JCTA Journal of Computing Theories and Applications ISSN:3024-9104

Research Article

A Training-Free Symbolic Baseline for Interactive Medical Diagnosis: Automatic Rule-Slot Derivation and Weighted Evidence Combination

Naufal Denta Elsatya Nugraha [1], Dhendra Marutho [2,*], Muhammad Sam'an [2], Wendy Sarasjati [2], and De Rosal Ignatius Moses Setiadi [3]

[1] Undergraduate Program, Department of Informatics, Universitas Muhammadiyah Semarang, Semarang 50273, Indonesia; e-mail : naufaldenta.e.n@gmail.com
[2] Department of Informatics, Universitas Muhammadiyah Semarang, Semarang 50273, Indonesia; e-mail : dhendra@unimus.ac.id; muhammad92sam@unimus.ac.id; wendysar@unimus.ac.id
[3] Research Group for Quantum Computing and Materials Informatics, Faculty of Intelligent Technology, Universitas Dian Nuswantoro, Semarang 50131, Indonesia; e-mail : moses@dsn.dinus.ac.id
* Corresponding Author : Dhendra Marutho

Abstract: Interactive medical diagnosis requires sequential symptom elicitation before a diagnosis can be established. Although recent approaches employ reinforcement learning and transformer-based agents, their performance has not been adequately contextualized against a classical probabilistic baseline. This study introduces a training-free symbolic baseline that derives conditional symptom probabilities and three rule slots—must-have, exclusion, and critical—from co-occurrence counts in simulated patient records. Candidate diseases are ranked using observed evidence, while subsequent questions are selected by expected information gain. Evaluation on five benchmarks from a published reinforcement learning study reveals distinct performance regimes. On MedlinePlus (893 diseases), the proposed method achieves 99.27 ± 0.09% accuracy with an average of 1.75 questions, outperforming the published baseline (96.2 ± 0.1%; 3.41 questions). Across four SymCat variants, it improves accuracy by 3.03–7.70 percentage points but requires approximately twice as many questions; at a matched budget of 8.98 questions, the reference method achieves higher accuracy, highlighting a trade-off between diagnostic accuracy and interaction cost. Ablation experiments across six datasets show that removing weighted evidence combination reduces accuracy by 35.4–65.4 percentage points, while eliminating symptom inquiry reduces it by 23.3–66.8 percentage points. The safety veto removes up to 50.2% of candidates on average but changes the final diagnosis in at most 0.17% of consultations, indicating limited incremental impact on predictive accuracy under the evaluated configurations. Rule-base derivation takes 1.8 s for 893 diseases on a CPU. As all benchmarks rely on simulated patients, these findings establish benchmark-level performance rather than clinical effectiveness.

Keywords: Expected information gain; Explainable clinical decision support; Medical automatic diagnosis; Naive Bayes; Reinforcement learning; Rule-based expert system; Symptom inquiry; Training-free diagnosis.

Received: July, 30th 2026
Revised: October, 1st 2026
Accepted: October, 10th 2026
Published: October, 25th 2026

Copyright: © 2026 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) licenses (https://creativecommons.org/licenses/by/4.0/)

1. Introduction

Medical automatic diagnosis treats the clinical interview as a sequential decision problem. A system receives an initial complaint, asks about further symptoms one at a time, and stops when it is confident enough to name a disease [1]. Two costs are traded against each other. Symptoms collected beyond the initial complaint raise diagnostic accuracy [2], but long interviews burden the patient and, when examinations are involved, raise the cost of care. The task therefore has two natural metrics: the accuracy of the final diagnosis and the number of questions needed to reach it.

Research on this problem has converged on reinforcement learning (RL). Early formulations merged asking and diagnosing into one policy [3]–[7], later designs separated the two

decisions [8], and Yuan and Yu [9] aligned them through an adaptive stopping criterion, reporting the strongest published results on five benchmarks. Transformer formulations followed a similar path [10], [11].

Four limitations of this line of work motivate the present study. First, the inquiry policy resides in network weights, so the system cannot state why it asked about fever rather than cough, although explainability is regarded as a precondition for medical artificial intelligence [12], [13] and the evaluation of explanations has itself become the subject of systematic review [14]. Second, training consumes large numbers of simulated episodes, which is impractical in domains where no patient simulator exists and rule bases are still written by hand, as they are for maternal care in low-resource settings [15]. Third, no mechanism prevents a clinically impossible diagnosis from receiving a high probability, because a softmax over learned logits cannot express a hard constraint; hybrid designs therefore keep such knowledge outside the network, in symbolic rules or knowledge graphs [16], [17]. Fourth, the reported gains have not been measured against a classical alternative, so it is unknown how much of the margin belongs to the learned policy and how much to the probabilistic formulation that any reasonable baseline would share. The size of that margin matters in practice, because online symptom checkers name the correct primary diagnosis in only 19% to 37.9% of cases [18], and systematic reviews warn that relying on them could compromise patient safety and call for validation before wide adoption [18], [19].

Sequential diagnosis by probabilistic inference is decades older than this literature. The Bayesian formulation, question selection by an information measure, and large probabilistic knowledge bases were established between 1956 and 1992 [20]–[27], as Section 2.2 reviews. None of these methods appears as a baseline in recent comparative studies of symptom inquiry. This omission matters because carefully specified baselines have repeatedly closed gaps that had been attributed to architectural progress. Reported advances in metric learning and in neural recommendation shrank or disappeared once simple methods received equal tuning effort [28], [29], and reported gains in deep RL proved hard to separate from variance and implementation detail [30]. This study applies the same scrutiny to symptom inquiry.

This study constructs a training-free symbolic system from established components and evaluates it with the benchmarks, protocol, and reporting conventions of Yuan and Yu [9], whose study and method are referred to below as the reference study and the reference method. Table 1 distinguishes established components from the contributions of this study, which are fourfold: (i) a procedure that derives must-have, exclusion, and critical rule slots from corpus counts rather than expert elicitation; (ii) a weighted evidence-combination method that scores diseases using observed evidence, replacing the match-ratio scoring of conventional expert systems; (iii) an evaluation of the assembled baseline on six datasets using a simulation protocol verified against the reference implementation, with remaining differences explicitly documented, including ablations of every component and rule slot; and (iv) a matched-budget analysis of the trade-off between diagnostic accuracy and interaction cost relative to the reference method. The central claim is not that these components are novel individually, but that their combination has been missing from comparative evaluations and that including it changes how the benchmark results should be interpreted.

The remainder of this paper is organized as follows. Section 2 reviews symptom-inquiry methods and the classical probabilistic tradition and explains the choice of components. Section 3 describes the data flow, the rule derivation, the scoring function, the safety veto, the question selector, the consultation loop, and the evaluation protocol. Section 4 reports the comparison with published methods, the ablation studies, and the trade-off between accuracy and interaction cost, and then discusses the findings and their limitations. Section 5 concludes.

2. Related Work

2.1. Sequential Symptom Inquiry

Early systems merged asking and diagnosing into a single action space, which forced the agent to search a larger space than the task requires [3]–[6]. Xia et al. [7] kept a single policy but regularized it with a generative adversarial objective and rewarded the most discriminative symptoms through a mutual-information term. Bipartite designs separated the two decisions: Lin et al. [8] added an introspective module that intervenes on candidate responses and stops once the diagnosis becomes stable, and He et al. [31] replaced the agent altogether with a

bipartite framework that selects symptoms by an information-theoretic reward and diagnoses with a variational encoder.

Yuan and Yu [9] made the coupling adaptive. Their agent is trained with a composite reward and their classifier with a cross-entropy loss, and the transition from asking to diagnosing is governed by a per-disease entropy threshold that is updated whenever a diagnosis proves correct. They also released a benchmark built from the MedlinePlus knowledge base, which contains 893 diseases, is substantially larger than the SymCat subsets used previously, and includes medical examinations.

Two properties make [9] the primary comparator. It reports the strongest published accuracy on all five benchmarks, so exceeding it implies exceeding the other methods on the same data, and it released the knowledge bases, the patient generator, and the training code, which allows the simulation protocol to be verified rather than assumed. The three methods it compares against, REFUEL [3], INS-DS [8], and FIT [31], published later as BSODA, are included in the present comparison with their published figures. REFUEL and INS-DS were re-implemented by [9], whereas the FIT figures are quoted there from its own paper.

Transformer variants replace the agent with sequence models. DxFormer [10] decouples inquiry from diagnosis in a decoder-encoder design with dense symptom representations, and Zhang et al. [11] inject records of physician diagnostic behavior into the model. Recent benchmarks of questioning efficiency in large language models [32] show that the number of questions required for a confident diagnosis remains an open problem, and sequential diagnosis agents built on language models are now evaluated on the same trade-off between accuracy and interaction cost [33].

2.2. Probabilistic and Rule-Based Diagnosis

The probabilistic treatment of diagnosis is older than the neural one. Ledley and Lusted [20] expressed diagnosis as inference over disease probabilities given the observed symptoms. Gorry and Barnett [21] added the sequential element, choosing each next question by a decision-theoretic criterion instead of a fixed order. The value of an observation had been formalized earlier by Lindley [26], as the expected reduction of entropy, and by Howard [27]; that measure is the criterion adopted in this paper. De Dombal et al. [22] applied the conditional-independence approximation to acute abdominal pain and reported an accuracy exceeding that of the attending clinicians.

