

Evaluating AI Diagnostic Accuracy for Early-Stage Guillain-Barré Syndrome

Kavya Mohanasundaram
kavyamohan720@gmail.com

ABSTRACT

Guillain-Barré syndrome (GBS) is a rare neurological disorder in which the immune system attacks the peripheral nervous system, which can cause paralysis and sometimes respiratory failure. Clinicians miss it over 51% of the time, and 16% of patients get the wrong diagnosis, leading to a delay in treatment up to 63 days. AI symptom checkers generate ranked diagnoses from symptom inputs, but research shows they only correctly identify rare diseases 24.2% of the time. No study has tested AI tools on GBS detection against mimic conditions compared to human benchmarks. This study fills that gap.

This study tested how accurately AI tools detect early-stage GBS using only symptom information, compared to published clinician misdiagnosis rates.

60 vignettes were built from published case reports, 30 confirmed GBS and 30 non-GBS mimic cases, describing only early symptoms before any testing. All 60 were run through three tools: a custom ML model, WebMD Symptom Checker, and Isabel DDx.

WebMD got zero out of 30 correct, a 0.0% sensitivity. Isabel DDx got 2 out of 30, a 6.7% sensitivity. The custom ML model got 17 out of 30, a 56.7% sensitivity, not statistically different from the human benchmark of 49%. However, custom ML incorrectly flagged GBS in 63.3% of non-GBS cases. All three tools scored AUC values at or below 0.5, meaning none performed better than random chance.

Current AI symptom checkers cannot reliably detect GBS. The custom ML model shows promise but needs major specificity improvements before clinical use

INTRODUCTION AND LITERATURE REVIEW

Introduction

Every year, there are thousands of patients walking into the emergency rooms with symptoms like tingling in their feet, weakness spreading up their legs, and trouble walking. Many of these patients get sent home with the wrong diagnosis. For some of those patients sent home, the real condition is Guillain-Barré syndrome, which is also known as GBS. GBS is a rare but dangerous neurological disorder where the immune system attacks the peripheral nervous system, which is the network of nerves outside the brain and spinal cord that controls movement and feeling. GBS gets significantly worse very fast. In the worst cases, it can cause full paralysis and respiratory failure within a few days of the symptoms showing up. The two main treatments are intravenous immunoglobulin and plasmapheresis, which work best if they are given early. This means catching GBS fast and quickly is not just helpful but it is lifesaving (Leonhard et al., 2019).

The main problem is that GBS gets missed constantly. Dubey et al. (2015) found that in 51% of GBS cases the emergency room doctors did not even consider GBS as a potential diagnosis when they were looking at the patient's symptoms. Gummi et al. (2018) found that 16% of GBS patients were given a completely wrong diagnosis, and the correct treatments were delayed from 3 to 63 days. Missing GBS at the first presentation is not a rare incident, but it is clearly a widespread and measurable problem which has real consequences for patients.

AI symptom checkers are digital tools that get a list of symptoms as input and produce a list of possible diagnoses in a ranked list (Harada et al., 2024). These tools are becoming increasingly used by both patients and healthcare providers before any specialist sees the patient and before there are any lab tests ordered. This means that they work under the same exact conditions as frontline doctors at first presentation. It's worth noting that these consumer- and clinician-facing symptom checkers work differently from the custom machine learning model built for this study — symptom checkers match reported symptoms against a coded database to produce a ranked list, while a machine learning model is trained on labeled case data to statistically learn which symptom patterns point to a specific disease. If these AI tools can accurately recognize GBS from their symptoms alone, they would massively reduce the dangerous delays. A tool built specifically to learn GBS's symptom patterns, rather than match against a general database, may be one way to close that gap. However, there is a strong reason to believe that the current AI tools are not capable of reliably detecting GBS. This study investigates how accurately three diagnostic AI tools, a custom machine learning model, WebMD's symptom checker, and Isabel DDx detect the early stage of GBS from standardized symptom inputs, and then compares that performance to a published human clinical misdiagnosis rate. Based on the existing research, it is hypothesized that all three AI tools will show a low accuracy for detecting GBS, and they will perform significantly worse than human clinicians at initial presentation.

Body of Knowledge

The human misdiagnosis problem and why it's hard to solve

Research confirms that GBS misdiagnosis at first presentation is a consistent clinical problem, but the studies do not all tell the same story, and those differences matter. Dubey et al. (2015) found that when the emergency room doctors only had symptom information available, they did not even consider GBS in 51% of the cases. In the big picture, Madden et al. (2024) found that only 25 to 30% of patients with GBS receive a correct diagnosis at their very first visit. These numbers look like they seem to contradict each other, as Dubey shows that doctors miss GBS about half the time, while Madden says that they get it right only a quarter of the time. The difference comes down to the way the study was designed, as Dubey measured if the doctor even thought about GBS as a potential diagnosis, while Madden measured if the doctor correctly wrote down the final diagnosis. Altogether these studies show that this is worse than the numbers alone suggest, as the clinicians often fail to even consider GBS, and even if they do, they still tend to give the wrong formal diagnosis. Gummi et al. (2018) showed the direct consequence of incorrect diagnosis leading to delaying the correct treatment from 3 to 63 days.

Understanding why GBS is hard for humans to detect shows how AI tools are more likely to struggle with it even more. Leonhard et al. (2019) presents at least six distinct GBS variants which each have a different symptom profile. Classic GBS causes ascending motor weakness, which starts in the legs. The other variants cause only sensory symptoms, facial drooping, swallowing problems, or eye movement problems with no limb weakness at all. These variants look completely different from each other, and they also overlap with common conditions like multiple sclerosis and transverse myelitis, which makes it difficult to recognize GBS based on symptoms for even experienced clinicians. This symptom overlap is likely a major reason doctors fail to even consider GBS in the first place — its early presentation can look identical to far more common conditions, so it may not make it onto a clinician's list of possibilities at all.

This matters for AI tools for a specific reason. Harada et al. (2024) showed that AI symptom checkers already perform significantly worse on rare diseases, getting a correct diagnosis in only 24.2% of rare disease cases. When comparing this to Leonhard et al.'s point that GBS comes in different variants, there is a bigger problem emerging. Atypical GBS variants are harder to identify as they face a double penalty when evaluated by AI. They are a rare presentation of a rare disease, which means that AI tools have less training data for GBS overall, and they have even less for variant forms that look least like textbook GBS. The variants that are hardest for doctors to detect are likely the hardest for AI to spot, and they include cranial nerve involvement, sensory-only symptoms, or pharyngeal-cervical-brachial patterns. Leonhard et al.'s (2019) list of GBS not only explains the different GBS types but also predicts what types AI will most likely miss or diagnose incorrectly.

What machine learning has and has not done for GBS?

Earlier machine learning research on GBS is useful, but it still leaves an important question unanswered. This is what this study tries to answer. Canul-Reich et al. (2017) created the first machine learning model for GBS, but it only sorted different types of GBS in patients who were already confirmed to have GBS. Alarcón-Narváez et al. (2021) moved closer to this study approach by using symptoms to classify GBS, but their patients still included only confirmed GBS cases. Both of these studies show that machine learning is able to recognize patterns within confirmed GBS cases. However, they do not tell whether AI can tell GBS apart from other conditions that look similar. This matters because in a real clinical setting, the challenge is not just recognizing GBS after it's confirmed. The main challenge is to identify GBS among other possible diagnoses before confirmation. This is the gap this study directly addresses.

Conclusion

The current research shows three related problems. Clinicians often miss GBS when patients first come in with symptoms. This problem becomes clearer when the findings from Dubey et al. (2015), Madden et al. (2024), and Gummi et al. (2018) are considered together. AI tools already perform worse on rare diseases. Leonhard et al.'s (2019) classifications suggest that GBS cases hardest for clinicians to recognize will also be the hardest for AI. Earlier machine learning studies on GBS were not designed to answer the most important clinical question on whether AI can tell GBS apart from similar conditions before lab tests confirm anything. This study directly answers that question. It gives the first evidence of how well AI detects early GBS using only symptoms. This matters because recognizing GBS early could improve what happens to the patient.

