With Missed Opportunities to Diagnose Herpes Simplex Encephalitis: A Population-Based Longitudinal Study. *Open Forum Infect Dis* 2021 Sep 1;8(9):ofab400. doi: 10.1093/ofid/ofab400
15. Nath A. Neuroinfectious Diseases: A Crisis in Neurology and a Call for Action. *JAMA Neurol* 2015 Feb 1;72(2):143–144. doi: 10.1001/jamaneurol.2014.3442
16. Nesher L, Hadi CM, Salazar L, Wootton SH, Garey KW, Lasco T, Luce AM, Hasbun R. Epidemiology of meningitis with a negative CSF Gram stain: under-utilization of available diagnostic tests. *Epidemiol Infect* 2016 Jan;144(1):189–197. PMID:25989841
17. Piantadosi A, Mukerji SS, Ye S, Leone MJ, Freimark LM, Park D, Adams G, Lemieux J, Kanjilal S, Solomon IH, Ahmed AA, Goldstein R, Ganesh V, Ostrem B, Cummins KC, Thon JM, Kinsella CM, Rosenberg E, Frosch MP, Goldberg MB, Cho TA, Sabeti P. Enhanced Virus Detection and Metagenomic Sequencing in Patients with Meningitis and Encephalitis. *mBio American Society for Microbiology*; 2021 Aug 31;12(4):e01143-21. doi: 10.1128/mBio.01143-

18. Wilson MR, Sample HA, Zorn KC, Arevalo S, Yu G, Neuhaus J, Federman S, Stryke D, Briggs B, Langelier C, Berger A, Douglas V, Josephson SA, Chow FC, Fulton BD, DeRisi JL, Gelfand JM, Naccache SN, Bender J, Dien Bard J, Murkey J, Carlson M, Vespa PM, Vijayan T, Allyn PR, Campeau S, Humphries RM, Klausner JD, Ganzon CD, Memar F, Ocampo NA, Zimmermann LL, Cohen SH, Polage CR, DeBiasi RL, Haller B, Dallas R, Maron G, Hayden R, Messacar K, Dominguez SR, Miller S, Chiu CY. Clinical Metagenomic Sequencing for Diagnosis of Meningitis and Encephalitis. *N Engl J Med Massachusetts Medical Society*; 2019 Jun 13;380(24):2327–2340. PMID:31189036
19. Jørgensen LK, Dalgaard LS, Østergaard LJ, Andersen NS, Nørgaard M, Mogensen TH. Validity of the coding for herpes simplex encephalitis in the Danish National Patient Registry. *Clin Epidemiol* 2016 May 31;8:133–140. PMID:27330328
20. Farzandipour M, Sheikhtaheri A, Sadoughi F. Effective factors on accuracy of principal diagnosis coding based on International Classification of Diseases, the 10th revision (ICD-10). *Int J Inf Manag* 2010 Feb 1;30:78–84. doi: 10.1016/j.ijinfomgt.2009.07.002
21. O'Malley KJ, Cook KF, Price MD, Wildes KR, Hurdle JF, Ashton CM. Measuring Diagnoses: ICD Code Accuracy. *Health Serv Res* 2005;40(5p2):1620–1639. doi: 10.1111/j.1475-6773.2005.00444.x
22. Samannodi M, Hansen M, Hasbun R. Lack of accuracy of the international classification of disease, ninth (ICD-9) codes in identifying patients with encephalitis. *J Neurol* 2019 Apr 1;266(4):1034–1035. doi: 10.1007/s00415-019-09229-9
23. Stausberg J, Lehmann N, Kaczmarek D, Stein M. Reliability of diagnoses coding with ICD-10. *Int J Med Inf* 2008 Jan;77(1):50–57. PMID:17185030
24. Guzmán E, Belmonte M-V, Lelis VM. Ensemble methods for meningitis aetiology diagnosis. *Expert Syst* 2022;39(8):e12996. doi: 10.1111/exsy.12996
25. Freedman SB, Marrocco A, Pirie J, Dick PT. Predictors of Bacterial Meningitis in the Era After *Haemophilus influenzae*. *Arch Pediatr Adolesc Med* 2001 Dec 1;155(12):1301–1306. doi: 10.1001/archpedi.155.12.1301
26. Wu J, Shi T, Yue Y, Kong X, Cheng F, Jiang Y, Bian Y, Tian J. Development a prediction model for identifying bacterial meningitis in young infants aged 29–90 days: a retrospective analysis. *BMC Pediatr* 2023 Feb 9;23:69. PMID:36759812
27. Cheng X, Zhang Q, Fu Z, Shi Z, Xia P, Zhang Y, Mao F, Xu Q, Yan X, Wang L. Establishment of a predictive model for purulent meningitis in preterm infants. *Transl Pediatr* 2022 Jun;11(6):1018–1027. PMID:35800267
28. Wang P, Cheng S, Li Y, Liu L, Liu J, Zhao Q, Luo S. Prediction of Lumbar Drainage-Related Meningitis Based on Supervised Machine Learning Algorithms. *Front Public Health* 2022 Jun 28;10:910479. PMID:35836985
29. Mago VK, Mehta R, Woolrych R, Papageorgiou EI. Supporting meningitis diagnosis amongst infants and children through the use of fuzzy cognitive mapping. *BMC Med Inform Decis Mak*

