RDmaster: a novel phenotype-oriented Q&A dialogue system assisting differential diagnosis of rare diseases

# Abstract

Differential diagnosis of rare diseases is of paramount importance due to the extensive overlap of their phenotypic characteristics, yet there are few relevant studies. Drawing inspiration from the recent proliferation of dialogue diagnostic systems, we propose a phenotype-oriented Q&A dialogue system, known as RDmaster, which endeavors to collect additional phenotypes and aid clinicians in the differential diagnosis of rare diseases. Two Bayesian diagnostic approaches, PheLR for phenotypic analysis and LIRICAL for genotypic analysis are integrated into our system to improve clinical interpretability. To achieve an efficient dialogue strategy, a novel metric called Adaptive Information Gain and Gini Index (AIGGI) is proposed to evaluate the expected gain of interrogated phenotypes within real-time diagnostic states. An extensive simulation evaluation has validated the feasibility of our dialogue strategy from multiple perspectives. A dialogue diagnostic trial involving 238 published patients revealed that RDmaster outperforms most competing tools and can enhance diagnostic accuracy through human-computer interaction. We also innovatively compared the performance of professional rare disease diagnostic tools and large language models in generating differential diagnosis lists for rare diseases. RDmaster has been implemented as an online web-based platform and offers clinicians exceptional visualization support.

# Introduction

Rare diseases (RDs), mostly genetic in nature, are defined by the World Health Organization (WHO) as conditions or disorders that affect a prevalence of 6.5 to 10 per 10,000 individuals [1]. Despite being referred to as “rare”, there are over 7,000 RDs that collectively affect tens of millions of people in the general population, creating a pressing clinical demand for RD diagnosis and treatment [2, 3]. RDs pose a great challenge for the clinical interview and further diagnosis due to their low individual frequency, complex genetic profiles, and the high heterogeneity and overlap of their phenotypic features [4, 5], which lead to many RD patients being misdiagnosed for years or remaining undiagnosed [6].

In clinical scenarios, the clinical interview is a vital step in the differential diagnosis (DDX) process, which plays a crucial role in decreasing diagnosis errors, enhancing diagnostic accuracy, and ultimately reducing healthcare costs [7, 8]. DDX involves conducting a systematic evaluation of the patient's symptoms and test results, comparing them to a list of potential diseases that can cause similar symptoms, and ultimately arriving at the correct diagnosis by narrowing down the possible options. The implementation of electronic differential diagnostic support (EDS) system for RDs is urgently needed for several reasons [9]. Firstly, unlike common diseases, clinicians often have limited experience in diagnosing RDs. Secondly, the broad phenotypic overlap of RDs often presents clinicians with multiple potential diagnoses. Lastly, the diversity of RDs, with thousands of conditions associated with nearly ten thousand phenotypes, can make clinical interviews challenging.

Currently, various knowledge bases (KBs) have been established to improve the management and treatment of RDs, e.g., OMIM [10], Orphanet [11], and Human Phenotype Ontology (HPO) [12, 13]. With the aid of these KBs, numerous diagnostic tools and algorithms have been developed to assist clinicians in accurately identifying the diseases and causative genes for RD patients [14-20]. A majority of these mainstream tools utilize the HPO to standardize patient signs and symptoms as input, sometimes supplemented by patient sequencing data in VCF format, and ultimately output a prioritized list of potential target diseases or disease-causing genes through various algorithms. The diagnostic pattern of these tools is to obtain a one-time diagnosis based on restricted patient information and does not include subsequent differential diagnosis procedures in instances where multiple competing diagnoses exist, i.e., they are DDX generators (only generate a candidate list) rather than DDX executors [21].

Nowadays, artificial intelligence (AI) research is increasingly attracting attention in the field of medical diagnosis [22]. ChatGPT, a recently most-popular general chatbot based on the generative pre-trained transformer (GPT) architecture [23-25], has shown great potential in multiple domains, e.g., assisting clinical diagnosis according to common chief complaints [26, 27]. However, these large language models (LLMs) cannot guarantee absolute correctness and lacks authoritative credibility when conversing about specialized domain knowledge and has not been validated in the field of RD diagnosis. Many recent studies have explored reinforcement learning (RL) techniques for dialogue diagnosis, mainly the deep Q-network (DQN) algorithm [28], which simulate clinical interview in user-agent interaction to collect additional patient symptoms and thus improve diagnostic accuracy [29-31]. Specifically, the DQN algorithm calculates Q-values for candidate symptoms, which measure rewards for its agent asking for different symptoms. This approach represents an innovative way to assist DDX but has limitations when applied to RD diagnosis. One major limitation is the scarcity of conversation corpus for RDs. Additionally, most of these studies only involve a few diseases and several dozens of symptoms. For RDs, the vast expansion of the research object scale leads to a reduction in the applicability and efficiency of DQN. But the form of human-computer interaction that simulates clinical interviews and the idea of questioning symptoms by agents based on calculated rewards have inspired us.