Knowledge-based systems followed a different route. INTERNIST-1 [23] encoded manually curated disease profiles for internal medicine, QMR-DT [24] recast that knowledge base in probabilistic form, and Pathfinder [25] combined expert elicitation with normative decision theory. The recurring obstacle was knowledge acquisition, because every disease profile consumed expert time. Contemporary rule-based clinical systems face the same constraint, including systems developed for maternal complications in low-resource settings [15].

Neuro-symbolic designs have renewed interest in symbolic reasoning where auditability is required. Prenosil et al. [34] linked a large language model to a rule-based verifier to produce traceable labels from clinical reports, and a recent review positions this combination as a viable direction for medicine [35]. Knowledge graphs serve a similar purpose, either as the substrate of interpretable hybrid reasoning [16] or as a factual constraint on a generative model [17], and large medical knowledge graphs are now assembled from several sources to supply such knowledge [36]. A related strand distills a trained network into a rule set whose decisions can be read [37]. These systems perform passive information extraction, answer generation, or post hoc explanation rather than interactive diagnosis.

2.3. Rationale for the Choice of Components

Each of the three components answers a specific requirement of the task, and each alternative that was considered fails a requirement that matters here.

Scoring has to produce a calibrated distribution over diseases rather than a ranking, because the question selector uses that distribution as its prior. A naive Bayes log-likelihood provides one in closed form from the conditional probabilities that the derivation step already estimates, and each term remains interpretable as a prevalence. A Bayesian network with learned structure would relax the independence assumption, but it requires structure learning from data that these knowledge bases do not provide and would forfeit the property that motivates the baseline, namely that no parameter is fitted. Discriminative classifiers such as

gradient-boosted trees perform well on complete feature vectors, but they degrade under the pattern of missing values that dominates the first turns of an interview and provide no per-symptom accounting.

Question selection has to be computable at run time without a policy network and has to justify each question in a unit that a clinician can read. Expected information gain (IG) satisfies both requirements: it follows from the same conditional probabilities, costs one pass over the candidate pool, and is measured in bits. A decision tree induced offline would also avoid training at run time, but it fixes the question order in advance and cannot react to an unexpected answer. A learned policy is what the reference method already provides, so Section 4.3 compares IG with a heuristic discriminative selector instead.

Rule slots express what a probabilistic score cannot. A posterior can be made arbitrarily small but not zero, so a purely probabilistic system cannot exclude a diagnosis that contradicts the evidence. The slots provide that exclusion, and deriving them from counts rather than from clinicians makes them usable at the scale of 893 diseases, where elicitation is impractical.

2.4. Position of the Present Study

The review identifies two gaps: symptom-inquiry studies report no classical sequential diagnosis baseline, and the rule-based tradition still depends on manual knowledge acquisition. This study addresses both by deriving rule slots from data and by evaluating the resulting system against published RL results under a verified protocol. Table 1 lists the elements adopted from existing work and those contributed here.

Table 1. Established components and contributions of this study

Element	Status	Source or section
Log-likelihood combination of observed evidence	Established	Naive Bayes under conditional independence [20], [22]
Expected information gain as the question criterion	Established	Value of information [26], [27]; sequential diagnosis [21]
Hard constraints and safety filtering	Established	Expert systems [23], [25]
Conditional probabilities estimated from a simulated patient corpus	Contributed	Section 3.2
Rule slots derived from corpus counts with a specificity requirement	Contributed	Section 3.3; ablation in Section 4.4
Scoring from observed evidence only, replacing the match ratio	Contributed	Section 3.4
Evaluation of the assembled system on six datasets under a verified protocol	Contributed	Sections 3.8, 4.1, and 4.2
Component and rule-slot ablations	Contributed	Sections 4.3 to 4.5
Matched-budget analysis of accuracy and interaction cost	Contributed	Section 4.6

3. Proposed Method

3.1. Overview and Terminology

The proposed system consists of two stages: an offline stage for constructing the symbolic rule base and an online stage for interactive diagnosis. In the offline stage, a single pass over a patient-record corpus is used to estimate conditional probabilities for each disease–symptom pair and derive three rule slots representing necessity, contradiction, and characteristic association. In the online stage, the system iteratively scores candidate diseases based on the evidence collected, applies a safety veto, evaluates the stopping criterion, and selects the next question when further inquiry is required. The overall workflow and the interactions between these components are illustrated in Figure 1. Since the benchmarks used in this study provide knowledge bases rather than actual patient records, it is important to distinguish the original knowledge sources from the simulated patient records generated from them. To ensure consistent interpretation of the experimental setup and findings, the key terms used throughout this paper are defined in Table 2.

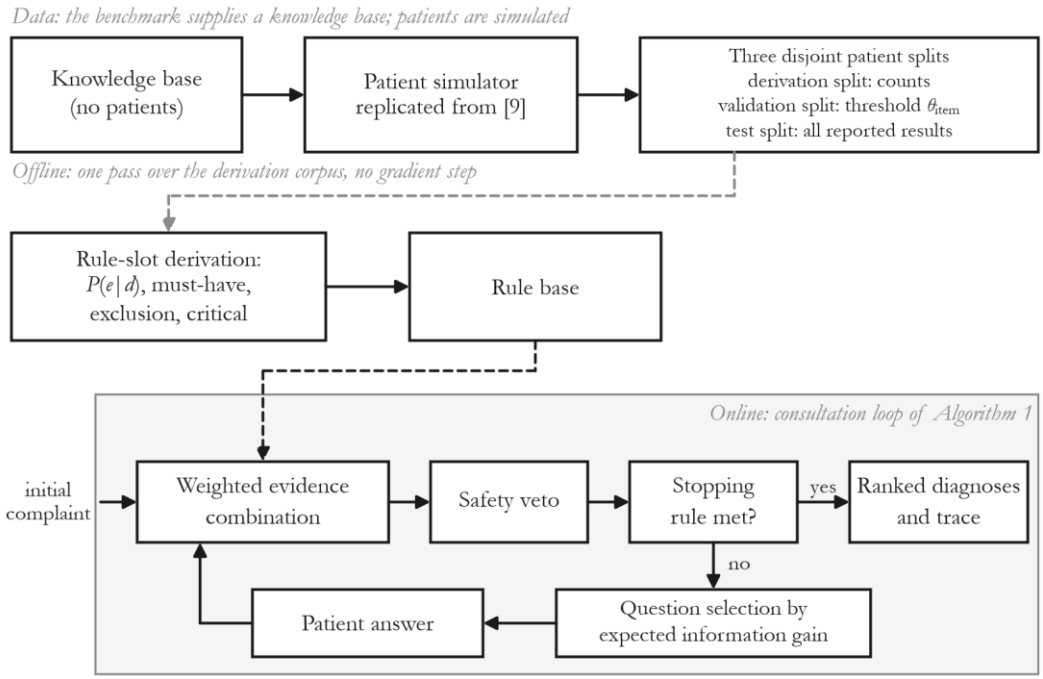


Figure 1. Data flow and components of the proposed system

Table 2. Terminology used in this paper

Term	Meaning in this paper
Knowledge base	The released mapping from each disease to its symptoms and examinations, with symptom prevalences on SymCat. It contains no patient records.
Patient simulator	The procedure that draws a disease and a set of symptoms from the knowledge base, replicated from the reference implementation (Section 3.8).
Simulated patient	One record produced by the simulator: a true disease, an initial complaint, and the symptoms that questions can still reveal.
Derivation, validation, and test splits	Three disjoint sets of patients, used for counting, for choosing the rule-item threshold θitem (Section 3.3), and for reporting results, respectively.
Derivation corpus	The patients of the derivation split are scanned once to obtain the counts of Section 3.3. On DDXPlus, this is the released training file.
Rule base	The output of rule derivation: P(e|d) for every disease-symptom pair and the must-have, exclusion, and critical slots.

3.2. From Knowledge Base to Derivation Corpus

The conditional probabilities of Section 3.3 are estimated from patient records rather than read from the knowledge base. MedlinePlus and SymCat provide no patient records, so the simulator described in Section 3.8 generates them, and the counts $n(d)$, $n(e)$, and $n(e,d)$ are accumulated over the simulated patients. DDXPlus provides released patient records, which are used directly.

Estimating from simulated patients rather than from the knowledge base is deliberate. A knowledge base lists the symptoms of a disease but not how often each appears in a simulated patient. On SymCat, the two differ because every symptom is drawn independently at its own prevalence; on MedlinePlus, they differ because the simulator draws a Poisson-distributed subset of symptoms but attaches every examination deterministically, so examinations reach $P(e|d)$ close to one, whereas individual symptoms do not. Counting over simulated patients recovers these frequencies, which makes the estimator consistent with the generative process of each benchmark. Section 4.7 discusses what this implies for the interpretation of the results.

The three splits (Table 2) are drawn with different random seeds and do not overlap, and no test-split result influenced any configuration choice. The splits share the knowledge

base from which they are generated. This is unavoidable for knowledge-base benchmarks and applies equally to the reference method, whose agent is trained on patients drawn from the same knowledge base. On DDXPlus, the splits are the released training, validation, and test files, which are disjoint by construction.