2012 Sep 4;12:98. PMID:22947265

30. Mentis A-FA, Garcia I, Jiménez J, Paparoupa M, Xirogianni A, Papandreou A, Tzanakaki G. Artificial Intelligence in Differential Diagnostics of Meningitis: A Nationwide Study. *Diagnostics Multidisciplinary Digital Publishing Institute*; 2021 Apr;11(4):602. doi: 10.3390/diagnostics11040602
31. Lelis VM, Guzmán E, Belmonte M-V. Non-Invasive Meningitis Diagnosis Using Decision Trees. *IEEE Access* 2020;8:18394–18407. doi: 10.1109/ACCESS.2020.2966397
32. Shi Y, Zhang C, Pan S, Chen Y, Miao X, He G, Wu Y, Ye H, Weng C, Zhang H, Zhou W, Yang X, Liang C, Chen D, Hong L, Su F. The diagnosis of tuberculous meningitis: advancements in new technologies and machine learning algorithms. *Front Microbiol* 2023 Oct 24;14:1290746. PMID:37942080
33. Ghaddaripouri K, Ghaddaripouri M, Mousavi AS, Mousavi Baigi SF, Rezaei Sarsari M, Dahmardeh Kemmak F, Mazaheri Habibi MR. The effect of machine learning algorithms in the prediction, and diagnosis of meningitis: A systematic review. *Health Sci Rep* 2024 Feb 14;7(2):e1893. PMID:38357491
34. Chen P-F, Chen L, Lin Y-K, Li G-H, Lai F, Lu C-W, Yang C-Y, Chen K-C, Lin T-Y. Predicting Postoperative Mortality With Deep Neural Networks and Natural Language Processing: Model Development and Validation. *JMIR Med Inform* 2022 May 10;10(5):e38241. PMID:35536634
35. Ge W, Alabsi H, Jain A, Ye E, Sun H, Fernandes M, Magdamo C, Tesh RA, Collens SI, Newhouse A, Moura LM, Zafar S, Hsu J, Akeju O, Robbins GK, Mukerji SS, Das S, Westover MB. Identifying Patients With Delirium Based on Unstructured Clinical Notes: Observational Study. *JMIR Form Res* 2022 Jun 24;6(6):e33834. doi: 10.2196/33834
36. Lee RY, Kross EK, Torrence J, Li KS, Sibley J, Cohen T, Lober WB, Engelberg RA, Curtis JR. Assessment of Natural Language Processing of Electronic Health Records to Measure Goals-of-Care Discussions as a Clinical Trial Outcome. *JAMA Netw Open* 2023 Mar 2;6(3):e231204. doi: 10.1001/jamanetworkopen.2023.1204
37. Chen P-F, Wang S-M, Liao W-C, Kuo L-C, Chen K-C, Lin Y-C, Yang C-Y, Chiu C-H, Chang S-C, Lai F. Automatic ICD-10 Coding and Training System: Deep Neural Network Based on Supervised Learning. *JMIR Med Inform* 2021 Aug 31;9(8):e23230. doi: 10.2196/23230
38. Fernandes M, Cardall A, Jing J, Ge W, Moura LMVR, Jacobs C, McGraw C, Zafar SF, Westover MB. Identification of patients with epilepsy using automated electronic health records phenotyping. *Epilepsia* 2023;64(6):1472–1481. doi: 10.1111/epi.17589
39. Fernandes MB, Valizadeh N, Alabsi HS, Quadri SA, Tesh RA, Bucklin AA, Sun H, Jain A, Brenner LN, Ye E, Ge W, Collens SI, Lin S, Das S, Robbins GK, Zafar SF, Mukerji SS, Brandon Westover M. Classification of neurologic outcomes from medical notes using natural language processing. *Expert Syst Appl* 2023 Mar 15;214:119171. doi: 10.1016/j.eswa.2022.119171
40. McGill F, Heyderman RS, Michael BD, Defres S, Beeching NJ, Borrow R, Glennie L, Gaillemine O, Wyncoll D, Kaczmarek E, Nadel S, Thwaites G, Cohen J, Davies NWS, Miller A, Rhodes A, Read RC, Solomon T. The UK joint specialist societies guideline on the diagnosis