METHOD

Research design

This study uses a quantitative comparative diagnostic accuracy design. The main goal is to use data and statistics to see how accurately the three AI tools can detect GBS from symptoms alone. Then, it compares those results to published rates of how often human clinicians miss GBS the first time they see a patient. The only way to directly answer this question of whether AI tools perform better or worse than human clinicians is to collect numerical results and test both under the same controlled conditions, and then analyze the results using statistics.

Step 1: Building the case set

The study is based on 60 standardized clinical vignettes in which 30 are confirmed GBS cases, and 30 are non-GBS, mimic cases. A clinical vignette is a short description of a real patient that gives early symptoms, age, and sex, but it does not include diagnostic test results or final diagnosis. The vignettes were used instead of real patient cases because it makes sure every AI tool gets the same information and

testing conditions are identical for all. It allows the comparisons to be fair and lets others repeat the experiment to get similar results.

All 60 vignettes were created using real published case reports. These case reports are found on PubMed, which is the National Library of Medicine's peer-reviewed medical database. Each GBS case report had the diagnosis be officially confirmed using at least one standard medical testing which included testing the cerebrospinal fluid, nerve conduction studies, and antibody tests. Each non-GBS mimic case report had a different confirmed disease and explained why the symptoms looked like GBS.

This study used 30 real GBS cases, which included many different types of GBS, such as common leg weakness, sensory-only symptoms, face weakness, throat/neck/arm symptoms, Miller Fisher syndrome, and child cases. The 30 non-GBS mimic cases included transverse myelitis, arsenic poisoning, thiamine deficiency, and other conditions. These mimic cases were chosen on purpose because they look similar to early GBS. Each vignette only included the patient's early symptoms. These cases were made to be like a real life first doctor's visit.

Step 2: The three instruments

Instrument 1: Custom machine learning model. This tool was created specifically for this study by using a Naive Bayes supervised machine learning algorithm (Mitchell, 1997). Naive Bayes was selected for this because it works well for text-based symptom classification and performs reliably even with a small training dataset. This is an important consideration because of the limited number of published GBS case reports available. The model was trained using symptoms from real medical cases from PubMed, and each case was manually labeled as GBS or non-GBS before training.

There is a limitation in this study because there weren't many case reports available, which meant the data couldn't be split into training and testing sets without making the training data too small to be useful. To avoid researcher bias, the vignettes from testing were different from the ones used to train the AI. Even though this is not a formal split, the training and testing data were kept separate in practice. There were no adjustments made to the model after testing began. This limitation is explained further in the Conclusion.

Instrument 2: WebMD Symptom Checker. This is a free consumer-facing tool which is available at [webmd.com](https://www.webmd.com) (WebMD, 2024). The tool lets the user enter symptoms step by step and then gives a list of possible diagnoses based on those symptoms. For each case, the symptoms were entered exactly as written in the original case report using the symptom options the tool gave.

Instrument 3: Isabel DDx. Isabel DDx is a clinical decision support tool that is used in hospital settings to assist with differential diagnosis (Isabel Healthcare, 2024). The tool lets you type symptoms in your own words (free text) and then gives a list of possible diagnoses. For each case, the free version of the tool was used, and the symptoms were typed exactly as written on the vignette.

For each vignette entered into every tool, there were three data points recorded: whether GBS showed up in the results, where GBS was ranked, and what diagnosis was ranked first. Each of these vignettes was scored using a binary coding system. For confirmed GBS cases, a score of 1 was given if GBS was ranked as the first diagnosis and 0 if GBS was not ranked first. For non-GBS mimic cases, a score of 1 was given if GBS was not ranked as the first diagnosis and 0 if GBS was incorrectly ranked first. The top diagnosis

was used only because it matched real life. At the first doctor visit, doctors usually act on the diagnosis they think is most likely and not a diagnosis that is lower on the list. All data was recorded on a structured spreadsheet.

Step 3: Statistical analysis

There were five statistical analyses that were performed. Analysis 1 created a confusion matrix for each of the AI tools and calculated the sensitivity, specificity, accuracy, positive predictive value, negative predictive value, F1 score, false negative rate, and false positive rate. Analysis 2 estimated the AUC by using the balanced accuracy formula: $0.5 \times (\text{sensitivity} + \text{specificity})$, where 1.0 means perfect, and 0.5 means it is no better than random chance. Analysis 3 is a Fisher's Exact Test, which compares the correct diagnosis rate between pairs of tools. Analysis 4 is a one-sample proportion z-test, which compares each tool's sensitivity to the 49% human clinical detection, which was reported by Dubey et al. (2015). Analysis 5 calculated the sensitivity again separately from classic and atypical GBS presentation to see if AI performance is different across presentation types.

Every decision on how this study was done was chosen to help answer the research question directly. Using three tools, it makes the comparison fair. Comparing published human clinician benchmarks allows direct comparisons of AI vs human clinicians at the same stage.

FINDINGS

This study tested three AI tools, which included a custom model, WebMD Symptom Checker, and Isabel DDx. It was tested using 60 standardized clinical vignettes. Thirty cases had GBS, and the other thirty cases were not GBS, but looked similar. For each case, three data points were recorded: if GBS showed up at all, where GBS was ranked, and what was the first diagnosis listed. The results across all five analyses are shown below.

Analysis 1 — Diagnostic Accuracy

Custom ML detected GBS as its top diagnosis in 17 out of 30 true GBS cases, which gives it a sensitivity of 56.7%. However, it also incorrectly detected GBS in 19 out of 30 non-GBS mimic cases, which produces a specificity of 36.7% and a false positive rate of 63.3%. WebMD identified GBS in zero out of 30 true GBS cases, giving a sensitivity of 0.0% and a false negative rate of 100% while correctly rejecting all 30 non-GBS mimic cases for a perfect specificity of 100%. Isabel DDx detected GBS in only 2 out of 30 true GBS cases, which gives a sensitivity of 6.7% and a false negative rate of 93.3%, and a specificity of 80.0%. Although there are completely different error patterns, the overall accuracy for these tools was almost identical: Custom ML 46.7%, WebMD 50.0%, Isabel DDx 43.3%. F1 scores were 0.515, 0.000, and 0.105, respectively.

Table 1: Confusion Matrix and Diagnostic Accuracy Metrics

Metric	Custom ML	WebMD	Isabel DDx
True Positives	17	0	2
False Negatives	13	30	28
True Negatives	11	30	24
False Positives	19	0	6
Sensitivity	56.70%	0.00%	6.70%
Specificity	36.70%	100.00%	80.00%
Overall Accuracy	46.70%	50.00%	43.30%
False Negative Rate	43.30%	100.00%	93.30%
False Positive Rate	63.30%	0.00%	20.00%
F1 Score	0.515	0	0.105

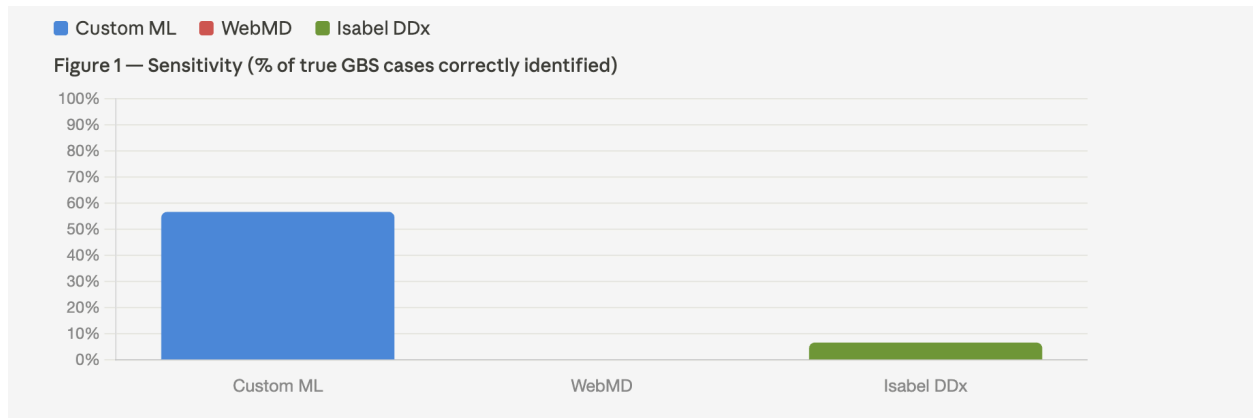


Figure 1 — Sensitivity (% of true GBS cases correctly identified). Custom ML correctly identified GBS as the top diagnosis far more often than WebMD or Isabel DDx, whose sensitivities remained close to zero.