In this study, we present a novel EDS system assisting the differential diagnosis of rare diseases, referred to as RDmaster, which enables the capture of additional crucial symptoms beyond the patient's initial medical record through phenotype-oriented question-and-answer (Q&A). Two clinically interpretable algorithms based on the likelihood ratio paradigm are integrated into our system to support the multi-omics diagnosis of RDs [32], specifically PheLR for phenomic analysis and LIRICAL for genomic [33, 34]. We also propose a novel indicator, AIGGI, by reference to two classic decision tree construction metrics (information gain and Gini index) [35, 36], to measure the comprehensive benefits of users affirming or negating interrogated phenotypes within a real-time diagnostic status. RDmaster is implemented as an online web tool and provides good visualization support (http://rdmaster.nbscn.org/). We conducted a simulation experiment to qualitatively analyze our dialogue strategy and confirm its feasibility. We further conducted a comprehensive diagnostic test (on both phenomic and genomic) on a cohort of 238 online published RD patients and validated that: (i) our dialogue strategy is effective in collecting diagnostically beneficial symptoms and improving the prioritization of target diseases and causative genes; (ii) LLMs, particularly GPT-3.5 and GPT-4, shows promising performance in RD diagnosis and have tremendous potential; (iii) compared to many KB-based diagnostic tools and the state-of-the-art LLM (GPT-4), our system has significant advantages in diagnostic accuracy and differential diagnosis execution of RDs.

# Results

## RDmaster overview: clinical scenarios and workflow

RDs are numerous and mostly present complex, overlapping phenotypic features, making the aids for DDX of RDs clinically significant. To address this, we proposed an EDS system called RDmaster, which aims to capture diagnostically beneficial phenotypes (signs and symptoms) in the form of human-computer interaction and make better diagnoses based on patient sequencing data and enriched phenotypic information.

An overview of RDmaster 's workflow is illustrated in Fig. 1. The procedure begins with an initial clinical report, in which users report the patient's phenotypic abnormalities (expressed in HPO terms) and genetic variation data in a Variant Call Format (VCF) [37]. RDmaster then engages in a multi-round Q&A dialogue with users to collect additional phenotypes beyond their initial report, to execute a DDX. Each turn of the dialogue starts with asking users if the patient has a specific phenotype, to which users can respond with “*Yes*”, “*No*”, or “*Not sure*”. Users are also allowed to proactively report phenotypic information during the conversation. User response is then fed-back to our agent to update the diagnostic status and take further actions, i.e., our agent will make a further diagnosis according to enriched patient information and continue questioning in the next turn. The human-computer interactions outlined above perform through RDmaster's user interface (UI) module. RDmaster's core functions include an interpretable multi-omics diagnostic method and an efficient dialogue strategy, both designed based on authoritative KBs and tools.

The current implementation of RDmaster can diagnose 4,257 rare diseases (including 2,380 rare genetic diseases), which collectively have 8,161 different phenotypic abnormalities annotated (an average of 26 annotations per disease). A total of 16,226 phenotypic abnormalities of HPO constitute the controlled vocabulary to normalize signs and symptoms of patients, of which 9,087 RD-related phenotypes are candidate questioning terms during Q&A.

## Implementation of RDmaster: an online dialogue diagnostic platform for RDs

RDmaster is an online web-based tool (http://rdmaster.nbscn.org/) that comprises two main steps: clinical information collection and dialogue diagnosis, as depicted in Fig. 2. The UI for clinical information collection consists of several components (Fig. 2a), including fields for patient age and gender, a file uploader for variant data in VCF format, and a character-matching input box that prompts for matched HPO terms as a drop-down list. The SciGraph, developed by the Monarch Initiative as an HPO annotator [38], has been integrated to annotate user-provided free text with HPO concepts. To provide a comprehensive description of patient phenotypic features, entered HPO terms can be flexibly organized as “Present Phenotypes” or “Absent Phenotypes”. Only the “Present Phenotypes” field is mandatory while other fields are optional. If a VCF file is uploaded, a phenotype-genotype integrated diagnosis will be performed. If not, RDmaster will conduct a phenotype-driven diagnosis. Upon completion of the initial report, users can proceed to the dialogue diagnosis step. The RDmaster's dialogue diagnosis is supported by comprehensive and multi-dimensional data visualization (Fig. 2b). The UI of this phase is mainly divided into four modules:

* Dialogue Q&A module (Fig. 2b, part 1), which facilitates human-computer interaction through phenotype-oriented Q&A. Users need to judge the interrogated phenotype in each round based on the patient's phenotype features and can also proactively report symptoms. A three-layer network, constructed based on HPO hierarchy, is designed to illustrate the semantic relationship between the current interrogated phenotypes and top-10 ranked diseases.
* Disease evaluation module (Fig. 2b, part 2), which provides detailed phenotypic and genotypic (if a VCF file is uploaded) analysis results for the top-10 ranked diseases. Specifically, the analysis results mainly include the prior probability (evaluated by its point prevalence) and posterior probability of tested disease, the disease-patient phenotypic match ratio, and likelihood ratio scores of collected phenotypes and disease-causing genes.
* Diagnostic state module (Fig. 2b, part 3), which indicates the real-time diagnostic status, mainly includes a four-dimensional (the posterior probability and match ratio of top 1 disease, the entropy and Gini index calculated from posterior probability distribution of all tested diseases) radar chart, where a larger area represents a lower diagnostic uncertainty.
* Patient phenotypic state module (Fig. 2b, part 4), mainly a sunburst chart of patient phenotype features drawn based on the HPO structure and annotation propagation rule [39].

These modules are updated in real-time during Q&A, leveraging Spring context and Caffeine caching technologies in our back-end service to minimize request/response time and enable real-time conversations, allowing our agent to ask a new question within a few seconds after receiving user feedback.

## Analysis of RDmaster's dialogue strategy

RDmaster is designed to assist in the DDX of RDs through a phenotype-oriented, multi-turn Q&A dialogue, which aims to reduce diagnostic uncertainty and narrow down the possible diseases. During each turn of the dialogue, the phenotype with the greatest expected gain (indicated by our proposed AIGGI, see details in the Methods section) will be asked by our agent. To evaluate the validity and plausibility of the RDmaster's dialogue strategy, we designed simulation experiments and tested tens of thousands of cases to investigate the association between candidate phenotypes for interrogation and candidate diseases in different diagnostic statuses (see detailed analysis result in Supplementary Note 3).

### A questioning strategy consistent with the differential diagnosis

The association between interrogated phenotype and the top 10 candidate diseases within different diagnostic statuses is illustrated in Fig. 3a. When diagnostic uncertainty is high ( greater than 0.9), the interrogated phenotype has similar association scores with all top 10 RDs. In this case, our system focuses on information gain, which helps minimize diagnostic uncertainty by asking for the phenotype with the highest differentiation of candidate diseases, no matter whether users answer yes or no. As diagnostic uncertainty decreases ( less than 0.9 and more than 0.5), the top-ranked diseases start showing diagnostic advantages over others. In this stage, the system tends to ask for phenotypes in the order of priority of candidate diseases, with the goal of validating the patient's potential phenotype to accentuate the target disease, which is likely in the first few positions. When diagnostic uncertainty is low ( less than 0.5), at which point the first-ranked disease has a significant diagnostic advantage over others, the system focuses on asking for phenotypes related to the second and more inferior diseases, with the expectation that users will exclude these phenotypes, i.e., excluding crucial phenotypes of non-first diseases (especially the second-ranked disease) will better improve diagnostic purity than verifying phenotypes of the first-ranked disease in this phase. Overall, as the diagnostic status changes, our dialogue strategy can adaptively adjust to follow DDX by maximizing the expected Q&A gain.

### Two intention-considered questioning strategies

When diagnostic uncertainty is low, the expected gain calculated from AIGGI may be overly focused on excluding lower-ranked disease phenotypes. To address this, we further analyzed two intention-considered indicators, AIGGI-P and AIGGI-N, which take user intentions into account and serve as complements to AIGGI, as shown in Fig. 3a. For AIGGI-P, which is oriented towards verifying phenotypes, as diagnostic uncertainty decreases, this strategy increasingly focuses on the first-ranked disease. This strategy is logical as validating the phenotype of the first-ranked disease is more likely to reduce diagnostic uncertainty than focusing on others. For AIGGI-N, which is oriented towards denying phenotypes, as diagnostic uncertainty decreases, the disease with the highest association score with the questioned phenotype changes in the order of diagnostic ranking, from 5th to 4th and finally to second, while the association score for the first-ranked disease is low. This strategy is reasonable when the user intends to exclude phenotypes. AIGGI-P and AIGGI-N provide two different questioning strategies by considering user intentions, which address the shortcomings of the original AIGGI strategy of rigidity and low user participation. Combining AIGGI-P and AIGGI-N with AIGGI can provide users with flexible and reliable phenotype questioning strategies.