- and management of acute meningitis and meningococcal sepsis in immunocompetent adults. *J Infect* 2016 Apr 1;72(4):405–438. doi: 10.1016/j.jinf.2016.01.007
41. 2017 Infectious Diseases Society of America's Clinical Practice Guidelines for Healthcare-Associated Ventriculitis and Meningitis* | Clinical Infectious Diseases | Oxford Academic. Available from: <https://academic.oup.com/cid/article/64/6/e34/2996079?login=true> [accessed Jan 7, 2025]
 42. Diagnosis and Treatment of Neurocysticercosis: 2017 Clinical Practice Guidelines by the Infectious Diseases Society of America (IDSA) and the American Society of Tropical Medicine and Hygiene (ASTMH) | Clinical Infectious Diseases | Oxford Academic. Available from: <https://academic.oup.com/cid/article/66/8/e49/4885412?login=true> [accessed Jan 7, 2025]
 43. Primary Care Guidance for Persons With Human Immunodeficiency Virus: 2020 Update by the HIV Medicine Association of the Infectious Diseases Society of America | Clinical Infectious Diseases | Oxford Academic. Available from: <https://academic.oup.com/cid/article/73/11/e3572/5956736> [accessed Jan 7, 2025]
 44. Bird S, Loper E, Klein E. *Natural Language Processing with Python*. O'Reilly Media Inc.; 2009.
 45. Fernandes M, Sun H, Jain A, Alabsi HS, Brenner LN, Ye E, Ge W, Collens SI, Leone MJ, Das S, Robbins GK, Mukerji SS, Westover MB. Classification of the Disposition of Patients Hospitalized with COVID-19: Reading Discharge Summaries Using Natural Language Processing. *JMIR Med Inform* 2021 Feb 10;9(2):e25457. PMID:33449908
 46. Liu Z, Bondell HD. Binormal Precision–Recall Curves for Optimal Classification of Imbalanced Data. *Stat Biosci* 2019 Apr 1;11(1):141–161. doi: 10.1007/s12561-019-09231-9
 47. Movahedi F, Antaki JF. Limitation of ROC in Evaluation of Classifiers for Imbalanced Data. *J Heart Lung Transplant Elsevier*; 2021 Apr 1;40(4):S413. doi: 10.1016/j.healun.2021.01.1160
 48. Kim M, Hwang K-B. An empirical evaluation of sampling methods for the classification of imbalanced data. *PLOS ONE Public Library of Science*; 2022 Jul 28;17(7):e0271260. doi: 10.1371/journal.pone.0271260
 49. Saito T, Rehmsmeier M. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLOS ONE Public Library of Science*; 2015 Mar 4;10(3):e0118432. doi: 10.1371/journal.pone.0118432
 50. Lo Re III V, Carbonari DM, Jacob J, Short WR, Leonard CE, Lyons JG, Kennedy A, Damon J, Haug N, Zhou EH, Graham DJ, McMahon-Walraven CN, Parlett LE, Nair V, Selvan M, Zhou Y, Pocobelli G, Maro JC, Nguyen MD. Validity of ICD-10-CM diagnoses to identify hospitalizations for serious infections among patients treated with biologic therapies. *Pharmacoepidemiol Drug Saf* 2021;30(7):899–909. doi: 10.1002/pds.5253
 51. Wiese AD, Griffin MR, Stein CM, Schaffner W, Greevy RA, Mitchel Jr EF, Grijalva CG. Validation of discharge diagnosis codes to identify serious infections among middle age and older adults. *BMJ Open* 2018 Jun 19;8(6):e020857. PMID:29921683
 52. Bonsu BK, Harper MB. Differentiating acute bacterial meningitis from acute viral meningitis