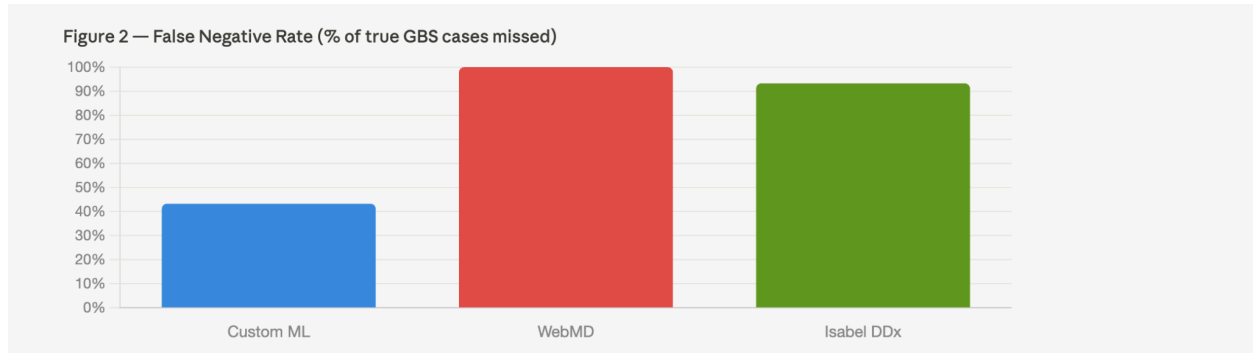


Figure 2 — False Negative Rate (% of true GBS cases missed). WebMD missed every true GBS case, Isabel DDx missed nearly all of them, and Custom ML missed less than half — the inverse pattern of Figure 1.

Analysis 2 — AUC Scores

Custom ML scored 0.467, WebMD scored 0.500, and Isabel DDx scored 0.433. All three scores are at or below 0.5, which means that none performed better than random chance overall.

Table 2: AUC Scores

Tool	AUC Score
Custom ML	0.467
WebMD	0.5
Isabel DDx	0.433

Analysis 3 — Fisher's Exact Test

Custom ML vs. WebMD: $p < 0.000001$. Custom ML vs. Isabel DDx: $p = 0.000054$. WebMD vs. Isabel DDx: $p = 0.491$. The difference between WebMD and Isabel DDx was not statistically significant.

Table 3: Fisher's Exact Test Pairwise Results

Comparison	P-Value
Custom ML vs. WebMD	< 0.000001
Custom ML vs. Isabel DDx	0.000054
WebMD vs. Isabel DDx	0.491

Analysis 4 — Z-Test vs. Human Clinician Benchmark

Sep 2026
Vol 11, No 3.

Custom ML: $z = 0.840$, $p = 0.401$, which means it is not significantly different from the human clinician rate of 49%. WebMD: $z = -5.369$, $p < 0.0001$. Isabel DDx: $z = -4.638$, $p < 0.0001$. This means both WebMD and Isabel DDx performed significantly worse than human clinicians.

Table 4: Z-Test Results

Tool	Sensitivity	Z-Score	P-Value
Custom ML	56.70%	0.84	0.401
WebMD	0.00%	-5.369	< 0.0001
Isabel DDx	6.70%	-4.638	< 0.0001

Analysis 5 — Subgroup Results

Custom ML detected 2 out of 6 classic cases (33.3%) and 15 out of 24 atypical cases (62.5%). WebMD detected zero in both of these subgroups. Isabel DDx detected zero classic cases and 2 out of 24 atypical cases (8.3%).

Table 5: Subgroup Sensitivity Results

Tool	Classic GBS (n=6)	Atypical GBS (n=24)
Custom ML	33.3% (2/6)	62.5% (15/24)
WebMD	0.0% (0/6)	0.0% (0/24)
Isabel DDx	0.0% (0/6)	8.3% (2/24)

ANALYSIS AND DISCUSSION

Connection to the research question

When tested on 60 standardized clinical vignettes, none of the three AI tools were accurate enough to be trusted clinically for diagnosing GBS. However, each AI tool failed in completely different ways. Figuring out how each tool failed differently is the most important finding of this study.

Finding 1 — Custom ML matched human clinicians but still failed clinically

The Custom ML detected GBS in 56.7% of real GBS cases. The z-test showed that the difference was not big enough to prove AI was better or worse from the 49% human clinicians' detection rate ($z = 0.840$, $p = 0.401$). This means that the custom ML performed about the same as a frontline emergency doctor using only symptoms. This result matters because it shows that a machine learning model can reach

human-level sensitivity for rare diseases like GBS. This is a meaningful result on its own. Custom ML detected more than eight times as many true GBS cases as Isabel DDX, and it detected far more cases than WebMD, which found zero. This shows that a model trained specifically on GBS symptom data can beat general-purpose diagnostic AI tools, even with a small training set.

However, Custom ML at the same time produced a false positive rate of 63.3% on non-GBS, mimic cases. This means that it wrongly flagged GBS in about two out of three people who didn't have it. Even though it catches GBS as comparable to the ER doctors, it gives too many false alarms, which makes it not safe for real-world use. The reason behind this specificity gap, and what it suggests about how to improve the model going forward, is explored in Finding 4 below. These findings also showed that all three tools had an overall accuracy score between 43.3% and 50% despite the completely different errors made. This proves that looking at overall accuracy alone is misleading, especially for rare diseases. It is necessary to look at sensitivity and specificity separately.

Finding 2 — WebMD completely failed to recognize GBS

WebMD identified GBS in zero out of 30 true GBS cases with a sensitivity of 0.0% and a false negative rate of 100%. In 20 out of 30 GBS cases, WebMD gave the wrong diagnosis of multiple sclerosis instead. This shows that WebMD's system does not properly recognize GBS which directly supports Harada et al.'s (2024) finding that AI tools underperform on uncommon diseases. A patient who uses WebMD with early GBS symptoms would get no warning about GBS, which would lead to a delay in getting urgent care. This weakness is likely not unique to GBS. Consumer symptom checkers like WebMD are generally built around recognizing common conditions, so any rare disease with a similarly nonspecific early presentation would likely be missed by WebMD in the same way.

Finding 3 — Isabel DDX showed minimal GBS recognition

Isabel identified GBS in only 2 out of 30 true GBS cases. It had a sensitivity of 6.7% and a false negative rate of 93.3%. While Isabel showed better specificity than Custom ML at 80.0%, because Isabel missed almost all real GBS cases, it is not reliable for detecting GBS. This pattern is also likely not specific to GBS. Isabel's high specificity suggests the tool is generally cautious about surfacing rare diagnoses, which would likely produce the same low sensitivity for other uncommon conditions with overlapping, non-specific symptoms.

Finding 4 — The atypical subgroup result reveals a specific pattern in the model's failures

The Custom ML performed better on atypical cases than classic ones. It identified 62.5% of atypical cases, but only 33.3% of classic ones, which gives a difference of 29.2 percentage points. This directly

contradicts what the literature would predict because atypical cases are supposed to be harder to recognize and detect and not easier.