### KB-based prioritization of candidate phenotypes

The importance of disease-related phenotypes varies depending on their annotation frequencies to a particular disease. Specifically, for several competing diseases, phenotypes annotated as high frequency are more valuable to be interrogated than those annotated as low frequency, which is involved in our dialogue strategy. Although our agent only asks for a phenotype with the maximum expected gain in each turn, analyzing more candidate phenotypes can further reveal our dialogue strategy. We conducted simulations to count the association scores of the top 100 candidate phenotypes for interrogation with the top 10 candidate diagnoses in the initial conversation turn under our proposed three expected gain indicators and different diagnostic states, as shown in Fig. 3b. In most situations, the top-ranked phenotypes have higher association scores than the lower-ranked phenotypes, which is reasonable since phenotypes with stronger associations are more worth to be validated or excluded. Generally, for a top-ranked disease, our questioning strategy prioritizes more efficient phenotypes which have higher association scores with the disease, regardless of user intention or diagnostic uncertainty.

### A strategy implicating phenotypic hierarchical semantics: a case study

The application of the annotation propagation rule to our dialogue strategy makes it possible to imply phenotypic hierarchical semantic relationships beyond the disease-phenotype annotation file. We tested a published female patient (Patient No 8 in PMID:17056636) diagnosed with cardiofaciocutaneous syndrome [40], using all her phenotypic features as an initial clinical report. In her differential diagnosis list, with is 0.59, the posterior probability of the first-ranked disease *Cardiofaciocutaneous syndrome* (ORPHA:1340) is 99% while the second-ranked disease *Autosomal recessive cutis laxa type 2, classic type* (ORPHA:357074) is 79%, which has an unreported phenotype *Delayed closure of the anterior fontanelle* (HP:0001476) annotated as always present, and the probabilities of others are very low. In this situation, excluding always-present phenotypes of the second-ranked disease would be most helpful in reducing diagnostic uncertainty, therefore, our agent places it first among all candidate phenotypes. Meanwhile, for the first-candidate phenotype in this Q&A turn, its two parent terms (*Abnormality of the anterior fontanelle* and *Delayed cranial suture closure*) are placed in the 2nd and 3rd positions, its two grandparent terms (*Abnormality of fontanelles* and *Abnormality of cranial sutures*) are 4th and 5th and its great-grandparent term (*Abnormality of the fontanelles or cranial sutures*) is 6th. When a phenotype is always present for a disease, its ancestor terms are also obligate. Therefore, our agent ranks *Delayed closure of the anterior fontanelle*'s ancestor terms at the top position in order of semantic distance, in case of users answer “*Not sure*” in this turn, then our agent will continue to ask its broader and more general ancestor terms.

## Diagnostic performance of RDmaster

A cohort of 238 online published rare disease patients was built to test the diagnostic performance of RDmaster, of which 225 cases had a rare genetic disease. A summary of collected cases is provided in Table 1. In the benchmarking performance of RDmaster, we introduced several competing approaches, including four phenotype-driven RD diagnosis tools (Phenomizer [14], BOQA [15], PhenoPro [20], and LIRICAL [34]), a previous state-of-the-art genomic analysis tool Exomiser [18, 41], and large language models GPT-3.5 and GPT-4.

### Dialogue diagnosis performance evaluation of RDmaster

We first tested the dialogue diagnostic performance (phenotype analysis only or combined with genotype analysis) of collecting additional patient phenotype information following our default dialogue strategy, i.e., calculating the expected gain of candidate interrogated phenotypes by our proposed AIGGI. For each patient, half of their phenotypic features were randomly chosen as an initial report (with or without variant data) with a phenotype-oriented multiple-turn Q&A dialogue performed without intention-considered. Ten experiments were performed for each case to consider the variance of the initial report in realistic clinical operations. During the Q&A dialogue, answer “*Yes*” to the patient's present phenotypes, answer “*No*” to excluded phenotypes, and answer “*Not sure*” if interrogated phenotype is not mentioned in the case report. In addition to analyzing the pre- and post-dialogue diagnostic performance of our approach, we also compared it with several competing tools in this benchmark.

For phenotype-only diagnosis, our approach significantly outperformed other competing phenotype-driven diagnostic tools both pre- and post-dialogue, as shown in Fig. 4a and Fig. 4b. As the turn of Q&A increases, the ranking distribution of the target disease becomes increasingly advanced. After 50 turns of Q&A, the cumulative number of target diseases ranked in the top 10 (56.9%, 1354 out of 2380 conversations) was markedly higher than pre-dialogue (52.7%, 1255 out of 2380 conversations). For disease-causing variant prioritization, the test result (shown in Fig. 4c) indicates that RDmaster has a remarkable advantage over Exomiser. After 50 turns of Q&A, the number of disease-causing variants placed in the first rank is the highest (67.9%, 1527 out of 2250 conversations), compared to 65.9% before the dialogue and 52.6% for Exomiser. In conclusion, RDmaster has excellent capabilities for RD diagnosis and disease-causing gene prioritization. This evaluation experiment demonstrates that our proposed phenotype-oriented Q&A dialogue diagnosis can assist clinicians in collecting additional phenotypic information that facilitates diagnosis and leads to better diagnoses.