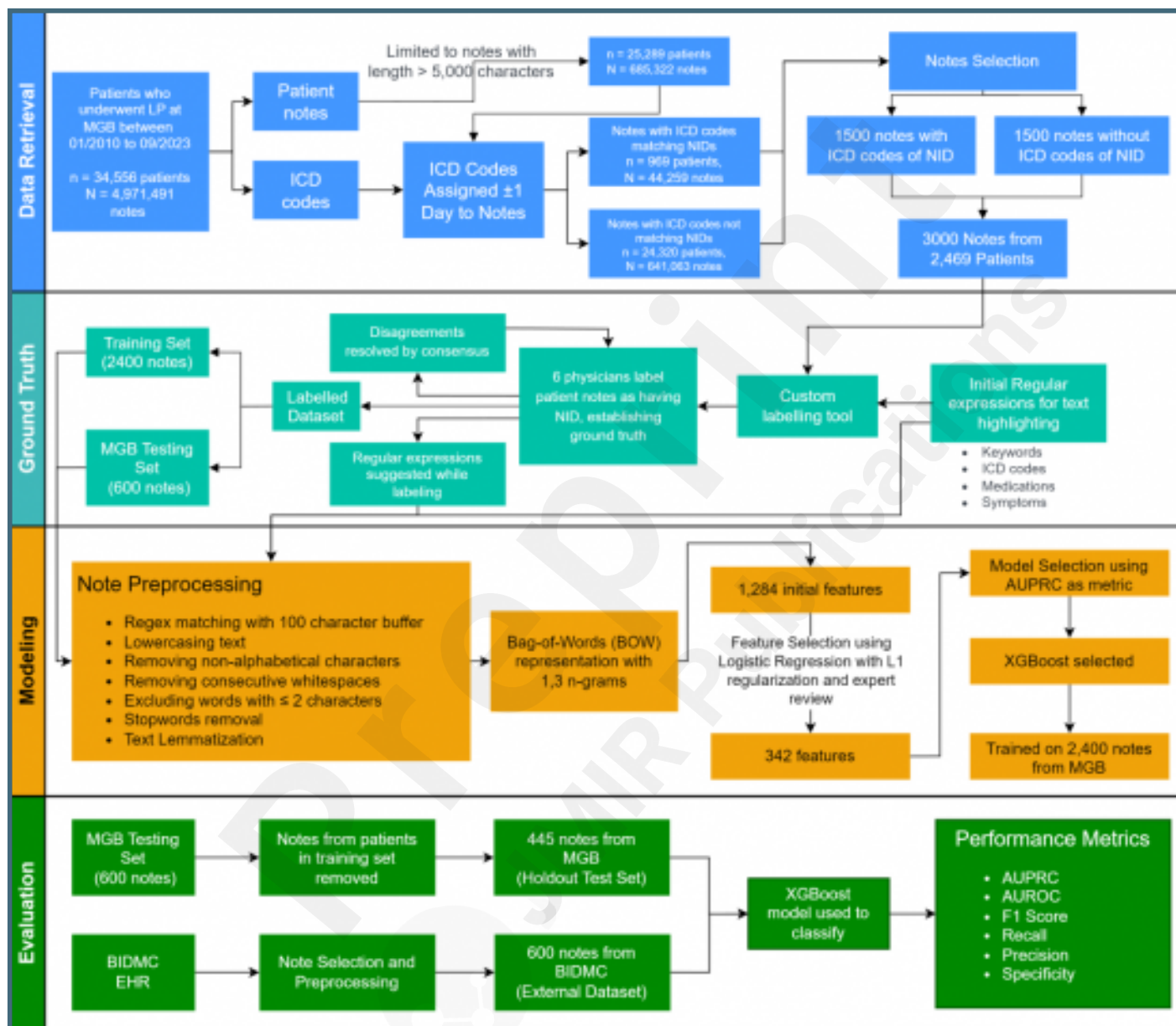
- among children with cerebrospinal fluid pleocytosis: a multivariable regression model. *Pediatr Infect Dis J* 2004 Jun;23(6):511–517. PMID:15194831
53. D'Angelo G, Pilla R, Tascini C, Rampone S. A proposal for distinguishing between bacterial and viral meningitis using genetic programming and decision trees. *Soft Comput* 2019 Nov 1;23(22):11775–11791. doi: 10.1007/s00500-018-03729-y
 54. Ene M, Gorunescu M, Gorunescu F, Revett K, Ene M, Gorunescu M, Gorunescu F, Revett K. A Machine Learning Approach to Differentiating Bacterial From Viral Meningitis. *IEEE John Vincent Atanasoff 2006 Int Symp Mod Comput JVA06* 2006. p. 155–162. doi: 10.1109/JVA.2006.2
 55. Gowin E, Januszkiewicz-Lewandowska D, Slowinski R, Blaszczyński J, Michalak M, Wysocki J. With a little help from a computer: discriminating between bacterial and viral meningitis based on dominance-based rough set approach analysis. *Medicine (Baltimore)* 2017 Aug;96(32):e7635. doi: 10.1097/MD.00000000000007635
 56. Jeong Y-S, Jeon M, Park JH, Kim M-C, Lee E, Park SY, Lee Y-M, Choi S, Park SY, Park K-H, Kim S-H, Jeon MH, Choo EJ, Kim TH, Lee MS, Kim T. Machine-Learning-Based Approach to Differential Diagnosis in Tuberculous and Viral Meningitis. *Infect Chemother* 2021 Mar;53(1):53–62. PMID:33538134
 57. Nigrovic LE, Kuppermann N, Malley R. Development and Validation of a Multivariable Predictive Model to Distinguish Bacterial From Aseptic Meningitis in Children in the Post-Haemophilus influenzae Era. *Pediatrics* 2002 Oct 1;110(4):712–719. doi: 10.1542/peds.110.4.712