3.3. Automatic Rule-Slot Derivation

Rule-based diagnostic systems require an expert to state, for every disease, which symptoms must be present, which symptoms exclude the disease, and which are strongly characteristic of it. Obtaining these statements is the knowledge acquisition bottleneck, which this step replaces with estimation from counts. Let N be the number of patients in the derivation corpus, $n(d)$ the number with disease d, $n(e)$ the number with symptom e, and $n(e,d)$ the number with both. Then

$$P(e|d) = \frac{n(e,d)}{n(d)}, \qquad P(e) = \frac{n(e)}{N} \tag{1}$$

$$P(e|\neg d) = \frac{[n(e) - n(e,d)]}{[N - n(d)]} \tag{2}$$

where $P(e)$ is the prevalence of e in the corpus. Two derived measures separate frequency from specificity: the lift expresses how much more often a symptom accompanies a disease than it occurs in general, and the discriminative margin expresses how strongly the symptom distinguishes the disease from all others,

$$\text{lift}(e,d) = \frac{P(e|d)}{P(e)}, \qquad \text{disc}(e,d) = P(e|d) - P(e|\neg d) \tag{3}$$

The must-have slot $\text{must}(d)$, the exclusion slot $\text{excl}(d)$, the critical slot $\text{crit}(d)$, and the set of rule $\text{items}(d)$ are thresholded sets:

$$\text{must}(d) = \{e : P(e|d) \geq \theta_{must} \wedge disc(e,d) \geq \delta_{must}\} \tag{4}$$

$$\text{excl}(d) = \{e : P(e|d) \leq \theta_{\text{excl}} \wedge P(e) \geq \pi_{\min} \wedge disc(e,d) \leq -\delta_{\text{excl}}\} \tag{5}$$

$$\text{crit}(d) = \{e : \text{lift}(e,d) \geq \lambda \wedge P(e|d) \geq \theta_{\text{crit}}\} \tag{6}$$

$$\text{items}(d) = \{e : P(e|d) \geq \theta_{\text{item}}\} \tag{7}$$

Table 3. Parameters of the proposed system

Symbol	Value	Role
θ_{must}, δ_{must}	0.95, 0.10	Must-have slot, Equation (4); at most 4 symptoms, largest disc first
θ_{excl}, $\pi_{\min}$, δ_{excl}	0.02, 0.05, 0.10	Exclusion slot, Equation (5); at most 8 symptoms, most negative disc first
λ, θ_{crit}	3.0, 0.10	Critical slot, Equation (6); at most 8 symptoms, largest lift first
θ_{item}	0.10 (MedlinePlus, DDXPlus); 0.005 (SymCat)	Rule items, Equation (7)
β	0.01	Weight of a symptom outside items(d), Equation (8)
ε	10^{-6}	Lower bound inside every logarithm, Equations (9) and (12)
$\|C\|$	50	Size of the candidate pool, Section 3.6
B	30	Question budget, Section 3.7

A symptom enters $\text{must}(d)$ when it is almost always present under d and also specific to d, enters $\text{excl}(d)$ when it is common in general but almost never accompanies d, and enters $\text{crit}(d)$ when it is far more frequent under d than overall. The specificity condition in Equation (4) is essential. Without it, frequent but uninformative symptoms satisfy the must-have threshold for many diseases at once: on DDXPlus, removing the condition made an uninformative context feature a must-have symptom for 37 of the 49 pathologies, which

raised the scores of all candidates together and removed the margin between them. Equations (4) to (6) define the sets before truncation. Each slot then keeps only its most extreme members, and a symptom that qualifies for both $\mathrm{must}(d)$ and $\mathrm{crit}(d)$ is kept only in $\mathrm{must}(d)$, so that no evidence is counted twice. Table 3 lists the numerical parameters of the system; the stopping thresholds, which are fixed constants of the consultation loop, are given in Section 3.7.

3.4. Weighted Evidence Combination

Conventional rule-based scoring computes the fraction of the rule items of a disease that the patient has confirmed. This formula treats a symptom that has not yet been asked about exactly like a symptom that the patient has denied. Because almost every symptom is unknown at the start of an interview, the ranking is then driven by the length of each rule list rather than by the evidence. The proposed score uses observed evidence only. Let E^+ and E^- denote the sets of confirmed and denied symptoms. The evidence weight $w(e,d)$ of symptom e for disease d is

$$w(e,d) = P(e|d) \text{ for } e \in \mathrm{items}(d), \qquad w(e,d) = \beta \text{ for } e \notin \mathrm{items}(d) \tag{8}$$

and the score of d is an approximate log-likelihood of the observed evidence under conditional independence, in which denied symptoms outside $\mathrm{items}(d)$ are ignored,

$$\mathrm{score}(d) = \sum_{e \in E^+} \log \max[w(e,d), \varepsilon] + \sum_{e \in E^- \cap items(d)} \log \max[1 - P(e|d), \varepsilon] \tag{9}$$

where log is the natural logarithm, and ε keeps every logarithm finite. The two sums treat a symptom outside $\mathrm{items}(d)$ differently. A confirmed symptom that d does not list receives the weight β, so a disease that cannot account for a reported symptom is penalized without being eliminated. A denied symptom that d does not list contributes nothing, because a disease that never produces the symptom gains nothing from its absence.

Equation (9) is an approximate log-likelihood of a naive Bayes model and plays the role that certainty factors play in classical expert systems, with weights estimated from data rather than assigned by experts. Each term remains interpretable: a contribution of log 0.95 states that the symptom occurs in 95% of the cases of that disease. Because each question adds only a few observations, the scores are updated incrementally, which reduces the cost of a turn from the product of the number of rules and observations to the number of rules.

3.5. Safety Veto and Posterior

A candidate is vetoed when the patient denies one of its must-have symptoms or confirms two of its exclusion symptoms. With D the set of all diseases, the set of vetoed candidates is

$$V = \{d \in D : \mathrm{must}(d) \cap E^- \neq \emptyset \vee |\mathrm{excl}(d) \cap E^+| \geq 2\} \tag{10}$$

Scores are converted into a posterior distribution by a softmax with unit temperature over the candidates that are not vetoed,

$$p(d) = \frac{\exp[\mathrm{score}(d)]}{\Sigma d'} \in D \backslash V \ \exp[\mathrm{score}(d')] \text{for } d \notin V, \qquad p(d) = 0 \text{ for } d \in V \tag{11}$$

which, for the approximate log-likelihoods of Equation (9), is up to that approximation, the Bayes posterior under a uniform prior over the remaining diseases. A vetoed candidate is removed from the normalization rather than given a small weight, so no amount of evidence can restore it. Applying the veto before normalization therefore makes the guarantee constructive rather than probabilistic. If every candidate were vetoed, the system would fall back to a uniform distribution over D and still return a ranking.

3.6. Question Selection by Expected Information Gain

The next question is the one whose answer is expected to reduce the uncertainty about the diagnosis the most. The computation is restricted to the candidate pool C of the $|C|$ diseases with the highest posterior, renormalized as

$$p_i = \frac{max[p(d_i), \varepsilon]}{\Sigma_{j \in C}\, max[p(d_j), \varepsilon]}, \qquad H(p) = -\Sigma_{i \in C} p_i \log_2 p_i \tag{12}$$

where p_i is the renormalized posterior of the i-th candidate, i and j index the diseases in C, $H(p)$ is the entropy of the pool distribution in bits, and 0 log2 0 is taken as 0, which matters for $H(p^+)$ and $H(p^-)$ in Equation (15), where an entry can be exactly zero. For a candidate symptom e, the predicted probability $q(e)$ of a positive answer and the pool distributions p^+ and p^- after a positive and a negative answer are

$$q(e) = \Sigma_{i \in C}\, p_i\, w(e, d_i) \tag{13}$$

$$p_i^+ = p_i \frac{w(e, d_i)}{q(e)}, \quad p_i^- = p_i[1 - w(e, d_i)]/[1 - q(e)] \tag{14}$$

and the expected information gain of asking about e is

$$\mathrm{IG}(e) = H(p) - q(e)H(p^+) - [1 - q(e)]H(p^-) \tag{15}$$

The selector uses the evidence weights of Equation (8) for both answers. For a denied symptom outside items(d), it therefore assigns the factor $1 - \beta = 0.99$, whereas Equation (9) assigns 1, a difference of at most one percent per symptom. Because the gain is computed rather than learned, every question carries a justification in bits, and the ranked alternatives can be inspected.

3.7. Consultation Loop

Algorithm 1 combines Sections 3.4 to 3.6 into the online consultation loop, and the remainder of this subsection specifies the details that the equations leave open.

Algorithm 1. Online consultation loop

INPUT: rule base (Section 3.3), initial evidence E_0, question budget B, pool size C
OUTPUT: candidate diseases ranked by p(d), with the questions asked and their gains

1: Set $\mathrm{E}^+ \leftarrow \mathrm{E}_0, \mathrm{E} \leftarrow \emptyset$, and $t \leftarrow 0$.
2: Compute score (d) for every $d \in D$ by Equation (9), the vetoed set V by Equation (10), and $p(d)$ by Equation (11).
3: If the stopping rule holds, go to Step 8.
4: Form the pool C of the $|C|$ candidates with the highest $p(d)$ and the set U of candidate symptoms.
5: Compute $IG(e)$ for every $e \in U$ by Equations (12) to (15).
6: If no answerable $e \in U$ has a positive gain, go to Step 8; otherwise, let e^* be the answerable symptom with the largest gain.
7: Ask the question for e^*, add each symptom that it covers to E^+ or E^- according to the answer, set $t \leftarrow t + 1$, and return to Step 2.
8: Return the candidate diseases ranked by $p(d)$.

Initial evidence. The loop starts from the initial evidence E_0 of the simulated patient. On MedlinePlus, E_0 is the initial complaint, a Poisson (2) +1 prefix of the sampled symptoms; on SymCat, E_0 consists of the first sampled symptom, which is the initial complaint, together with age and sex; on DDXPlus, it is the released initial evidence. Examinations are never part of E_0. On DDXPlus, one question can cover several atomic features of the same evidence code, and a confirmed value also confirms its parent evidence.