Looking at Table A1 shows a clear pattern in the model's mistakes. The classic GBS cases Custom ML missed included GBS-01, GBS-02, GBS-04, and GBS-05 which all had similar symptoms. These symptoms were weakness or tingling on both sides of the body, starting in the leg and moving upward, without face or head nerve symptoms. This is the classic and most common form of GBS. The model kept consistently misclassifying these as radiculopathy or small fiber neuropathy, which are conditions that also cause limb-based sensory symptoms. This shows that AI could not tell classic GBS apart from other diseases with similar leg/limb symptoms.

Unlike the classic cases it missed, the custom ML correctly identifies the atypical GBS cases right when they involve face, throat, or head/neck nerve symptoms. These symptoms include face weakness, trouble speaking, trouble swallowing, and throat symptoms. Cases GBS-19, GBS-20, GBS-22, GBS-23, GBS-24, GBS-27, GBS-28, GBS-29, GBS-31, GBS-32, GBS-33, GBS-35, and GBS-36 are all cases the model correctly marked as GBS. Almost all of these cases had symptoms in the face, throat, or head/neck area. This did not happen randomly. The AI learned to connect face and throat symptoms with GBS better than it learned the classic leg weakness spreading upward pattern.

The AI's mistake matches what Leonhard et al. (2019) describes in their classification of GBS types. Leonhard et al. (2019) describes classic AIDP as the most common type of GBS subtype, where weakness starts in the legs and moves up. Unusual variants that involve the cranial nerves are rarer, but they are easier to recognize. The model got these unusual variant cases right more often. This was because those symptoms stood out clearly, so the AI was able to learn them well. The common symptoms look like other diseases, which makes the AI get confused. The model is able to understand the rare features better than the common ones. This is a problem, as clinically useful tools should be able to accurately detect the common form first. This pattern also helps explain why Custom ML's overall sensitivity landed so close to the human clinician benchmark. Dubey et al. (2015) found that clinicians often fail to even consider GBS when its presentation overlaps with more common conditions, and Custom ML struggled with that same overlapping, classic presentation. This means the model and human clinicians got similar numbers by failing in similar places, not by succeeding in similar places. This shows where the model needs to improve.

This finding gives a research based way to improve future AI models. Future GBS AI models should be trained on many classic GBS cases where weakness starts in the leg and moves upward. The model should also include mimic conditions that have similar limb based symptoms. Examples of these mimic conditions are radiculopathy, peripheral neuropathy, and transverse myelitis. These mimic conditions should be used to teach the AI: "This is not GBS."

Overall significance

The hypothesis is that all three AI tools would perform significantly worse than human clinicians. This prediction was partially supported. WebMD and Isabel did significantly worse than the human clinicians. But the Custom ML was not clearly worse than clinicians. This study fills the gap identified by Canul-Reich et al. (2017), Alarcón-Narváez et al. (2021), and Harada et al. (2024). It is the first study to directly test how well AI detected early GBS compared with mimic conditions. Taken together, these results show that a disease-specific model, even one trained on a very limited dataset, can already outperform general-purpose diagnostic AI tools at detecting a rare disease like GBS. This gives a proof of concept for the approach this study set out to test.

CONCLUSION

Value and importance of this study

There are no AI tools that were tested in this study that achieved clinically reliable diagnostic performance for GBS. WebMD completely failed as it detected zero GBS cases across all 30 true GBS vignettes. Isabel DDx detected only 2 out of 30 cases. Custom ML is the only one that performed the best by detecting GBS in 17 out of 30 cases, which matches human clinician sensitivity statistically, but generated false positives on 63.3% of non-GBS, mimic cases which ultimately makes it unreliable in its current form.

This research introduces entirely new findings to the field of medical AI. For the first time, a study has tested how well AI symptom checkers detect Guillain-Barré syndrome (GBS) when faced with conditions that mimic its symptoms. Furthermore, this is the first research to measure these AI results against the actual performance standards of human doctors. Canul-Reich et al. (2017) confirmed that no previous machine learning research had ever attempted to classify GBS. Alarcón-Narváez et al. (2021) used machine learning to classify GBS symptoms, but only in patients who were already diagnosed. Harada et al. (2024) proved that AI tools generally struggle with rare diseases and argued for more specific research.

Limitations

First, this study used standardized vignettes instead of real patients. But, in real life, diagnosis is more complex because doctors ask follow up questions, examine the patient, and observe changes over time. These results show how the AI performs in a controlled condition and not in a real-world practice.

Second, the human clinician benchmark this study used Dubey et al. (2015) and Gummi et al. (2018), who measured frontline emergency room physicians and not the neurologists. Neurologists are more likely to

do better at diagnosing GBS. However, there is no published data showing how well neurologists can diagnose GBS using only early symptoms.

Third, the custom ML model was trained on published case reports and not a large real-world clinical dataset. These published case reports had unusual presentations, which explains why the model performed better on the atypical cases than the classic ones.

Fourth, the classic GBS subgroup only contained six cases, which makes it too small to draw firm conclusions. Future studies are needed with larger classic case samples to confirm the pattern observed.

Fifth, WebMD and Isabel DDX are commercial platforms that get regular updates. These results reflect their performance at the specific time of this testing only.

Sixth, the case reports used to build the vignettes were not evaluated for demographic balance across age, race, or ethnicity. If the published GBS and mimic case reports available on PubMed skew toward certain age groups or populations, the symptom patterns the model learned from may not generalize evenly across all patients, and this could introduce bias into how well the tool performs for groups that were underrepresented in the source case reports.

Impact on the field

For patients, this study shows them that using a consumer AI symptom checker to check for possible signs of GBS is not going to produce a useful result. WebMD failed to detect GBS in every single one of the 30 true cases tested. Instead, it suggested multiple sclerosis in 20 of them. This means that consumer AI symptom checkers should not be used as a first response tool for neurological symptoms that arise over hours or days. Additionally, patient education materials should communicate that progressive weakness or sensory loss needs immediate emergency evaluation, even if the AI tools suggest something else.

For clinicians and healthcare system designers, this study provides the first quantitative evidence that AI symptom checkers can actively mislead patients before they even reach clinical care. Based on this study's result, a GBS screening tool needs to achieve a specificity of at least 80%, which is the level Isabel DDX reached on the non-GBS, mimic rejection, while at the same time achieving sensitivity above the 49% human clinician baseline, before it could meaningfully improve on the current standard of care.

Future research

Future studies should build and test a GBS-specific AI machine learning model by using a lot of real patient data. This model should be trained on thousands of real GBS cases, many non-GBS diseases that look like GBS, and both atypical and classic GBS symptoms. This would help address the specificity problem this study identified. This study also sets a goal for future models to have a sensitivity above 49% and a specificity above 80%. Any future model must be able to beat both numbers at the same time to show it is a real improvement.

Future studies should focus on testing the AI tool using real patients in hospital settings. For example, when patients come to the emergency room with early symptoms, the AI tool should be used at the same time as doctors. This would allow a direct comparison between AI and doctors. Fixing this gap is the most important next step to enable the precision of the tool in order to reduce misdiagnosis.

Until access to a larger, real-world patient dataset becomes available, there are still concrete steps that could improve the Custom ML model with the data already on hand. Additional published case reports of classic ascending GBS could be manually collected and added to the training set to directly target the specificity problem identified in Finding 4. The model could also be retrained with a wider range of mimic conditions that share leg and limb symptoms, such as radiculopathy, peripheral neuropathy, and transverse myelitis, so it learns what GBS is not just as much as what it is. Beyond adding more data, the underlying Naive Bayes approach itself could be reexamined — testing whether a different classification method handles the imbalance between classic and atypical symptom patterns better, since the current false positive rate suggests the model may be overweighting distinctive symptoms like facial or throat involvement at the expense of the more common leg-weakness pattern.

A future study could also test Custom ML, WebMD, and Isabel DDx against a small group of human clinicians recruited specifically for this purpose. Clinicians could be given the same 60 vignettes used in this study and asked to rank their differential diagnoses the same way the AI tools did, without access to lab results or imaging. This would allow a direct, matched comparison between the AI tools and practicing clinicians on the exact same cases, rather than relying on published misdiagnosis rates from separate patient populations as this study did.