 58. Xiang Y, Zeng C, Liu B, Tan W, Wu J, Hu X, Han Y, Luo Q, Gong J, Liu J, Li Y. Deep Learning-Enabled Identification of Autoimmune Encephalitis on 3D Multi-Sequence MRI. *J Magn Reson Imaging* 2022;55(4):1082–1092. doi: 10.1002/jmri.27909
 59. Guzmán E, Belmonte M-V, Lelis VM. Ensemble methods for meningitis aetiology diagnosis. *Expert Syst* 2022;39(8):e12996. doi: 10.1111/exsy.12996
 60. Lélis V-M, Guzmán E, Belmonte M-V. A Statistical Classifier to Support Diagnose Meningitis in Less Developed Areas of Brazil. *J Med Syst* 2017 Aug 11;41(9):145. doi: 10.1007/s10916-017-0785-5
 61. Mukerji SS, Petersen KJ, Pohl KM, Dastgheyb RM, Fox HS, Bilder RM, Brouillette M-J, Gross AL, Scott-Sheldon LAJ, Paul RH, Gabuzda D. Machine Learning Approaches to Understand Cognitive Phenotypes in People With HIV. *J Infect Dis* 2023 Mar 15;227(Supplement_1):S48–S57. doi: 10.1093/infdis/jiac293
 62. Xu Y, Lin Y, Bell RP, Towe SL, Pearson JM, Nadeem T, Chan C, Meade CS. Machine learning prediction of neurocognitive impairment among people with HIV using clinical and multimodal magnetic resonance imaging data. *J Neurovirol* 2021 Feb;27(1):1–11. PMID:33464541
 63. Tu W, Chen PA, Koenig N, Gomez D, Fujiwara E, Gill MJ, Kong L, Power C. Machine learning models reveal neurocognitive impairment type and prevalence are associated with distinct variables in HIV/AIDS. *J Neurovirol* 2020 Feb;26(1):41–51. PMID:31520320