Candidate symptoms. The set U contains every symptom that appears in $\mathrm{items}(d)$ for some $d \in C$, except symptoms whose status is already known, symptoms whose evidence code has already been asked, and symptoms that appear in $\mathrm{items}(d)$ for every $d \in C$, which separate nothing. A symptom is answerable when the question bank contains a question for it; the demographic context features of SymCat are known from the start and have no question. The pool size $|C| = 50$ was fixed for all datasets; enlarging it to 200 during development improved accuracy by about one point at 5.7 times the running time

Ties. Gains are expressed as a percentage of $H(p)$ and rounded to two decimals before sorting, so exact ties occur, and symptoms whose rounded gain is zero are discarded. The sort is stable, and the candidates are enumerated in descending posterior with the rule items of each candidate in descending $P(e|d)$, so a tie resolve to the symptom that appears first under the highest-ranked candidate. The selection is therefore deterministic and independent of the random seed.

Stopping rule. Let p_1 and p_2 be the two largest posteriors, $m = p_1 - p_2$ their margin, and t the number of questions asked. The interview stops when $t \geq B$. Otherwise, provided that no must-have symptom of the leading candidate has been denied, it stops before the first question if $p_1 \geq 0.92$ and m ≥ 0.25, and after at least one question if $p_1 \geq 0.80$ and m ≥ 0.18 or if $p_1 \geq 0.88$ and m ≥ 0.15; after at least three questions, $p_1 \geq 0.70$ and $m \geq 0.12$ also suffice. With the veto enabled, the leading candidate never has a denied must-have symptom, so the proviso matters only when the veto is disabled. These thresholds are fixed constants, shared by all datasets and not tuned. Because the posteriors sum to one, $m \geq 2p_1 - 1$, so every margin condition above is implied by the corresponding condition on p_1, and the second alternative after one question is implied by the first; the rule therefore reduces to $p_1 \geq 0.92$ before the first question, $p_1 \geq 0.80$ after at least one question, and $p_1 \geq 0.70$ after at least three.

3.8. Datasets and Protocol Verification

The evaluation uses the five benchmarks of the reference study and DDXPlus [38] as an external benchmark. MedlinePlus and SymCat provide knowledge bases, so their patients are simulated. Because any deviation in the simulator would invalidate a comparison with published figures, the official pickle files were opened, and the generation procedure was replicated from the released source code rather than inferred from the paper. Table 4 summarizes the derivation corpora, and Table 5 lists every aspect of the protocol that was checked.

Table 4. Derivation corpora and rule-derivation time on a CPU

Dataset	Diseases	Distinct features	Patients scanned	θitem	Derivation time
MedlinePlus	893	2,112	100,000	0.10	1.8 s
SymCat 200	200	332	150,000	0.005	2.0 s
SymCat 300	300	356	150,000	0.005	3.2 s
SymCat 400	400	370	150,000	0.005	2.2 s
SymCat Common	210	332	150,000	0.005	1.8 s
DDXPlus	49	515	1,025,602	0.10	89.1 s

The verification in Table 5 corrected two errors in the first replication: examinations had to be attached to every MedlinePlus patient deterministically rather than sampled, and the initial complaint had to contain a Poisson-distributed number of symptoms rather than one. Before these corrections, the two protocols were not equivalent, and a comparison drawn from them would not have been meaningful. Such protocol drift is difficult for a reader to detect, because two systems evaluated on nominally the same benchmark may face different tasks. Two aspects of the present protocol differ from the reference implementation and are listed as such.

3.9. Evaluation Protocol and Statistical Treatment

The reporting protocol follows the reference study. Every configuration is run in three replicates of 1,000 test patients, and the mean and its standard error (SE) are reported. Replicate r draws its test patients with the split seed offset by r and sets the random seed to 20260717 + r. On MedlinePlus and SymCat, this yields three different patient samples. On DDXPlus, the test records are fixed, so replicates differ only in the random-question ablation, and the SE of every deterministic configuration is exactly zero. The primary metrics are top-1 accuracy and the mean number of questions, denoted Q. All results in Section 4 use this protocol with a budget of $B = 30$, except the budget and threshold analyses in Section 4.6.

Table 5. Protocol aspects verified against the reference implementation

Aspect	Finding in the reference implementation	Status
Disease sampling	Called without a prevalence argument, so it falls back to a uniform distribution	Replicated
MedlinePlus scale	893 diseases, 2,112 features, 8.31 symptoms and 5.30 examinations per disease	Matches
Examinations	All examinations of the true disease are included deterministically	Replicated
Initial complaint, MedlinePlus	The first Poisson(2)+1 sampled symptoms	Replicated
Initial complaint, SymCat	The first sampled symptom only	Replicated
SymCat symptoms	Each symptom is present independently at its own prevalence	Replicated
SymCat context	Age and sex are known from the outset	Replicated
Feature identity	Each item is truncated to its first ten characters, which form the concept code of the Unified Medical Language System	Replicated
Stopping criterion	Entropy threshold learned per disease during training	Differs
Question budget	No explicit cap; this study applies a cap of 30	Differs

No model parameters are learned through gradient descent, and the system does not train a policy, classifier, or embedding. The offline stage consists of a single pass over the derivation corpus to accumulate integer counts, while the online stage evaluates the closed-form expressions in Sections 3.4–3.6. The only data-dependent configuration choice is the threshold θ_{item} which was selected using the validation split (0.10 for MedlinePlus and DDXPlus, and 0.005 for SymCat). Section 4.6 examines the sensitivity of the results to this threshold. This selection constitutes hyperparameter tuning on held-out data rather than model training and is the only stage that requires evaluating alternative configurations rather than performing a single pass over the corpus.

Differences against published figures are summarized by

$$z = \frac{\Delta}{\sqrt{(\mathrm{SE12} + \mathrm{SE22})}} \tag{16}$$

where Δ is the difference between the two mean accuracies and SE1 and SE2 are their standard errors; $|z| > 1.96$ is read as a difference at the 95% level. The statistic compares two independently reported means from different implementations evaluated on separately drawn patient samples. It is not a paired test, because a head-to-head run on identical patients would require retraining the reference agent, which would no longer be a comparison with published figures. Each SE is estimated from three replicates and is therefore imprecise, and the statistic assumes approximately normal sampling distributions. The value of z therefore indicates that a difference is large relative to the run-to-run variability of both systems; it is not a hypothesis test with a calibrated error rate.

4. Results and Discussion

4.1. Comparison with Published Methods

Table 6 compares the proposed system with every method reported on the five reference benchmarks. Its accuracy is the highest on all five datasets, and every difference with respect to the reference method [9] exceeds the 95% threshold of Equation (16), with z values of 22.8, 3.6, 2.7, 11.7, and 5.5 for MedlinePlus, SymCat 200, SymCat 300, SymCat 400, and SymCat Common, respectively, each computed from the figures reported in Table 6. The number of questions, however, separates two regimes. On MedlinePlus, the proposed system is better on both metrics than the reference method, whereas on the four SymCat variants it asks about twice as many questions as the reference method. Section 4.6 analyzes this trade-off.

The advantage on MedlinePlus, the 893-disease benchmark contributed by the reference study, lies in the structure of the data rather than in the method: every MedlinePlus patient carries all examinations of the true disease, examinations are highly disease-specific, and conditional probabilities estimated from the simulated corpus recover this structure. A likelihood model that matches the generative process therefore needs few questions.

Table 6. Top-1 accuracy (%) and mean number of questions (Q) on the five reference benchmarks

Dataset	Classes	Proposed		Yuan and Yu [9]		REFUEL [3]		INS-DS [8]		FIT [31]	
		acc (%)	Q	acc (%)	Q	acc (%)	Q	acc (%)	Q	acc (%)	Q
MedlinePlus	893	99.27 ± 0.09	1.75	96.2 ± 0.1	3.41	0.0	15.0	71.4	1.34	n/r	n/r
SymCat 200	200	71.83 ± 0.74	18.95	68.8 ± 0.4	8.98	59.9	8.45	46.2	2.77	55.7	12.0
SymCat 300	300	65.33 ± 1.82	20.54	60.4 ± 0.2	9.39	48.5	9.63	39.7	4.47	48.2	13.1
SymCat 400	400	63.20 ± 0.65	21.96	55.5 ± 0.1	10.02	33.6	10.18	31.1	3.25	44.6	14.4
SymCat Common	210	62.10 ± 0.70	23.11	58.1 ± 0.2	9.61	47.6	8.66	36.1	2.99	n/r	n/r

Note: baseline values from [9]. ±, standard error; n/r, not reported

The weaker baselines mark the two extremes of interaction cost. INS-DS stops earliest, after 1.34 to 4.47 questions, and is 25.6 to 32.1 points less accurate than the proposed system. REFUEL never issues a diagnosis action on MedlinePlus and asks questions until the step limit, which yields an accuracy of zero. Both failures originate in the stopping rule rather than in the scoring.

4.2. Generalization to DDXPlus

The reference study did not report results on DDXPlus, but the dataset paper [38] evaluated two learned baselines in a configuration that measures the same quantities as this study: AARLC, the preprint version of the reference method [9] retrained on DDXPlus, and BASD, a neural agent trained by supervised learning to choose the next question and the diagnosis. Their ground-truth pathology accuracy at rank one (GTPA@1) is their top-1 accuracy. Their interaction length (IL) counts the initial evidence together with the inquired evidence, so the number of questions is $Q = IL - 1$. The comparison uses the configuration in which each agent is trained to predict the pathology of the patient; the configuration trained to predict a differential diagnosis optimizes a different objective. Table 7 reports the results.