### Comparison with LLMs for RD differential diagnosis

We also tested the performance of KB-based RD diagnostic tools (including our system) and large language models (LLMs, mainly GPT-3.5 and GPT-4) in generating differential diagnosis (DDX) lists for RDs based on the phenotype information of collected patients. We utilized the ChatGPT online web tool (<https://chat.openai.com/chat>, 2023-3-14 Version) to test GPT-3.5 and GPT-4. The ChatGPT prompt was designed as “Please give a ranked list with 10 differential diagnoses (rare diseases only) based on the HPOs provided below”, followed by a patient's phenotypes standardized as HPO terms. In this benchmark, for each patient, all existing phenotypes were taken as input, and the rankings of the target disease in ranked DDX lists generated by different approaches were manually verified and counted. A summary of DDX results of our cohort with 238 patients is in Table 2.

As shown in the comparison of the isolated methods in Table 2, GPT-4 showed promising performance in RD differential diagnosis, outperforming GPT-3.5 comprehensively. For the ability to put the target disease first, GPT-4 was better than Phenomizer, comparable to BOQA and PhenoPro, but weaker than LIRICAL and our system (which showed the best performance). When the size of the DDX list increased, KB-based methods showed a significant improvement in the recall for the target disease and were generally superior to GPTs when the list size reached 10. It is worth noting that GPT sometimes recommends broader concepts for the target disease, e.g., not considering disease subtypes or inheritance modes. If such recommendations are considered, the performance of GPT-4 in ranking the target disease as No.1 is only weaker than that of RDmaster.

Furthermore, KB-based tools and GPTs sometimes exhibit discrepancies in performance when dealing with different diseases. Specifically, in some cases, KB-based tools perform poorly while GPTs perform well, and there are also cases where all KB-based tools perform well while GPTs fail to diagnose the disease. We provide detailed testing results in the supplementary materials (Supplementary Note 5). To reflect this difference, we calculated the joint diagnostic performance of RDmaster and other methods and directly took the better diagnostic result of the two methods, as shown in the joint approaches in Table 2. The results showed that although the standalone diagnostic performance of GPT4 is mediocre, its combination with RDmaster yields significantly better results than the combination of other KB-based tools with RDmaster, suggesting a potential complementary diagnostic advantage between KB-based methods and GPTs.

# Discussion

A recent scoping review has revealed that most existing clinical decision support systems (CDSS) for RDs do not involve the execution of DDX [42]. Taking inspiration from the growing popularity of clinical dialogue systems, we proposed and implemented a novel phenotype-oriented Q&A dialogue system assisting the DDX of RDs, referred to as RDmaster. Our system is designed to collect additional crucial symptoms beyond the patient's initial medical record through human-computer interaction, differentiate between potentially multiple competing diagnoses, and ultimately identify target diseases and causative genes more accurately. Our system has four core advantages: (1) the application of an efficient and interpretable diagnostic approach, which is probably state-of-the-art in the field of RD diagnosis; (2) a metric to measure the expected benefit of inquiring about a specific phenotype in the real-time diagnostic state; (3) a well-designed online platform with rich visualization support; (4) an enrichment of phenotypic information in patient’s medical record from the perspective of differential diagnosis.

Two Bayesian methods based on the likelihood ratio paradigm are used for the RD diagnosis of our system [32]. Our proposed PheLR is a phenotype-driven analysis method [33], which incorporates prior probabilities computed with prevalence data of RDs, and provides an evaluation of the contribution of individual phenotype to each disease as well as the patient-to-disease phenotype matching rate, improving its clinical interpretability and diagnostic specificity. The annotation propagation rule is applied to PheLR to enhance its robustness to inaccurate phenotypes and noise. Additionally, we integrated LIRICAL for genomic analysis [34]. The performance of our integrated diagnostic approach in RDmaster was validated in this study, and the accuracy of the initial diagnosis before dialogue was found to be higher than a variety of current diagnostic methods. One important reason for adopting Bayesian analysis into our system is its logical simplicity, diagnostic robustness, and clinical acceptance. Another is that the posterior probability and phenotypic frequency of diseases are both “probability” which makes the combination of these metrics more reasonable in our expected benefit calculation of candidate phenotypes.

The decision tree (DT) in machine learning is a typical classification method that uses an inductive algorithm to generate interpretable rules and a binary-tree-formed decision process. If leaf nodes represent diseases and non-leaf nodes represent phenotypes being asked, then the execution of a decision tree can be seen as a dialogue diagnostic process. Two classic decision tree metric functions, information entropy (for ID3) and Gini index (for CART), both mathematically measure uncertainty or impurity for each split condition, and although widely proven to have similar accuracy for classification problems [43, 44], they have different expressions and measurement focuses. A previous study has shown that a linear fusion of these two metrics can achieve satisfactory classification performance [45]. We constructed these two classification decision trees based on disease phenotype associations from Orphanet (see details in Supplementary Note 2). Results showed that information entropy performed better for messier data, by selecting the most conditional discriminative phenotype at each node. For example, the root node of the information entropy decision tree was *Abnormality of the musculature* (HP:0003011), which was associated with 2117 ORPHAs and not with the remaining 2140. On the other hand, Gini index was more effective in rapidly purifying candidates, e.g., obtaining the fastest diagnosis after only two decisions in the Gini decision tree. To take advantage of both these strengths, we proposed an adaptive fusion metric, AIGGI, which focuses on reducing uncertainty when diagnostic uncertainty is high, primarily using information gain, and quickly obtaining a diagnosis when diagnostic purity is high, mainly using Gini index.