Preprint
JMIR Publications

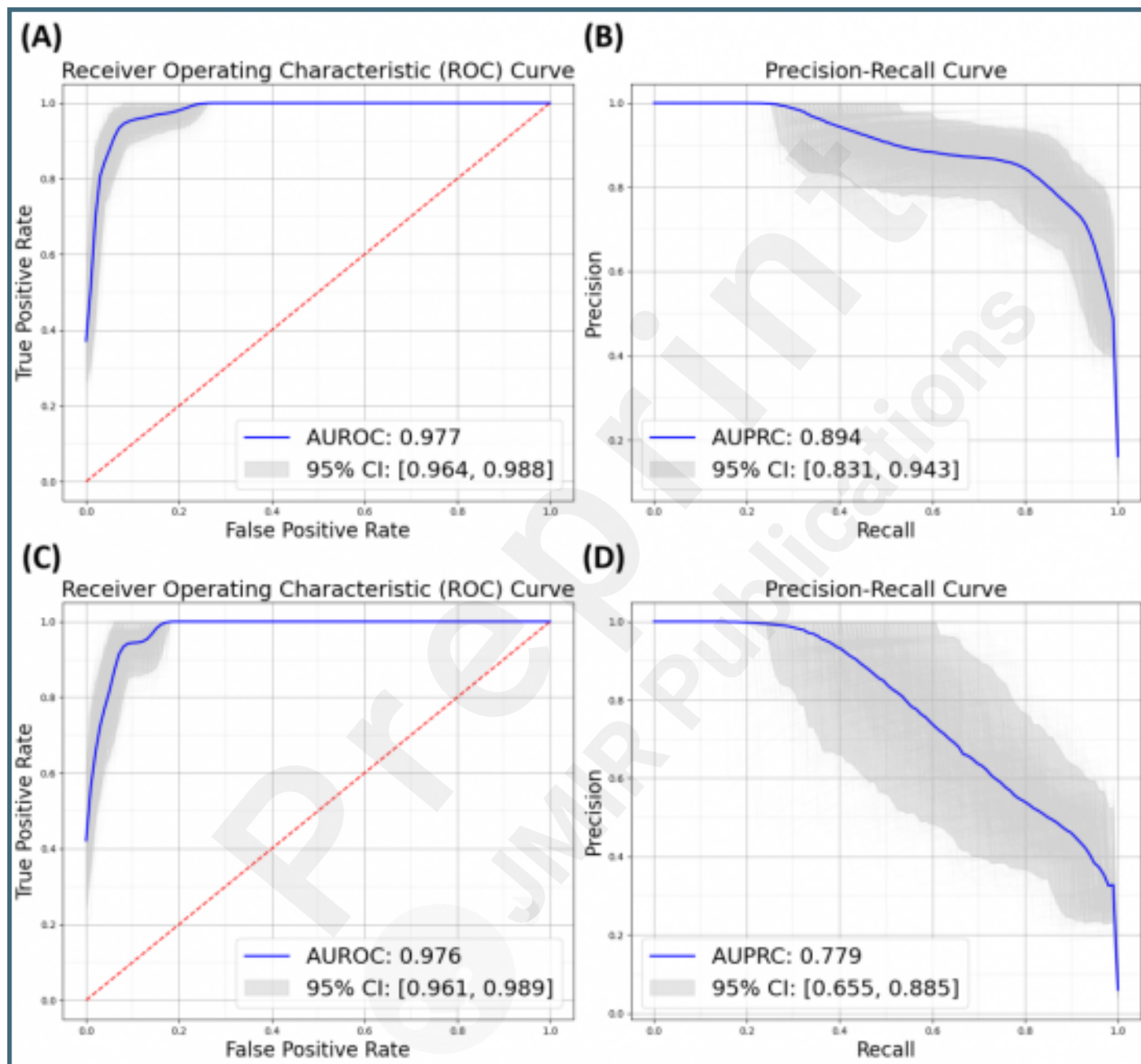
Supplementary Files

Figures

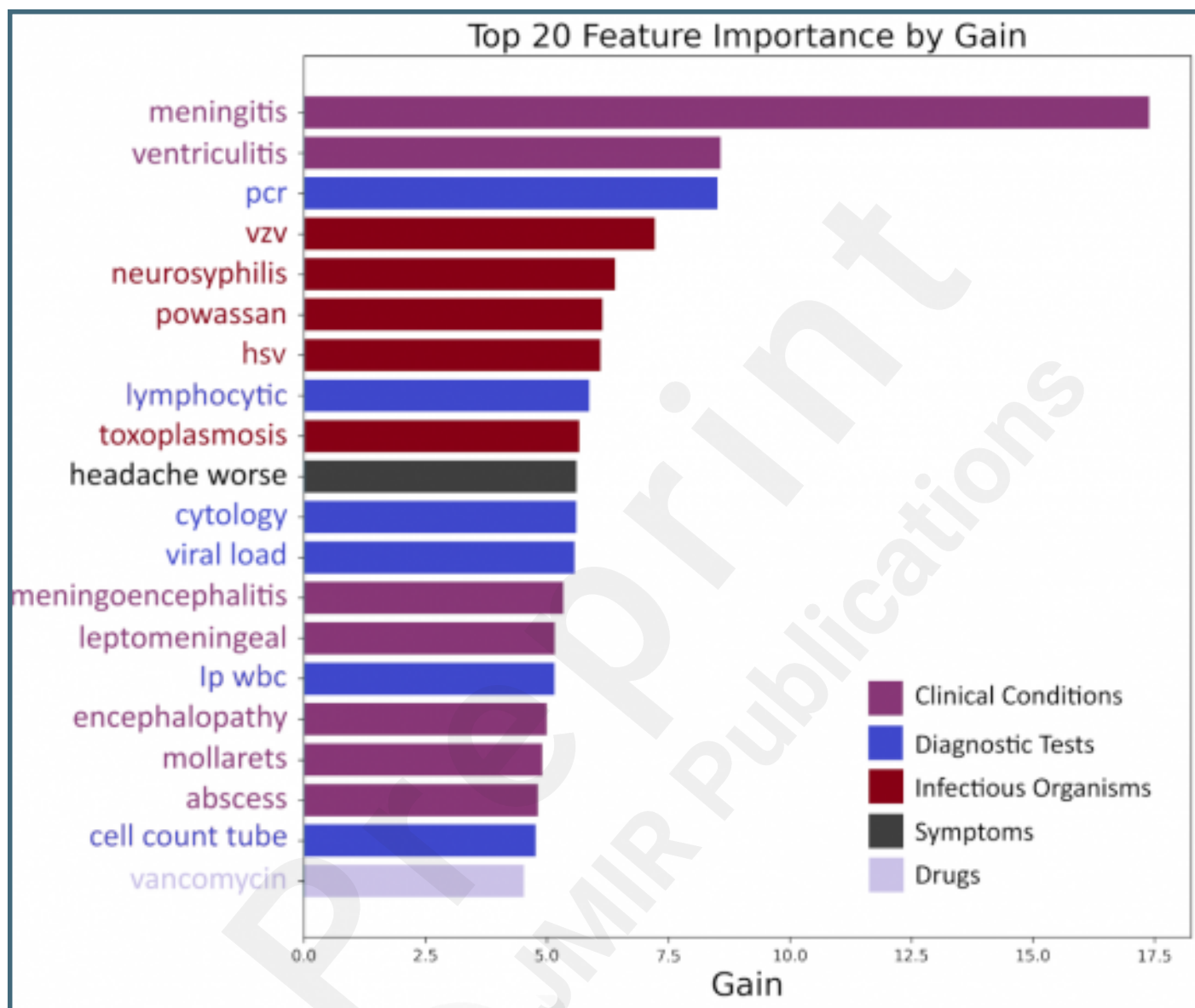
Methods Overview: Data Retrieval and Machine Learning Pipeline for NID Classification. Abbreviations: AUPRC (Area Under the Precision-Recall Curve), AUROC (Area Under the Receiver Operating Characteristic Curve), BIDMC (Beth Israel Deaconess Medical Center), BOW (Bag-of-Words), EHR (Electronic Health Records), ICD (International Classification of Diseases), LP (Lumbar Puncture), n (number of unique patients), N (number of unique notes), MGB (Mass General Brigham), NID (Neurological Infectious Diseases).



XGBoost model performance on two sets of unseen notes, as evaluated by AUROC and AUPRC. 95% CIs were estimated using 1000 bootstrapping iterations. (A) AUROC and (B) AUPRC on notes from MGB and (C) AUROC and (D) AUPRC on notes from BIDMC. AUROC, area under the receiver operating characteristic (ROC) curve; AUPRC, area under the precision recall curve; CI, confidence interval.



Top 20 features ranked by the feature importance (information gain measured by entropy), arranged in descending order (from top to bottom). The x-axis is the feature importance. It does not differentiate the association directions, i.e., positive, or negative association with the NID label.



Multimedia Appendixes

List of ICD-9 and ICD-10 codes for NID.

URL: <http://asset.jmir.pub/assets/90fd27a8c6123af59295bceebbb6e2ba.xlsx>

List of positive expressions.

URL: <http://asset.jmir.pub/assets/fd8554e580385b493027243d8d2dc5e7.xlsx>

List of negative expressions.

URL: <http://asset.jmir.pub/assets/6918872c21d321a2814ab068c6662507.xlsx>

List of antimicrobial expressions.

URL: <http://asset.jmir.pub/assets/496d5d92bf90e949d7376e029441fb39.xlsx>

List of stopwords.

URL: <http://asset.jmir.pub/assets/24d5bb9bb605de3ae210f2a6f2a57086.txt>