Table 7. Top-1 accuracy (%), interaction length (IL), and mean number of questions (Q) on DDXPlus

Method	Training	IL	Q	Top-1 accuracy (%)
AARLC [38]	RL and supervised learning	6.73	5.73	99.21 ± 0.78
BASD [38]	Supervised learning	17.99	16.99	97.15 ± 1.70
Proposed system	None (counting only)	3.84	2.84	95.60 ± 0.00

Note: baseline values from [38] ± 95% confidence interval for the baselines and standard error for the proposed system

Without modification, the proposed system reaches 95.60% with 2.84 questions across the 49 pathologies. AARLC is 3.6 points more accurate but asks 2.0 times as many questions, and BASD is 1.6 points more accurate with 6.0 times as many. The pattern mirrors SymCat with the roles reversed: here, the learned systems obtain slightly higher accuracy with more questions. Both baselines were trained on 1,025,602 patients, whereas the proposed system derived its rule base from the same file in 89.1 s without any gradient step. The interaction limit is comparable, because both baselines were run with at most 30 turns [38] and the budget here is $B = 30$. The comparison is nevertheless not an exact head-to-head result for two reasons: the categorical and multi-choice evidence of DDXPlus is binarized into atomic features here, and the evaluation uses the first 1,000 test records rather than the full test file. The zero SE of the proposed system follows from the fixed DDXPlus records.

4.3. Component Ablation

Table 8 and Figure 2 report the contribution of each component as the change in accuracy resulting from its removal or replacement. The variants replace weighted evidence with match-ratio scoring, stop before the first question, select questions randomly, use a heuristic discriminative score instead of information gain (IG), or disable the safety veto. The heuristic scores each symptom according to how evenly it divides the candidate pool, how few candidates list it, and whether it distinguishes the two leading candidates, following the key-differentiator tables used in expert systems. The first row of Table 8 reproduces the proposed system results reported in Tables 6 and 7.

Table 8. Ablation results on six datasets: top-1 accuracy (%) ± SE

Variant	MedlinePlus	DDXPlus	SymCat 200	SymCat 300	SymCat 400	SymCat Common
Proposed system	99.27 ± 0.09; Q = 1.75 ± 0.02	95.60 ± 0.00; Q = 2.84 ± 0.00	71.83 ± 0.74; Q = 18.95 ± 0.40	65.33 ± 1.82; Q = 20.54 ± 0.30	63.20 ± 0.65; Q = 21.96 ± 0.19	62.10 ± 0.70; Q = 23.11 ± 0.09
No weighted evidence	33.90 ± 1.59 (−65.37); Q = 13.70 ± 0.24	60.20 ± 0.00 (−35.40); Q = 8.95 ± 0.00	23.37 ± 0.74 (−48.47); Q = 29.80 ± 0.06	22.23 ± 0.69 (−43.10); Q = 29.85 ± 0.03	16.90 ± 0.72 (−46.30); Q = 29.99 ± 0.01	17.90 ± 0.52 (−44.20); Q = 29.82 ± 0.03
No inquiry	75.93 ± 0.29 (−23.33); Q = 0.00 ± 0.00	28.80 ± 0.00 (−66.80); Q = 0.00 ± 0.00	37.83 ± 0.73 (−34.00); Q = 0.00 ± 0.00	31.70 ± 1.70 (−33.63); Q = 0.00 ± 0.00	28.13 ± 0.28 (−35.07); Q = 0.00 ± 0.00	31.63 ± 0.96 (−30.47); Q = 0.00 ± 0.00
Random questions	89.13 ± 0.23 (−10.13); Q = 10.57 ± 0.16	91.23 ± 0.43 (−4.37); Q = 7.88 ± 0.08	48.97 ± 0.56 (−22.87); Q = 28.66 ± 0.05	40.67 ± 0.50 (−24.67); Q = 28.93 ± 0.09	38.53 ± 0.87 (−24.67); Q = 29.14 ± 0.10	39.70 ± 0.67 (−22.40); Q = 29.53 ± 0.08
Heuristic selector	97.93 ± 0.09 (−1.33); Q = 2.83 ± 0.05	94.70 ± 0.00 (−0.90); Q = 4.88 ± 0.00	57.00 ± 1.46 (−14.83); Q = 24.45 ± 0.18	52.90 ± 2.32 (−12.43); Q = 25.26 ± 0.30	49.83 ± 0.52 (−13.37); Q = 25.83 ± 0.31	47.27 ± 0.63 (−14.83); Q = 27.67 ± 0.12
No safety veto	99.47 ± 0.09 (+0.20); Q = 1.76 ± 0.03	95.60 ± 0.00 (0.00); Q = 2.84 ± 0.00	71.83 ± 0.74 (0.00); Q = 18.95 ± 0.40	65.37 ± 1.79 (+0.03); Q = 20.54 ± 0.30	63.30 ± 0.70 (+0.10); Q = 21.96 ± 0.20	62.13 ± 0.67 (+0.03); Q = 23.11 ± 0.09

Note: change with respect to the proposed system (in parentheses), and Q ± SE; changes are computed from unrounded values

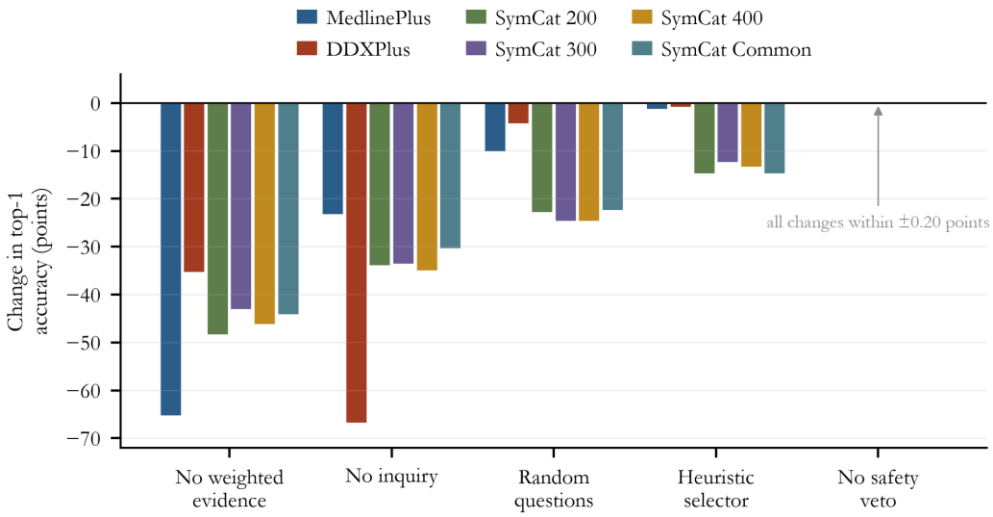


Figure 2. Change in top-1 accuracy with respect to the proposed system when a component is removed or replaced

Weighted evidence combination makes the largest marginal contribution on five of the six datasets. Replacing it with the match-ratio score costs 35.4 to 65.4 points on every dataset, which quantifies the defect described in Section 3.4 and indicates that the poor reputation of rule-based scoring on this task stems from the scoring formula rather than from the symbolic paradigm.

Inquiry is the second most important component overall and the most important on DDXPlus, where the initial evidence alone yields an accuracy of 28.80%. Asking questions at all is worth 23.3 to 66.8 points, and choosing them by IG rather than at random adds 4.4 to 24.7 points, with larger gains on the datasets on which the candidate set remains ambiguous for longer. The largest marginal contribution therefore does not mean that the scoring function carries the method alone: without the inquiry module, the system loses at least 30 points on five of the six datasets.

The heuristic selector stays close to IG only on MedlinePlus and DDXPlus, where it loses 1.33 and 0.90 points and asks 1.1 and 2.0 more questions. On the SymCat variants, it loses 12.43 to 14.83 points and asks 3.9 to 5.5 more questions, so it is worse on both metrics. The heuristic treats membership in a rule list as binary and discards the magnitude of P(e|d). MedlinePlus and DDXPlus tolerate this because their conditional probabilities are high and

similar, whereas SymCat prevalences have a median near 0.25 and a long tail, so ignoring magnitude discards most of the information.

The safety veto provides no measurable improvement in predictive accuracy under the evaluated configurations. Disabling it leaves accuracy unchanged on DDXPlus and SymCat 200, while increasing it by 0.03–0.20 percentage points on the other four datasets, as the veto occasionally eliminates the correct candidate. Sections 4.4 and 4.5 further examine the contribution of the rule slots and the behavior of the safety veto.

4.4. Contribution of the Individual Rule Slots

Because the derivation of the three rule slots is presented as a contribution, each slot was evaluated through a separate ablation. For each experiment, the rule base was rebuilt with the corresponding slot emptied for every disease, and the complete system was evaluated using the modified rule base. Table 9 reports the resulting top-1 accuracy and the mean number of entries in each slot.