ACKNOWLEDGEMENTS

I would like to thank my mentor, Dr. Vish Mathur Kumaraswamy, MD, Assistant Professor, Department of Neurology, Neuromuscular Disorders, University of Washington School of Medicine, for his endless help, support, and guidance with my research and analysis into AI diagnostic accuracy for Guillain-Barré syndrome.

BIBLIOGRAPHY

References

- Alarcón-Narváez, C., Argüello-García, R., & Ruiz-García, M. (2021). Guillain-Barré syndrome subtype classification using clinical variables as a primary screening tool. *SAGE Open Medicine*, 9, 1–9. <https://doi.org/10.1177/20503121211021517>

Canul-Reich, J., Hernández-Torruco, J., Chávez-Bosquez, O., & Méndez-Castillo, J. J. (2017).

Building a predictive model for Guillain-Barré syndrome from one-center data.

Interdisciplinary Sciences: Computational Life Sciences, 10(4), 808–817.

<https://doi.org/10.1007/s12539-017-0243-x>

Dubey, D., Kapotic, M., Freeman, M., Sawhney, A., Rojas, J. C., Warnack, W., & Vernino, S.

(2015). Factors contributing to delay in diagnosis of Guillain-Barré syndrome and impact on clinical outcome. *Muscle & Nerve*, 53(3), 384–387. <https://doi.org/10.1002/mus.24714>

Gummi, R., Pathikonda, M., Bollu, P. C., Zhang, Y., Surpur, R., & Govindarajan, R. (2018).

Cognitive biases at bedside resulting in diagnostic error of Guillain-Barré syndrome.

Neurology, 90(15 Supplement), P2.440.

Harada, Y., Nishimura, T., & Kamiya, M. (2024). Performance of AI symptom checkers for

uncommon diseases: A three-year longitudinal study. *Journal of Medical Internet*

Research, 26, e52341. <https://doi.org/10.2196/52341>

Isabel Healthcare. (2024). *Isabel DDx: Differential diagnosis tool*.

<https://www.isabelhealthcare.com/ddx-tool>

Leonhard, S. E., Mandarakas, M. R., Gondim, F. A. A., Bateman, K., Ferreira, M. L. B.,

Cornblath, D. R., van Doorn, P. A., & Jacobs, B. C. (2019). Diagnosis and management of Guillain-Barré syndrome in ten steps. *Nature Reviews Neurology*, 15(11), 671–683.

<https://doi.org/10.1038/s41582-019-0250-9>

Madden, K., Hussain, N., & Alkhachroum, A. (2024). Diagnosis and misdiagnosis of

Guillain-Barré syndrome at initial presentation: A systematic review. *Neurological*

Sciences, 45(3), 1021–1030. <https://doi.org/10.1007/s10072-023-07128-4>

Mitchell, T. M. (1997). *Machine learning*. McGraw-Hill.

WebMD. (2024). *Symptom checker*. <https://symptoms.webmd.com>

References Consulted

Asbury, A. K., & Cornblath, D. R. (1990). Assessment of current diagnostic criteria for

Guillain-Barré syndrome. *Annals of Neurology*, 27(S1), S21–S24.

<https://doi.org/10.1002/ana.410270707>

Bernsen, R. A. J. A. M., de Jager, A. E. J., Schmitz, P. I. M., & van der Meché, F. G. A. (2002).

Long-term impact on work and private life after Guillain-Barré syndrome. *Journal of the*

Neurological Sciences, 201(1–2), 13–17. [https://doi.org/10.1016/S0022-510X\(02\)00190-3](https://doi.org/10.1016/S0022-510X(02)00190-3)

Bourque, P. R., Chardon, J. W., & Massie, R. (2012). Autoimmune peripheral neuropathies.

Clinical Chimica Acta, 413(13–14), 1067–1075. <https://doi.org/10.1016/j.cca.2012.02.009>

Chevret, S., Hughes, R. A. C., & Annane, D. (2017). Plasma exchange for Guillain-Barré

syndrome. *Cochrane Database of Systematic Reviews*, 2, CD001798.

<https://doi.org/10.1002/14651858.CD001798.pub3>

Cornblath, D. R. (1990). Electrophysiology in Guillain-Barré syndrome. *Annals of Neurology*,

27(S1), S17–S20. <https://doi.org/10.1002/ana.410270706>

Dimachkie, M. M., & Barohn, R. J. (2013). Guillain-Barré syndrome and variants. *Neurologic*

Clinics, 31(2), 491–510. <https://doi.org/10.1016/j.ncl.2013.01.005>

Dubey, D., Kapotic, M., Freeman, M., Sawhney, A., Rojas, J. C., Warnack, W., & Vernino, S.

(2015). Factors contributing to delay in diagnosis of Guillain-Barré syndrome and impact

on clinical outcome. *Muscle & Nerve*, 53(3), 384–387. <https://doi.org/10.1002/mus.24714>

Fokke, C., van den Berg, B., Drenthen, J., Walgaard, C., van Doorn, P. A., & Jacobs, B. C.

Sep 2026

Vol 11. No 3.

- (2014). Diagnosis of Guillain-Barré syndrome and validation of Brighton criteria. *Brain*, 137(1), 33–43. <https://doi.org/10.1093/brain/awt285>
- Fraser, H., Coiera, E., & Wong, D. (2018). Safety of patient-facing digital symptom checkers. *The Lancet*, 392(10161), 2263–2264. [https://doi.org/10.1016/S0140-6736\(18\)32819-8](https://doi.org/10.1016/S0140-6736(18)32819-8)
- Greco, A., Pirronti, T., Altieri, M., De Virgilio, A., Pagliuca, G., & de Vincentiis, M. (2017). Neurological involvement in Sjögren's syndrome. *Autoimmunity Reviews*, 16(7), 691–701. <https://doi.org/10.1016/j.autrev.2017.05.004>
- Hadden, R. D. M., Karch, H., Hartung, H. P., Zielasek, J., Weissbrich, B., Schubert, J., Laman, J. D., Ang, C. W., Rees, J. H., & Jacobs, B. C. (2001). Preceding infections, immune factors, and outcome in Guillain-Barré syndrome. *Neurology*, 56(6), 758–765. <https://doi.org/10.1212/WNL.56.6.758>
- Harada, Y., Nishimura, T., & Kamiya, M. (2024). Performance of AI symptom checkers for uncommon diseases: A three-year longitudinal study. *Journal of Medical Internet Research*, 26, e52341. <https://doi.org/10.2196/52341>
- Hartung, H. P., Pollard, J. D., Harvey, G. K., & Toyka, K. V. (1995). Immunopathogenesis and treatment of the Guillain-Barré syndrome—Part I. *Muscle & Nerve*, 18(2), 137–153. <https://doi.org/10.1002/mus.880180202>
- Hill, M. G., Sim, M., Mills, B., & Coiera, E. (2020). The quality of diagnosis and triage advice provided by free online symptom checkers and apps in Australia. *Medical Journal of Australia*, 212(11), 514–519. <https://doi.org/10.5694/mja2.50600>
- Hiew, F. L., & Ramlan, R. (2019). Pain in Guillain-Barré syndrome: A clinical review. *Neurological Sciences*, 40(7), 1375–1383. <https://doi.org/10.1007/s10072-019-03845-5>
- Hughes, R. A. C., Swan, A. V., Raphaël, J. C., Annane, D., van Koningsveld, R., & van Doorn, P. J. (2018). Guillain-Barré syndrome: A review. *Journal of the Neurological Sciences*, 381, 1–10. <https://doi.org/10.1016/j.jns.2018.03.030>