The objectives of DQN-based dialogue diagnosis systems and a limited number of human-computer interaction DDX applications for RDs [29-31, 46, 47], are to capture additional symptoms existed in patients and ultimately improve diagnostic accuracy. However, this approach may not always align perfectly with the process of DDX, as sometimes excluding phenotypes belonging to competing diagnoses that patients do not possess can better aid in narrowing down the possible diseases. As decision trees always have binary branches of yes or no, our dialogue strategy considers the overall benefit of users’ answering yes or no, making it more comprehensive and in line with DDX. Furthermore, two intension-considered dialogue strategies were designed in case users explicitly want to verify or exclude phenotypes.

Many studies have explored LLMs, such as GPT-3, on their performance in diagnosing common diseases [26, 27]. It has been proved that GPT-3 is significantly better than a famous DDX generator Isabel Pro [48, 49]. But the effectiveness of LLM in the DDX of RDs has not been well studied before. In this work, we compared the performance of KB-based RD diagnostic tools and LLMs (GPT-3.5 and GPT4) in generating DDX lists for RDs and demonstrated the enormous potential of LLMs in RD diagnosis. However, the clinical application of LLMs for RD diagnosis needs to be cautious. Even the most advanced GPT-4 still has some native limitations: (i) instability, meaning that the DDX list may vary even when using the same prompt; (ii) inability to analyze sequencing data, while RDs are mostly genetic diseases; (iii) cannot recommend RDs well that have not been trained or have limited training data; (iv) bias of professional knowledge, e.g., confusing the ORPHA codes of RDs.

Nonetheless, there are several limitations to this work. Firstly, the diagnostic accuracy of RDmaster cannot be fully guaranteed, as patients usually do not present all phenotypes of the target disease and often exhibit symptoms that are not related to the target disease. If enough evidence to support a viable diagnosis is not in the initial report, the target disease may be placed in a relatively lower position in the DDX list, in which case our dialogue diagnosis will not perform well. Secondly, the diagnostic methods and Q&A strategies adopted by RDmaster are designed based on KBs. However, even in the most authoritative KB Orphanet, there are still many RDs that lack phenotypic and genetic evidence. As of February 2022, Orphanet only recorded 4257 RDs with phenotypic information and 3884 RDs with associated genes. With the advent of next-generation sequencing technology, hundreds of new disease-gene associations are revealed per year [50]. As more and more research and reporting on RDs are conducted, the relevant KBs will become increasingly complete.

# Methods

## Data sources

RDmaster adheres to evidence-based medicine and all evidence data is derived from authoritative knowledge bases. The Human Phenotype Ontology (HPO) is used to standardize clinically described symptoms and signs (collectively referred to as phenotypes) and serves as the entity for inquiry during the diagnostic process. The hp.obo file (version 2022-02-14), which models the HPO, and the phenotype.hpoa file (version 2022-02-14), which provides HPO frequency annotations for 4257 ORPHA diseases, can be obtained from the HPO website (<https://hpo.jax.org/app/>). Additionally, we downloaded a series of RD-related files from Orphadata (<https://www.orphadata.com/>), including the en\_product6.xml (version 2022-02-01), which provides ORPHA-gene relationships, en\_product9\_prev.xml (version 2022-02-01), which offers epidemiology data of ORPHAs, and en\_product9\_ages.xml (version 2022-02-01), which has onset and death ages and inheritance types of ORPHAs. These files give multivariate information and promote the diagnosis of RDs. To aid in genomic analysis, Exomiser is integrated to annotate, filter, and score likely causative variants, and Exomiser’s supporting data, the Jannovar-derived exome functional annotation files (2102\_hg19 and 2102\_hg38) [51], can be downloaded from <https://data.monarchinitiative.org/exomiser/data/index.html>.

## Diagnostic method of RDmaster

In order to perform multi-omics diagnosis, two interpretable RD diagnostic approaches with the likelihood ratio paradigm, PheLR for phenotypic analysis and LIRICAL for genotypic analysis, are integrated into our system. The likelihood ratio (LR) is the percentage of people with a specific test result divided by the percentage of healthy people with the same result. Theoretically, people with a specific disease should present more typical phenotypes and variants than people with other diseases or healthy people, which causes a relatively high LR. According to Bayesian principles, the prior odds of disease multiplied by the LR of patient test results gives the posterior odds, which can be converted to posterior probability by simple arithmetic [32]. Ultimately, candidate diseases and causative genes are prioritized based on calculated posterior probabilities in RDmaster.