Table 9. Top-1 accuracy (%) after removing individual rule slots from the rule base, with the change with respect to the complete rule base in parentheses

Rule base	MedlinePlus	DDXPlus	SymCat 200	SymCat 300	SymCat 400	SymCat Common
Complete (proposed)	99.27	95.60	71.83	65.33	63.20	62.10
Mean size of must / excl / crit	3.7 / 8.0 / 4.6	1.8 / 8.0 / 7.7	0.1 / 1.0 / 7.1	0.2 / 1.9 / 7.2	0.1 / 1.0 / 7.1	0.1 / 1.4 / 6.7
must(d) emptied	99.47 (+0.20)	95.60 (0.00)	71.83 (0.00)	65.37 (+0.03)	63.30 (+0.10)	62.13 (+0.03)
excl(d) emptied	99.27 (0.00)	95.60 (0.00)	71.83 (0.00)	65.33 (0.00)	63.20 (0.00)	62.10 (0.00)
crit(d) emptied	99.27 (0.00)	95.60 (0.00)	71.83 (0.00)	65.33 (0.00)	63.20 (0.00)	62.10 (0.00)
All three emptied	99.47 (+0.20)	95.60 (0.00)	71.83 (0.00)	65.37 (+0.03)	63.30 (+0.10)	62.13 (+0.03)

The ablation results distinguish the roles of the three rule slots. The must-have slot affects candidate filtering through the safety veto, although removing it slightly increases accuracy under the evaluated configurations. The exclusion slot produces no measurable change in accuracy when removed, while the critical slot does not enter the diagnostic scoring function or the veto and serves only an explanatory role. Therefore, the results do not establish a positive contribution of the individual rule slots to predictive accuracy.

The observed effects are consistent with the sizes of the derived slots. A must-have symptom requires $P(e|d) \geq \theta_{must} = 0.95$, which the MedlinePlus simulator produces for every examination and the DDXPlus records produce for a few characteristic findings, giving 3.7 and 1.8 entries per disease on average. On SymCat, where symptoms are drawn independently with a median prevalence near 0.25, the slot holds at most 0.2 entries per disease on average and is empty for most diseases, so its effect cannot exceed 0.10 points. A populated slot is necessary for an effect but not sufficient: DDXPlus has 1.8 must-have symptoms per disease but no effect, because the veto never removed the true disease there and never changed the returned diagnosis (Table 10).

The slot ablation and the veto ablation act on different parts of the system and agree. Emptying all three slots reproduces the accuracy obtained by disabling the veto (Table 8) on all six datasets, which confirms that the slots affect the diagnosis only through the veto. The two configurations differ in their stopping behavior on MedlinePlus: with the veto disabled, the stopping rule still waits while a must-have symptom of the leading candidate is denied, whereas emptying the must-have slot removes this proviso.

4.5. Direct Measurement of the Safety Veto

An accuracy ablation cannot distinguish a mechanism that never fires from one that fires often but redundantly. Table 10 therefore reports five measurements taken on the runs of Table 8: the share of consultations in which at least one candidate is vetoed; the mean share of candidates removed; the share of consultations whose returned diagnosis would change if the vetoed candidates were restored at the final evidence state, with the interview held fixed; the share in which the true disease is vetoed; and, on the runs without the veto, the share whose returned diagnosis belongs to the set V of Equation (10), that is, denies one of its own must-have symptoms or has two of its exclusion symptoms confirmed.

Table 10. Measurements of the safety veto on six datasets (%)

Dataset	Cases with a veto	Candidates removed	Diagnosis changed	True disease vetoed	Violation without veto
MedlinePlus	38.63	13.86	0.17	0.27	0.20
DDXPlus	98.90	50.24	0.00	0.00	0.00
SymCat 200	89.73	2.53	0.03	0.13	0.03
SymCat 300	96.57	4.11	0.03	0.07	0.03
SymCat 400	97.73	3.18	0.07	0.10	0.07
SymCat Common	87.43	3.08	0.03	0.07	0.00

The veto is active but rarely decisive. It fires in 38.6% to 98.9% of consultations and removes, on average, 2.5% to 50.2% of the candidates, yet it changes the returned diagnosis in at most 0.17% of cases. The evidence score of Equation (9) has already ranked constraint violating candidates low before the veto is applied, so the veto mostly removes candidates that would not have been returned anyway. This agreement between the two mechanisms, rather than inactivity of the veto, explains the null result in Table 8.

The last two columns quantify the cost and the benefit. The veto removes the true disease in at most 0.27% of cases, which is the source of the small accuracy loss on MedlinePlus. Without the veto, the returned diagnosis violates one of its own derived constraints in up to 0.20% of cases; with the veto, this share is zero by construction on every dataset, except when every candidate is vetoed and the system falls back to a uniform ranking (Section 3.5). Whether such a guarantee is worth up to 0.20 accuracy points is a deployment decision rather than a benchmark result.

4.6. Trade-off Between Accuracy and Interaction Cost

The accuracy values in Table 6 were obtained with a budget of 30 questions, whereas the reference method asks about nine to ten questions on SymCat. The trade-off was therefore measured directly by rerunning the proposed system on SymCat 200 with six question budgets under the protocol of Section 3.9. Table 11 lists the results, and Figure 3(a) plots accuracy against the number of questions actually asked; the 30-question point is the SymCat 200 entry of Table 6.

Table 11. Top-1 accuracy (%) ± SE and mean number of questions (Q) of the proposed system on SymCat 200 for six question budgets

Budget B	5	9	12	15	20	30
Q	4.65	7.71	9.71	11.50	14.19	18.95
Top-1 accuracy	51.00 ± 0.90	57.93 ± 0.62	61.63 ± 1.01	64.87 ± 0.77	68.30 ± 0.84	71.83 ± 0.74

At the 8.98 questions used by the reference method, the proposed system reaches approximately 60.3%, interpolated between the neighboring budgets, whereas the reference method reaches 68.8%; at a matched budget, the reference method is therefore about 8.5 points more accurate. The proposed system needs about 14.9 questions, 1.7 times as many, to reach the same accuracy. On SymCat, the two systems therefore occupy different points of one trade-off: the proposed system reaches a higher accuracy with longer interviews, whereas the learned stopping rule of the reference method stops earlier at a lower accuracy. On MedlinePlus, no such trade-off arises, because the proposed system is better on both metrics. The stopping rule, not the scoring function, is the component that would have to improve to move the SymCat trade-off.

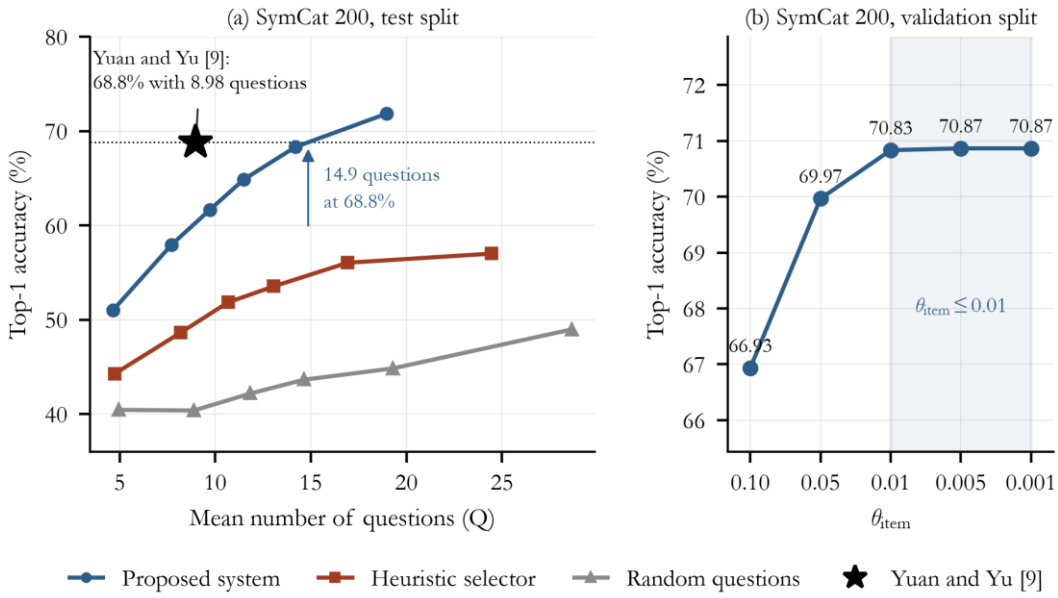


Figure 3. (**a**) Top-1 accuracy against mean number of questions on SymCat 200, with the result of [9] marked by a star. (**b**) Validation accuracy on SymCat 200 as a function of θ_{item}

Figure 3(b) examines whether the results depend on a finely tuned threshold. The rule base of SymCat 200 was derived with five values of θ_{item} between 0.10 and 0.001, and each version was evaluated on the validation split under the protocol of Section 3.9. Accuracy is 66.93% for θ_{item} = 0.10 and 69.97% for 0.05, whereas the three values at or below 0.01 lie within 0.04 points of one another, far below their standard errors of 1.23 to 1.25 points. The response is a plateau rather than a peak: the threshold only needs to be low enough to keep low-prevalence symptoms in the rule base. On MedlinePlus, the threshold has no effect, because the simulator places every conditional probability well above 0.10.

4.7. Discussion

The results demonstrate that a classical probabilistic formulation developed more than half a century ago [20], [21], when supplied with conditional probabilities estimated automatically rather than elicited from experts, can match or outperform published RL methods in top-1 accuracy on five reference benchmarks, although it generally requires more questions to achieve this performance. On DDXPlus, the proposed system remains less accurate than the learned baselines. These findings suggest that the performance gap attributed to learned inquiry policies may partly reflect the underlying probabilistic formulation rather than the inquiry policy alone.