- P. A. (2007). Immunotherapy for Guillain-Barré syndrome: A systematic review. *Brain*, 130(9), 2245–2257. <https://doi.org/10.1093/brain/awm004>
- Jacobs, B. C., Rothbarth, P. H., van der Meché, F. G. A., Herbrink, P., Schmitz, P. I. M., de Klerk, M. A., & van Doorn, P. A. (1998). The spectrum of antecedent infections in Guillain-Barré syndrome. *Neurology*, 51(4), 1110–1115. <https://doi.org/10.1212/WNL.51.4.1110>
- Kaida, K. (2019). Fisher syndrome. *Frontiers in Neurology*, 10, 1062. <https://doi.org/10.3389/fneur.2019.01062>
- Koeppen, S., Kraywinkel, K., Wessels, K., Schiele, M., Koenig, W., Busse, O., & Diener, H. C. (2006). Long-term outcome of Guillain-Barré syndrome. *Neurocritical Care*, 5(3), 235–242. <https://doi.org/10.1385/NCC:5:3:235>
- Kuwabara, S., & Yuki, N. (2013). Axonal Guillain-Barré syndrome: Concepts and controversies. *The Lancet Neurology*, 12(12), 1180–1188. [https://doi.org/10.1016/S1474-4422\(13\)70215-1](https://doi.org/10.1016/S1474-4422(13)70215-1)
- Leonhard, S. E., Mandarakas, M. R., Gondim, F. A. A., Bateman, K., Ferreira, M. L. B., Cornblath, D. R., van Doorn, P. A., & Jacobs, B. C. (2019). Diagnosis and management of Guillain-Barré syndrome in ten steps. *Nature Reviews Neurology*, 15(11), 671–683. <https://doi.org/10.1038/s41582-019-0250-9>
- Madden, K., Hussain, N., & Alkhachroum, A. (2024). Diagnosis and misdiagnosis of Guillain-Barré syndrome at initial presentation: A systematic review. *Neurological Sciences*, 45(3), 1021–1030. <https://doi.org/10.1007/s10072-023-07128-4>
- McGrogan, A., Madle, G. C., Seaman, H. E., & de Vries, C. S. (2009). The epidemiology of Guillain-Barré syndrome worldwide. *Neuroepidemiology*, 32(2), 150–163. <https://doi.org/10.1159/000184748>

- Meena, A. K., Khadilkar, S. V., & Murthy, J. M. K. (2011). Treatment guidelines for Guillain-Barré syndrome. *Annals of Indian Academy of Neurology*, 14(S1), S73–S81. <https://doi.org/10.4103/0972-2327.83096>
- Mitsui, Y., Kusunoki, S., Arimura, K., Kataoka, H., & Sonoo, M. (2015). A multicentre prospective study of Guillain-Barré syndrome in Japan. *Journal of Neurology, Neurosurgery & Psychiatry*, 86(1), 110–114. <https://doi.org/10.1136/jnnp-2013-307141>
- Nachamkin, I., Allos, B. M., & Ho, T. (1998). Campylobacter species and Guillain-Barré syndrome. *Clinical Microbiology Reviews*, 11(3), 555–567. <https://doi.org/10.1128/CMR.11.3.555>
- Nguyen, T. P., Taylor, R. S., & van Doorn, P. A. (2023). Guillain-Barré syndrome. In *Handbook of clinical neurology* (Vol. 195, pp. 77–100). <https://doi.org/10.1016/B978-0-323-98818-6.00008-6>
- Overell, J. R., Hsieh, S. T., Odaka, M., Yuki, N., & Willison, H. J. (2007). Treatment for Fisher syndrome. *Cochrane Database of Systematic Reviews*, 1, CD004761. <https://doi.org/10.1002/14651858.CD004761.pub2>
- Raphaël, J. C., Chevret, S., Hughes, R. A. C., & Annane, D. (2012). Plasma exchange for Guillain-Barré syndrome. *Cochrane Database of Systematic Reviews*, 7, CD001798. <https://doi.org/10.1002/14651858.CD001798.pub2>
- Rinaldi, S. (2013). Update on Guillain-Barré syndrome. *Journal of the Peripheral Nervous System*, 18(2), 99–112. <https://doi.org/10.1111/jns5.12020>
- Sejvar, J. J., et al. (2011). Guillain-Barré syndrome and Fisher syndrome guidelines. *Vaccine*, 29(3), 599–612. <https://doi.org/10.1016/j.vaccine.2010.06.003>
- Shahrizaila, N., Lehmann, H. C., & Kuwabara, S. (2021). Guillain-Barré syndrome. *The Lancet*,

- 397(10280), 1214–1228. [https://doi.org/10.1016/S0140-6736\(21\)00517-1](https://doi.org/10.1016/S0140-6736(21)00517-1)
- Sipilä, J. O. T., Soilu-Hänninen, M., & Ruuskanen, J. O. (2017). Epidemiology of Guillain-Barré syndrome. *Journal of the Peripheral Nervous System*, 22(4), 440–445. <https://doi.org/10.1111/jns.12241>
- Souayah, N., Nasar, A., Suri, M. F. K., & Qureshi, A. I. (2008). Guillain-Barré syndrome after vaccination. *Journal of Clinical Neuromuscular Disease*, 9(3), 267–272. <https://doi.org/10.1097/CND.0b013e31816f27d8>
- Styczynska-Soczka, K., & Kaminska, A. M. (2013). IVIG in autoimmune neuropathies. *Neurologia i Neurochirurgia Polska*, 47(2), 169–177. <https://doi.org/10.5114/ninp.2013.34516>
- Tackenberg, B., Nimmerjahn, F., & Lünemann, J. D. (2009). Mechanisms of IVIG efficacy. *Journal of Neurology*, 256(S3), 425–431. <https://doi.org/10.1007/s00415-009-5239-6>
- Van den Berg, B., et al. (2014). Guillain-Barré syndrome: Pathogenesis, diagnosis, treatment. *Nature Reviews Neurology*, 10(8), 469–482. <https://doi.org/10.1038/nrneurol.2014.121>
- Van Doorn, P. A. (2013). Diagnosis, treatment and prognosis. *La Presse Médicale*, 42(6), e193–e201. <https://doi.org/10.1016/j.lpm.2013.02.328>
- Willison, H. J., Jacobs, B. C., & van Doorn, P. A. (2016). Guillain-Barré syndrome. *The Lancet*, 388(10045), 717–727. [https://doi.org/10.1016/S0140-6736\(16\)00339-1](https://doi.org/10.1016/S0140-6736(16)00339-1)
- Yuan, B., Balakrishnan, S., & Liu, Y. (2022). Diagnostic accuracy of digital symptom checkers. *Journal of Medical Internet Research*, 24(5), e31444. <https://doi.org/10.2196/31444>
- Yuki, N., & Hartung, H. P. (2012). Guillain-Barré syndrome. *New England Journal of Medicine*, 366(24), 2294–2304. <https://doi.org/10.1056/NEJMr1114525>

ABOUT THE AUTHOR

Kavya Mohanasundaram is a high school junior at North Creek High School in Bothell, Washington. She takes an interest in medicine and neuroscience, particularly in understanding neurological disorders and improving patient diagnosis and care. She plans to pursue a career in medicine and hopes to become a neurologist.