PheLR [33], a phenotype-driven LR analysis method developed in one of our previous studies, is integrated into RDmaster to facilitate phenotype-based diagnosis. It utilizes epidemiological data and onset & death age data from Orphanet to estimate the prior probabilities of RDs. Multiple computational models of LR for both present and absent phenotypes are designed in PheLR based on the annotation propagation rule [39], which quantifies the contribution of individual phenotypic observations to candidate diagnoses. LIRICAL [34], a recently developed method for interpretable clinical genomics and concurrently providing phenotypic analysis, is integrated into RDmaster for genomic analysis only. It models the expected counts of pathogenic variants in disease-causing genes as a Poisson distribution based on different modes of inheritance. The LR of a genotype is the probability of observing the number of pathogenic variants within a Poisson distribution of this genotype in a sequenced individual compared to the situation in which the individual does not have a genotype-associated disease and the variants originate from the population background. The pathogenicity of variants in LIRICAL is determined according to Clinvar database or pathogenicity scores calculated from Exomiser [52]. In our system, phenotypic and genotypic analysis can be performed separately or simultaneously, but separate single genome analysis is not recommended.

## Dialogue strategy of RDmaster

During phenotype-oriented Q&A dialogue, the selection of candidate phenotypes is a crucial task. We propose a novel metric, Adaptive Information Gain and Gini Index (AIGGI), to evaluate the expected gain of candidate phenotypes in real-time diagnostic states. The AIGGI metric is inspired by the metric functions of two decision tree algorithms (ID3 and CART). The strategy aims to reduce diagnostic uncertainty (primarily through information gain) when multiple candidate diseases have relatively high posterior probabilities and to improve diagnostic specificity (primarily through Gini index) when there are a few significant candidate diseases.

As with the binary-tree decision process, our dialogue strategy allows users to respond “*Yes*” or “*No*” to interrogated phenotypes, corresponding to the positive or negative branches of the decision tree nodes, respectively. Rather than a dichotomous distribution (whether a phenotype is associated with a disease or not), we use a numerical probability , which represents the degree to which phenotype is related to disease , to measure the weights of two branches. We take the median value (refer to as ) of the frequency range (e.g., 0.9 for “Very frequent (99-80%)”) in the HPO annotation file as the numerical probability, and apply the annotation propagation rule on HPO to exploit implied phenotypic features of diseases. can be calculated as:

where is ’s annotated phenotypes and represents ancestor terms of phenotype . Real-time diagnostic status is indicated as the uncertainty or purity of the diagnostic result, which can be measured using the information entropy and Gini index calculated from the posterior probability of all candidate diseases:

and

where is a set of numerical weight and is the sum of the elements in . represents all candidate RDs and is the posterior probability of disease . and can represent the uncertainty and purity of the diagnostic status. For a candidate phenotype within real-time diagnostic status, its conditional entropy and information gain can be calculated as:

and ’s conditional Gini index and Gini gain can be calculated as:

In Equation 6 and 8, the first term corresponds to the positive branch and the second term to the negative branch. equal 1 minus . is the weight of positive branch and is calculated as the sum of divided by the sum of .

To balance the strengths of information gain (better performance on messier data) and Gini index (superior performance in pure data classification), we propose a novel metric AIGGI to measure the expected gain of candidate phenotype within real-time diagnostic status:

where is an indicator reflecting the uncertainty or clutter of current posterior probability distribution, which can be taken directly as .

AIGGI is entirely agent-dominated and the expected gain of candidate phenotypes is a composite gain of answering “*Yes*” or “*No*”. To account for user intention, we also propose two intention-considered AIGGI-based indicators. The first indicator, AIGGI-P, is used when users expect to validate phenotypes (answer “*Yes*”). In this case, we minimize the gain from answering “*No*” by setting the conditional entropy and Gini index of the negative branch in Equations 6 and 8 to a constant value of 1. Another indicator, AIGGI-N, is used when users expect to exclude phenotypes (answer “*No*’). In this case, we minimize the gain of answering “*Yes*” by setting the conditional entropy and Gini index of the positive branch to 1.

In each turn of Q&A, our agent will select a phenotype with the highest expected gain to interrogate users. A response of “Yes” or “No” will prompt a re-diagnosis, updating the diagnostic status accordingly. If users respond with “Not sure”, the diagnostic status remains unchanged and our agent will interrogate the next-ranked phenotype. To enhance the efficiency of our dialogue strategy, previously reported and interrogated phenotypes will not be re-asked. Additionally, following the annotation propagation rule, ancestors of known existing phenotypes and descendants of excluded phenotypes will not be further considered by our agent in future rounds.