This performance advantage is partly structural. SymCat patients are generated by independently sampling symptoms conditioned on the disease, while MedlinePlus patients are generated under an approximately similar assumption. This setting aligns with the conditional-independence assumption underlying Equation (9). With a uniform disease prior and conditional probabilities estimated from the same generative process, the resulting probabilistic score consequently approaches the optimum for these simulators. This optimality is specific to the simulation settings and should not be generalized to clinical diagnosis, where symptom dependencies, nonuniform disease prevalences, and violations of conditional independence are common. The potential headroom can be illustrated using SymCat 200: a scorer with access to the status of all 322 symptoms in the knowledge base—corresponding to the 332 features reported in Table 4, excluding the ten demographic context features—achieves 76.00% accuracy, compared with 71.83% for the proposed system and 68.8% for the reference method. These results place both systems relatively close to the full-evidence reference score, suggesting that the remaining improvement over the proposed system is limited to a few percentage points under this comparison.

Adding a disease is computationally inexpensive but requires recomputation. Specifically, N, $n(e)$, $P(e)$, and $P(e|\neg d)$ in Equations (1) and (2) change, affecting the lift and discriminative margin in Equation (3) for existing diseases and necessitating the re-derivation of the

rule slots defined in Equations (4)–(6). This process requires rerunning the single-pass procedure described in Table 4, which takes 1.8–89.1 s, rather than conducting a model-training cycle. Although an incremental update is theoretically possible because the system stores counts, this capability was neither implemented nor evaluated in the present study.

The proposed formulation also provides a transparent representation of its diagnostic decisions: the rules can be inspected directly, the evidence weights are derived from symptom probabilities, and each question is associated with an expected information gain measured in bits. However, this transparency is an intrinsic property of the formulation rather than evidence of demonstrated auditability or clinical usefulness. Recent studies on explainable medical AI and AI-based disease detection highlight the importance of considering explanation quality, validation, and clinical applicability beyond predictive performance alone [39], [40]. No clinician evaluation was conducted to determine whether the exposed reasoning is usable, no user study assessed whether the justifications support decision-making, and existing frameworks for evaluating explanations [14] were not applied. Accordingly, the practical value of this transparency remains to be established.

Finally, the effects reported in Tables 8 and 9 should be interpreted as conditional marginal contributions rather than independent component effects, because each ablation removes a component while retaining the remainder of the system. The safety veto provides no measurable improvement in predictive accuracy under the evaluated configurations: disabling it leaves accuracy unchanged on DDXPlus and SymCat 200 and increases it slightly on the other four datasets (Table 8). Nevertheless, the veto removes candidates that violate the encoded constraints and can prevent such candidates from being returned. Its value may therefore extend beyond predictive accuracy, although this potential benefit requires further evaluation.

4.8. Limitations

Several limitations constrain the interpretation and generalizability of these findings. First, all patients in this study are simulated. MedlinePlus and SymCat patients are generated from knowledge bases, while DDXPlus, although publicly released as a dataset, was generated using a rule-based medical model rather than collected from clinical encounters. Across all benchmarks, simulated patients answer questions correctly and consistently. Consequently, the results do not establish clinical effectiveness, and the reported accuracies may overestimate performance in real-world settings, where patients may misremember symptoms, provide inconsistent answers, or report symptoms absent from the underlying knowledge base.

Second, the conditional-independence assumption that makes Equation (9) computationally tractable is also largely reflected in the data-generating processes of the simulators. The observed performance advantage may therefore diminish on clinical records containing correlated symptoms. Third, the stopping rule relies on fixed confidence thresholds rather than a learned or explicitly optimized criterion. This design contributes to the long consultations observed on SymCat and represents a major opportunity for improvement.

Fourth, the natural-language interface was not evaluated because all benchmarks provide structured symptom codes. The present results therefore do not establish whether the system can reliably interpret patient descriptions or sustain effective interaction in natural-language consultations. Fifth, θ_{item} differs across datasets, although its value was selected using the validation split and the sensitivity analysis indicates a performance plateau. This dataset-specific setting may affect reproducibility and warrants consideration when extending the system to new datasets. Sixth, the transparency of the diagnostic reasoning has not been evaluated with clinicians. Although the formulation exposes its rules, evidence weights, and question-selection rationale, whether these elements are understandable, useful, or appropriate for clinical decision support remains unverified.

5. Conclusions

This study provides the classical baseline missing from evaluations of RL methods for interactive symptom inquiry: a training-free symbolic system whose parameters are estimated from counts, whose diagnoses are based on observed evidence only, and whose questions are selected by expected information gain. The comparison with the reference method reveals two distinct performance regimes. On MedlinePlus, where the simulated data are highly structured, the proposed baseline achieves better accuracy and lower interaction cost. On SymCat,

it achieves higher accuracy only with longer interviews, while the reference method performs better at a matched question budget, indicating a trade-off rather than a uniform advantage. On DDXPlus, the learned baselines achieve slightly higher accuracy but require more questions, mirroring the SymCat trade-off. The ablation results show that weighted evidence combination and symptom inquiry account for most of the predictive performance. The must-have and exclusion rules enable the safety veto to enforce the encoded constraints, except when all candidates are vetoed and the system falls back to a uniform ranking. However, the ablation results do not show a measurable improvement in predictive accuracy from these constraints. The critical slot, by contrast, serves an explanatory role and does not directly affect diagnostic decisions.

For practitioners, the proposed system demonstrates that a diagnostic approach whose steps can be decomposed into inspectable terms can be implemented in under two minutes on commodity hardware and remain competitive on published benchmarks. This may be relevant in settings where training infrastructure or labeled interaction data are unavailable. For researchers, the results on SymCat and MedlinePlus underscore the importance of including a classical baseline, as substantial benchmark performance can be achieved without model training. However, because all benchmarks use simulated patients, these findings do not establish clinical effectiveness. Future work should evaluate the system on patient records not generated from knowledge bases, assess the interpretability and usefulness of its reasoning with clinicians, relax the conditional-independence assumption through structured probabilistic models, replace fixed stopping thresholds with a criterion that optimizes the trade-off between diagnostic accuracy and interaction cost, and develop a natural-language interface for free-text complaints.

Author Contributions: Conceptualization, D.M. and N.D.E.N.; Methodology, D.M. and N.D.E.N.; Software, N.D.E.N.; Validation, N.D.E.N.; Formal Analysis, N.D.E.N.; Investigation, N.D.E.N.; Data Curation, N.D.E.N.; Writing—original draft preparation, N.D.E.N.; Visualization, N.D.E.N.; Writing—review and editing, D.M., M.S., W.S., and D.R.I.M.S.; Supervision, D.M.; Project administration, D.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The DDXPlus dataset is publicly available under a CC BY license. The MedlinePlus and SymCat knowledge bases are distributed in the public repository of the reference study. The conversion, derivation, and evaluation code developed for this work, together with the derived rule bases and the per-run experiment records behind every table, is available from the corresponding author upon reasonable request.

Acknowledgments: The authors thank the maintainers of the DDXPlus benchmark and the authors of the reference study for making their datasets and source code publicly available, which enabled verification of the simulation procedure and identification of protocol differences reported in Section 3.8. Generative artificial intelligence (AI) tools were used to assist with language editing and improving the clarity of the manuscript. The authors reviewed and verified the final content and take full responsibility for its accuracy and integrity.

Conflicts of Interest: The authors declare no conflict of interest.

References

[1] C. Yu, J. Liu, S. Nemati, and G. Yin, "Reinforcement Learning in Healthcare: A Survey," *ACM Comput. Surv.*, vol. 55, no. 1, pp. 1–36, Jan. 2023, doi: 10.1145/3477600.

[2] Z. Wei *et al.*, "Task-oriented Dialogue System for Automatic Diagnosis," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2018, pp. 201–207. doi: 10.18653/v1/P18-2033.

[3] Y.-S. Peng, K.-F. Tang, H.-T. Lin, and E. Chang, "REFUEL: exploring sparse features in deep reinforcement learning for fast disease diagnosis," in *Advances in Neural Information Processing Systems 31 (NeurIPS 2018)*, 2018, vol. 31, pp. 7322–7331. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2018/hash/b5a1d925221b37e2e399f7b319038ba0-Abstract.html

[4] H.-C. Kao, K.-F. Tang, and E. Chang, "Context-Aware Symptom Checking for Disease Diagnosis Using Hierarchical Reinforcement Learning," *Proc. AAAI Conf. Artif. Intell.*, vol. 32, no. 1, Apr. 2018, doi: 10.1609/aaai.v32i1.11902.

[5] C. Zhong *et al.*, "Hierarchical reinforcement learning for automatic disease diagnosis," *Bioinformatics*, vol. 38, no. 16, pp. 3995–4001, Aug. 2022, doi: 10.1093/bioinformatics/btac408.