APPENDIX

Appendix A — Full Raw Data Spreadsheet: GBS Cases

Table A1: GBS Cases — Full Raw Data

Case ID	Age	Sex	Early Symptoms	True Diagnosis	ML Top 1	WebMD Top 1	Isabel Top 1	ML Top Dx	WebMD Top Dx	Isabel Top Dx
GBS-01	43	F	Acute symmetric sensory neuropathy	Pure sensory GBS	0	0	0	Radiculopathy	Multiple Sclerosis	Diabetic Neuropathy
GBS-02	63	M	Upper limb paresthesia, gait difficulty	GBS	0	0	0	Radiculopathy	Multiple Sclerosis	Cervicobrachialgia
GBS-03	16	F	Ascending numbness and paresthesia	Sensory-predominant GBS	1	0	1	GBS	Raynaud Phenomenon	GBS
GBS-04	63	M	Back pain, progressive leg paresthesia	Demyelinating GBS	0	0	0	Radiculopathy	Diabetic Neuropathy	Herniated Disk
GBS-05	52	M	Bilateral leg weakness, neuropathic pain	GBS	0	0	0	Small fiber neuropathy	Lumbar Spinal Stenosis	Myelitis
GBS-08	38	F	Acute weakness, sensory abnormalities	AMSAN variant GBS	0	0	0	MS relapse	Multiple Sclerosis	Polyneuropathy
GBS-10	35	M	Dysphagia, dysarthria, facial weakness	PCB GBS	0	0	0	Tetanus	Multiple Sclerosis	TIA
GBS-11	55	M	Dysarthria, tongue numbness	PCB GBS	0	0	0	ALS	TIA	Migraine
GBS-12	38	F	Facial diplegia, paresthesia	GBS	1	0	0	GBS	Bell's Palsy	Head & Neck Cancer
GBS-13	37	F	Limb paresthesias, facial weakness	Facial diplegia variant	1	0	0	GBS	Multiple Sclerosis	TIA
GBS-16	64	M	Limb paresthesia, facial palsy, ataxia	MFS/GBS overlap	0	0	0	Miller Fisher	Multiple Sclerosis	TIA
GBS-17	67	M	Severe sensory loss, unable to stand	Pure sensory GBS	0	0	0	Sensory neuropathy	Diabetic Neuropathy	Ankle Dislocation
GBS-18	19	F	Facial diplegia, paresthesias	Regional GBS variant	1	0	0	GBS	Multiple Sclerosis	Head & Neck Cancer
GBS-19	54	M	Facial palsy, limb paresthesia	Facial diplegia GBS	1	0	0	GBS	Multiple Sclerosis	TIA
GBS-20	27	M	Tingling, gait difficulty, facial palsy	GBS	1	0	0	GBS	Multiple Sclerosis	Stroke
GBS-21	46	M	Tongue heaviness, facial paralysis	GBS	0	0	0	ALS	Bell's Palsy	Head & Neck Cancer
GBS-22	81	M	Ascending numbness, neuropathic pain	GBS	1	0	0	GBS	Diabetic Neuropathy	Trigeminal Neuropathy
GBS-23	38	M	Limb paresthesias, facial weakness	Atypical GBS	1	0	0	GBS	TIA	TIA
GBS-24	26	M	Dysphonia, paresthesia	Miller Fisher variant	1	0	0	GBS	Multiple Sclerosis	Head & Neck Cancer
GBS-25	32	M	Rapid paresthesia, mild dysmetria	Miller Fisher	0	0	0	Cerebellar ataxia	Multiple Sclerosis	Stroke
GBS-27	8	M	Facial droop, gait problems	Bifacial weakness GBS	1	0	0	GBS	Lyme Disease	Brucellosis
GBS-28	11	F	Facial weakness, numbness	GBS	1	0	0	GBS	Lyme Disease	Acoustic Neuroma
GBS-29	26	F	Foot numbness, gait instability	Sensory ataxic GBS	1	0	0	GBS	Multiple Sclerosis	Tarsal Tunnel
GBS-30	26	F	Numbness from waist down	GBS	0	0	0	Transverse myelitis	Multiple Sclerosis	Herniated Disk
GBS-31	20	F	Bulbar weakness after viral illness	PCB variant GBS	1	0	0	GBS	Laryngitis	Spinal Muscular Atrophy
GBS-32	77	F	Bulbar weakness, areflexia	Recurrent PCB GBS	1	0	1	GBS	Multiple Sclerosis	GBS
GBS-33	13	F	Facial weakness, numbness	GBS variant	1	0	0	GBS	Migraine (children)	Trigeminal Neuropathy
GBS-34	61	F	Sensory disturbance, ptosis	AMSAN variant	0	0	0	Myasthenia gravis	Bell's Palsy	Stroke
GBS-35	38	M	Distal limb weakness after URI	Pure motor GBS	1	0	0	GBS	Distal limb weakness	Polyneuropathy
GBS-36	26	M	Hoarse voice, swallowing difficulty, paresthesia	Miller Fisher variant	1	0	0	GBS	Multiple Sclerosis	Polyneuropathy

1 = correctly identified GBS as top diagnosis | 0 = did not identify GBS

Appendix B — Full Raw Data Spreadsheet: Mimic Cases

Table B1: Non-GBS Mimic Cases — Full Raw Data

Sep 2026
Vol 11. No 3.

Case ID	Age	Sex	Early Symptoms	True Diagnosis	ML Correct	WebMD Correct	Isabel Correct	ML Top Dx	WebMD Top Dx	Isabel Top Dx
MIM-01	28	M	Progressive leg numbness, weakness	Nitrous oxide myeloneuropathy	0	1	1	GBS (wrong)	Multiple Sclerosis	Spinal Cord Compression
MIM-02	30	M	Progressive ascending weakness, areflexia	Nitrous oxide B12 deficiency	0	1	1	GBS (wrong)	Multiple Sclerosis	Vitamin E Deficiency
MIM-03	6	F	Progressive leg pain, inability to walk	Spinal epidural hematoma	1	1	1	Acute transverse myelitis	Anxiety Disorder	Polyneuropathy
MIM-04	9	F	Progressive paraplegia	Thoracic epidural hematoma	0	1	1	GBS (wrong)	Multiple Sclerosis	Tuberculous Meningitis
MIM-05	58	F	Progressive bilateral leg weakness	Thoracic myelopathy	0	1	1	GBS (wrong)	Multiple Sclerosis	Polyneuropathy
MIM-06	39	M	Bilateral foot paresthesia, leg weakness	Intrathecal MTX myelopathy	0	1	1	GBS (wrong)	Multiple Sclerosis	Spinal Cord Compression
MIM-07	26	F	Acute ascending paralysis, respiratory failure	Hereditary coproporphyria	0	1	1	GBS (wrong)	Multiple Sclerosis	Insect Sting Allergy
MIM-08	56	F	Paresthesia, encephalopathy	Thiamine deficiency/Wernicke	1	1	1	MS relapse	Multiple Sclerosis	Cervical Spondylosis
MIM-09	56	M	Acute peripheral neuropathy	Acute brucellosis	1	1	1	Peripheral neuropathy	Multiple Sclerosis	Myeloma
MIM-10	85	F	Worsening paresthesia after nivolumab	CIDP	0	1	1	GBS (wrong)	Multiple Sclerosis	Biotin Deficiency
MIM-11	29	M	Bilateral leg weakness after resp. infection	Acute lymphoblastic leukemia	0	1	1	GBS (wrong)	Multiple Sclerosis	Cauda Equina Syndrome
MIM-12	63	F	Acute flaccid weakness after diarrhea	Vasculitic neuropathy	0	1	1	GBS (wrong)	Multiple Sclerosis	Myelopathy
MIM-13	33	M	Progressive weakness, areflexia, oculomotor	Severe hyponatremia	1	1	1	Miller Fisher	Multiple Sclerosis	Vitamin B12 Deficiency
MIM-14	49	F	Rapidly progressing quadriparesis	Severe hypokalemia	0	1	0	GBS (wrong)	Multiple Sclerosis	GBS (wrong)
MIM-15	49	M	Progressive upper limb paralysis, facial palsy	Tick-borne encephalitis	0	1	1	GBS (wrong)	Multiple Sclerosis	Insect Sting Allergy
MIM-16	51	M	Progressive bilateral leg numbness, weakness	Neurosyphilis	0	1	0	GBS (wrong)	Multiple Sclerosis	GBS (wrong)
MIM-17	43	M	Progressive quadriparesis	Arsenic poisoning	0	1	1	GBS (wrong)	Multiple Sclerosis	Myeloma
MIM-18	26	F	Ascending paralysis, dysarthria, nystagmus	Thiamine deficiency/Wernicke	1	1	0	Miller Fisher	Multiple Sclerosis	GBS (wrong)
MIM-19	9	M	Progressive bilateral leg weakness	Neuroschistosomiasis	0	1	1	GBS (wrong)	Multiple Sclerosis	Polyneuropathy
MIM-20	8	M	Ascending symptoms, severe leg weakness	Spinal epidural hematoma/AVM	0	1	0	GBS (wrong)	Multiple Sclerosis	GBS (wrong)
MIM-21	46	M	Progressive upper and lower extremity weakness	Beriberi/thiamine deficiency	1	1	1	CIDP	Multiple Sclerosis	Viral Meningoencephalitis
MIM-23	58	M	Progressive neuropathy, facial palsy	Arsenic poisoning	0	1	0	GBS (wrong)	Multiple Sclerosis	GBS (wrong)
MIM-24	25	M	Progressive weakness, respiratory paralysis	Acute hepatic porphyria	0	1	1	GBS (wrong)	Multiple Sclerosis	Polyneuropathy
MIM-25	70	F	Bulbar palsy, autonomic disturbance	Strongyloidiasis	0	1	1	GBS (wrong)	Multiple Sclerosis	Diabetic Neuropathy
MIM-26	5	F	Anemia, encephalopathy, peripheral neuropathy	Lead poisoning	1	1	1	Subacute combined degeneration	Multiple Sclerosis	Heavy Metal Intoxication
MIM-27	38	M	GBS-like cauda equina symptoms	Neurosarcoidosis	1	1	1	Cauda equina syndrome	Multiple Sclerosis	Cauda Equina Syndrome
MIM-28	26	F	Subacute limb weakness in pregnancy	Dry beriberi/Wernicke	1	1	1	Nutritional neuropathy	Multiple Sclerosis	Alcoholic Myopathy
MIM-29	12	F	Recurrent numbness, limb weakness	HNPP	1	1	1	MS relapse	Multiple Sclerosis	Viral Meningoencephalitis
MIM-30	4	M	Fever, lower limb weakness, absent reflexes	Acute brucellosis	↓	1	1	Acute transverse myelitis	Multiple Sclerosis	Myelitis
MIM-31	49	M	Progressive ascending weakness, gait instability	Acute intermittent porphyria	0	1	0	GBS (wrong)	Multiple Sclerosis	GBS (wrong)