[6] L. Xu, Q. Zhou, K. Gong, X. Liang, J. Tang, and L. Lin, "End-to-End Knowledge-Routed Relational Dialogue System for Automatic Diagnosis," *Proc. AAAI Conf. Artif. Intell.*, vol. 33, no. 01, pp. 7346–7353, Jul. 2019, doi: 10.1609/aaai.v33i01.33017346.
[7] Y. Xia, J. Zhou, Z. Shi, C. Lu, and H. Huang, "Generative Adversarial Regularized Mutual Information Policy Gradient Framework for Automatic Diagnosis," *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 01, pp. 1062–1069, Apr. 2020, doi: 10.1609/aaai.v34i01.5456.
[8] J. Lin, L. Xu, Z. Chen, and L. Lin, "Towards a reliable and robust dialogue system for medical automatic diagnosis," *OpenReview.net*. 2020. [Online]. Available: https://openreview.net/forum?id=TCAmP8zKZ6k
[9] H. Yuan and S. Yu, "Efficient symptom inquiring and diagnosis via adaptive alignment of reinforcement learning and classification," *Artif. Intell. Med.*, vol. 148, p. 102748, Feb. 2024, doi: 10.1016/j.artmed.2023.102748.
[10] W. Chen, C. Zhong, J. Peng, and Z. Wei, "DxFormer: a decoupled automatic diagnostic system based on decoder–encoder transformer with dense symptom representations," *Bioinformatics*, vol. 39, no. 1, p. btac744, Jan. 2023, doi: 10.1093/bioinformatics/btac744.
[11] F. Zhang, G. Sang, Z. Liu, H. Lin, and Y. Zhang, "A doctor's diagnosis experience enhanced transformer model for automatic diagnosis," *Eng. Appl. Artif. Intell.*, vol. 134, p. 108675, Aug. 2024, doi: 10.1016/j.engappai.2024.108675.
[12] C. Combi *et al.*, "A manifesto on explainability for artificial intelligence in medicine," *Artif. Intell. Med.*, vol. 133, p. 102423, Nov. 2022, doi: 10.1016/j.artmed.2022.102423.
[13] K. Zhang, D. Wang, F. Lin, J. Xie, and W. Zhou, "A comprehensive review of explainable artificial intelligence in healthcare methods, evaluation, and clinical integration," *iScience*, vol. 29, no. 3, p. 115026, Mar. 2026, doi: 10.1016/j.isci.2026.115026.
[14] S. Roychowdhury, V. Lanfranchi, and S. Mazumdar, "Evaluating explanation performance for clinical decision support systems for non-imaging data: A systematic literature review," *Comput. Biol. Med.*, vol. 197, p. 110944, Oct. 2025, doi: 10.1016/j.compbiomed.2025.110944.
[15] B. M. Gebremariam, G. T. Aboye, A. A. Dessalegn, and G. L. Simegn, "Rule-based expert system for the diagnosis of maternal complications during pregnancy: For low resource settings," *Digit. Heal.*, vol. 10, Jan. 2024, doi: 10.1177/20552076241230073.
[16] M.-E. Vidal, Y. Chudasama, H. Huang, D. Purohit, and M. Torrente, "Integrating Knowledge Graphs with Symbolic AI: The Path to Interpretable Hybrid AI Systems in Medicine," *J. Web Semant.*, vol. 84, p. 100856, Jan. 2025, doi: 10.1016/j.websem.2024.100856.
[17] Y. Deng, S. Zhao, Y. Miao, J. Zhu, and J. Li, "MedKA: A knowledge graph-augmented approach to improve factuality in medical Large Language Models," *J. Biomed. Inform.*, vol. 168, p. 104871, Aug. 2025, doi: 10.1016/j.jbi.2025.104871.
[18] W. Wallace *et al.*, "The diagnostic and triage accuracy of digital and online symptom checker tools: a systematic review," *npj Digit. Med.*, vol. 5, no. 1, p. 118, Aug. 2022, doi: 10.1038/s41746-022-00667-w.
[19] E. Riboli-Sasco *et al.*, "Triage and Diagnostic Accuracy of Online Symptom Checkers: Systematic Review," *J. Med. Internet Res.*, vol. 25, p. e43803, Jun. 2023, doi: 10.2196/43803.
[20] R. S. Ledley and L. B. Lusted, "Reasoning Foundations of Medical Diagnosis," *Science (80-.).*, vol. 130, no. 3366, pp. 9–21, Jul. 1959, doi: 10.1126/science.130.3366.9.
[21] G. A. Gorry and G. O. Barnett, "Experience with a model of sequential diagnosis," *Comput. Biomed. Res.*, vol. 1, no. 5, pp. 490–507, May 1968, doi: 10.1016/0010-4809(68)90016-5.
[22] F. T. de Dombal, D. J. Leaper, J. R. Staniland, A. P. McCann, and J. C. Horrocks, "Computer-aided Diagnosis of Acute Abdominal Pain," *BMJ*, vol. 2, no. 5804, pp. 9–13, Apr. 1972, doi: 10.1136/bmj.2.5804.9.
[23] R. A. Miller, H. E. Pople, and J. D. Myers, "Internist-I , an Experimental Computer-Based Diagnostic Consultant for General Internal Medicine," *N. Engl. J. Med.*, vol. 307, no. 8, pp. 468–476, Aug. 1982, doi: 10.1056/NEJM198208193070803.
[24] M. A. Shwe *et al.*, "Probabilistic Diagnosis Using a Reformulation of the INTERNIST-1/QMR Knowledge Base," *Methods Inf. Med.*, vol. 30, no. 04, pp. 241–255, Feb. 1991, doi: 10.1055/s-0038-1634846.
[25] D. E. Heckerman, E. J. Horvitz, and B. N. Nathwani, "Toward Normative Expert Systems: Part I The Pathfinder Project," *Methods Inf. Med.*, vol. 31, no. 02, pp. 90–105, Feb. 1992, doi: 10.1055/s-0038-1634867.
[26] D. V. Lindley, "On a Measure of the Information Provided by an Experiment," *Ann. Math. Stat.*, vol. 27, no. 4, pp. 986–1005, Dec. 1956, doi: 10.1214/aoms/1177728069.
[27] R. Howard, "Information Value Theory," *IEEE Trans. Syst. Sci. Cybern.*, vol. 2, no. 1, pp. 22–26, 1966, doi: 10.1109/TSSC.1966.300074.
[28] K. Musgrave, S. Belongie, and S.-N. Lim, "A Metric Learning Reality Check," in *Proceedings of the European Conference on Computer Vision*, 2020, pp. 681–699. doi: 10.1007/978-3-030-58595-2_41.
[29] M. Ferrari Dacrema, P. Cremonesi, and D. Jannach, "Are we really making much progress? A worrying analysis of recent neural recommendation approaches," in *Proceedings of the 13th ACM Conference on Recommender Systems*, Sep. 2019, pp. 101–109. doi: 10.1145/3298689.3347058.
[30] P. Henderson, R. Islam, P. Bachman, J. Pineau, D. Precup, and D. Meger, "Deep Reinforcement Learning That Matters," *Proc. AAAI Conf. Artif. Intell.*, vol. 32, no. 1, Apr. 2018, doi: 10.1609/aaai.v32i1.11694.
[31] W. He, X. Mao, C. Ma, Y. Huang, J. M. Hernàndez-Lobato, and T. Chen, "BSODA: A Bipartite Scalable Framework for Online Disease Diagnosis," in *Proceedings of the ACM Web Conference 2022*, Apr. 2022, pp. 2511–2521. doi: 10.1145/3485447.3512123.
[32] M. Werthaim, M. Kimhi, A. Apartsin, and Y. Aperstein, "A benchmark for evaluating diagnostic questioning efficiency of LLMs in patient conversations," *Sci. Rep.*, vol. 16, no. 1, p. 6121, Jan. 2026, doi: 10.1038/s41598-026-37022-y.
[33] H. Nori *et al.*, "Sequential Diagnosis with Language Models," *arXiv preprint*. Jul. 02, 2025. [Online]. Available: http://arxiv.org/abs/2506.22405
[34] G. A. Prenosil *et al.*, "Neuro-symbolic AI for auditable cognitive information extraction from medical reports," *Commun. Med.*, vol. 5, no. 1, p. 491, Nov. 2025, doi: 10.1038/s43856-025-01194-x.
[35] M. Y. H. Wong, J. C. L. Ong, N. Shah, D. C. Klonoff, and B. Sheng, "Neuro-symbolic artificial intelligence in medicine," *Nat. Biomed. Eng.*, vol. 10, no. 8, pp. 1507–1512, Jul. 2026, doi: 10.1038/s41551-026-01728-1.
[36] P. Yang *et al.*, "LMKG: A large-scale and multi-source medical knowledge graph for intelligent medicine applications," *Knowledge-Based Syst.*, vol. 284, p. 111323, Jan. 2024, doi: 10.1016/j.knosys.2023.111323.

[37] X. Li and Q. Shen, "A hybrid framework based on knowledge distillation for explainable disease diagnosis," *Expert Syst. Appl.*, vol. 238, p. 121844, Mar. 2024, doi: 10.1016/j.eswa.2023.121844.

[38] A. F. Tchango, R. Goel, Z. Wen, J. Martel, and J. Ghosn, "DDXPlus: A New Dataset For Automatic Medical Diagnosis," in *NIPS'22: Proceedings of the 36th International Conference on Neural Information Processing Systems*, Oct. 2022, pp. 31306–31318. doi: 10.48550/arXiv.2205.09148.

[39] E. Winarno, I. M. Nur, A. Karim, S. Amri, I. E. Wirdati, and P. W. Adi, "An Explainable Multimodal Framework for Chest X-Ray Alert Classification Using Radiology Reports and Images," *J. Comput. Theor. Appl.*, vol. 3, no. 4, pp. 647–666, May 2026, doi: 10.62411/jcta.16023.

[40] A. S. Egbunu and A. M. Okedoye, "Harnessing Artificial Intelligence for Early Disease Detection: Opportunities and Challenges in Modern Healthcare," *J. Comput. Theor. Appl.*, vol. 3, no. 3, pp. 384–401, Feb. 2026, doi: 10.62411/jcta.15367.