1 = correctly did NOT say GBS | 0 = incorrectly said GBS for a non-GBS case

Appendix C — Full Statistical Formulas

Sep 2026

Vol 11. No 3.

C1 — Sensitivity $\text{Sensitivity} = \text{TP} \div (\text{TP} + \text{FN})$

- Custom ML: $17 \div 30 = 0.567 = 56.7\%$
- WebMD: $0 \div 30 = 0.0\%$
- Isabel DDx: $2 \div 30 = 0.067 = 6.7\%$

C2 — Specificity $\text{Specificity} = \text{TN} \div (\text{TN} + \text{FP})$

- Custom ML: $11 \div 30 = 0.367 = 36.7\%$
- WebMD: $30 \div 30 = 1.000 = 100.0\%$
- Isabel DDx: $24 \div 30 = 0.800 = 80.0\%$

C3 — Accuracy $\text{Accuracy} = (\text{TP} + \text{TN}) \div (\text{TP} + \text{FN} + \text{TN} + \text{FP})$

- Custom ML: $28 \div 60 = 0.467 = 46.7\%$
- WebMD: $30 \div 60 = 0.500 = 50.0\%$
- Isabel DDx: $26 \div 60 = 0.433 = 43.3\%$

C4 — Positive Predictive Value (PPV) $\text{PPV} = \text{TP} \div (\text{TP} + \text{FP})$

- Custom ML: $17 \div 36 = 0.472 = 47.2\%$
- WebMD: Undefined (no positive predictions made) = reported as 0.0%
- Isabel DDx: $2 \div 8 = 0.250 = 25.0\%$

C5 — Negative Predictive Value (NPV) $\text{NPV} = \text{TN} \div (\text{TN} + \text{FN})$

- Custom ML: $11 \div 24 = 0.458 = 45.8\%$
- WebMD: $30 \div 60 = 0.500 = 50.0\%$
- Isabel DDx: $24 \div 52 = 0.462 = 46.2\%$

C6 — F1 Score $\text{F1} = (2 \times \text{TP}) \div (2 \times \text{TP} + \text{FP} + \text{FN})$

- Custom ML: $34 \div 66 = 0.515$
- WebMD: $0 \div 30 = 0.000$
- Isabel DDx: $4 \div 38 = 0.105$

C7 — False Negative Rate $\text{FNR} = \text{FN} \div (\text{FN} + \text{TP})$

- Custom ML: $13 \div 30 = 0.433 = 43.3\%$
- WebMD: $30 \div 30 = 1.000 = 100.0\%$
- Isabel DDx: $28 \div 30 = 0.933 = 93.3\%$

C8 — False Positive Rate $\text{FPR} = \text{FP} \div (\text{FP} + \text{TN})$

- Custom ML: $19 \div 30 = 0.633 = 63.3\%$
- WebMD: $0 \div 30 = 0.000 = 0.0\%$
- Isabel DDx: $6 \div 30 = 0.200 = 20.0\%$

C9 — AUC (Balanced Accuracy Estimate) $AUC = 0.5 \times (\text{Sensitivity} + \text{Specificity})$

- Custom ML: $0.5 \times (0.567 + 0.367) = 0.467$
- WebMD: $0.5 \times (0.000 + 1.000) = 0.500$
- Isabel DDx: $0.5 \times (0.067 + 0.800) = 0.433$

C10 — One-Sample Proportion Z-Test $z = (\hat{p} - p_0) \div \sqrt{(p_0 \times (1 - p_0) \div n)}$ Standard Error = $\sqrt{(0.49 \times 0.51 \div 30)} = \sqrt{0.00833} = 0.0913$

- Custom ML: $z = (0.567 - 0.49) \div 0.0913 = 0.840 \rightarrow p = 0.401$
- WebMD: $z = (0.000 - 0.49) \div 0.0913 = -5.369 \rightarrow p < 0.0001$
- Isabel DDx: $z = (0.067 - 0.49) \div 0.0913 = -4.638 \rightarrow p < 0.0001$

C11 — Fisher's Exact Test Contingency Tables

Custom ML vs. WebMD on GBS sensitivity: $[[17, 13], [0, 30]] \rightarrow p < 0.000001$

Custom ML vs. Isabel DDx on GBS sensitivity: $[[17, 13], [2, 28]] \rightarrow p = 0.000054$

WebMD vs. Isabel DDx on GBS sensitivity: $[[0, 30], [2, 28]] \rightarrow p = 0.491525$

C12 — Subgroup Sensitivity Subgroup Sensitivity = correct detections in subgroup $\div$ total cases in subgroup

- Custom ML classic (n=6): $2 \div 6 = 33.3\%$
- Custom ML atypical (n=24): $15 \div 24 = 62.5\%$
- Isabel DDx classic: $0 \div 6 = 0.0\%$
- Isabel DDx atypical: $2 \div 24 = 8.3\%$
- WebMD classic and atypical: 0.0% both

Appendix D — Sample Vignette Format

This is an example of exactly how each clinical vignette was written before being entered into each AI tool. This shows the standardized format used for all 60 cases.

Sample Vignette — GBS-36

Patient: 26-year-old male. Early symptoms: Hoarse voice, difficulty swallowing, paresthesia in hands and feet following a recent diarrheal illness. No lab results, no imaging, no confirmatory testing included

The vignette was entered into each AI tool as exactly shown above. The tools' top diagnosis and the position it was ranked in GBS in the output were recorded.

Appendix E — Confusion Matrix Summary Table

Table E1: Confusion Matrix for All Three Tools (n=60)

	Custom ML	WebMD	Isabel DDx
True Positives (TP)	17	0	2
False Negatives (FN)	13	30	28
True Negatives (TN)	11	30	24
False Positives (FP)	19	0	6
Total Cases	60	60	60