## Implementation of RDmaster

RDmaster is implemented as an online web tool (http://rdmaster.nbscn.org/), which includes a RESTful back-end server that provides data access as well as dialogue and diagnostic services, developed using the Spring (<https://spring.io/>) framework, a front-end server developed using the React (<https://reactjs.org/>) framework, and a web-based UI that enables human-computer interaction. To reduce system response time and perform real-time diagnosis, the Spring context, storing preloaded and pre-calculated data, and the Caffeine plugin (<https://github.com/ben-manes/caffeine>), which caches current diagnostic results, are used on our back-end server. Two open-source JavaScript visualization libraries, D3 (<https://d3js.org/>) and ECharts (<https://echarts.apache.org/>) provide essential visualized decision-support information on UI [53, 54].

## Evaluation

The evaluation of RDmaster consists of two main tasks: (i) a qualitative assessment of its dialogue strategy; (ii) a quantitative analysis of before and after dialogue diagnostic performance of prioritizing target diseases and disease-causing genes.

To evaluate our dialogue strategy, we simulated tens of thousands of patients according to HPO frequency annotations from phenotype.hpoa file and randomly assigned inaccurate phenotypes (ancestor terms of annotated phenotypes) and noise to each patient. The phenotype information of each simulated patient was reported to our system to compute a candidate questioning phenotype list, and then we analyzed and quantified the relationship between candidate phenotypes and top-ranked diseases within different diagnostic states.

We further evaluated RDmaster’s diagnostic performance using real diagnosed patients published online. A total of 238 case reports involving 92 RDs were collected, of which 225 cases involving 86 RDs had documented causative genes in KB and were able to perform genetic diagnosis. For each case report, we endeavored to capture all recorded phenotypic features (present or absent) and ultimately generated standardized patient information within the Phenopacket schema [55]. See Supplementary Table 1 for the detailed information of all 238 published cases. Like many previous studies [56-58], we simulated exomes for the 86 RDs with causative genes by randomly adding known pathogenic variants from the Clinvar database (homozygous for recessive diseases, heterozygotes for dominant) to our internally randomly selected 86 exomes (see Supplementary Table 2). In this evaluation, several influential KB-based RD diagnostic tools (Phenomizer, BOQA, PhenoPro, and LIRICAL for phenotype-driven diagnosis; Exomiser for disease-causing genes prioritization) were introduced to compare diagnostic performance with RDmaster. In addition, we also explored the effectiveness of large language models, specifically GPT-3.5 and GPT-4, in RD differential diagnosis.

# Reference

1. Rare diseases and orphan drugs (2006)
2. The burden of rare diseases (2019)
3. Priority Diseases and Reasons for Inclusion (2013)
4. Progress, challenges and global approaches to rare diseases (2021)
5. RDmap: a map for exploring rare diseases (2021)
6. Rare diseases in general practice: recognising the zebras among the horses (2016)
7. Differential diagnosis: the key to reducing diagnosis error, measuring diagnosis and a mechanism to reduce healthcare costs (2014)
8. The key role of differential diagnosis in diagnosis (2017)
9. Should electronic differential diagnosis support be used early or late in the diagnostic process? A multicentre experimental study of Isabel (2022)
10. Online Mendelian Inheritance in Man (OMIM) (2000)
11. Orphanet: a European database for rare diseases (2008)
12. The Human Phenotype Ontology: A Tool for Annotating and Analyzing Human Hereditary Disease (2008)
13. The Human Phenotype Ontology in 2021 (2021)
14. Clinical Diagnostics in Human Genetics with Semantic Similarity Searches in Ontologies (2009)
15. Bayesian ontology querying for accurate and noise-tolerant semantic searches (2012)
16. eXtasy: Variant prioritization by genomic data fusion (2013)
17. Phen-gen: Combining phenotype and genotype to analyze rare disorders (2014)
18. Next-generation diagnostics and disease-gene discovery with the Exomiser (2015)
19. Phenolyzer: Phenotype-based prioritization of candidate genes for human diseases (2015)
20. PhenoPro: A novel toolkit for assisting in the diagnosis of Mendelian disease (2019)
21. Differential Diagnosis Generators: an Evaluation of Currently Available Computer Programs (2012)
22. Artificial intelligence in medical diagnosis: methods, algorithms and applications (2020)
23. Attention is all you need (2017)
24. Language models are unsupervised multitask learners (2019)
25. Language models are few-shot learners (2020)
26. The Diagnostic and Triage Accuracy of the GPT-3 Artificial Intelligence Model (2023)
27. Diagnostic Accuracy of Differential-Diagnosis Lists Generated by Generative Pretrained Transformer 3 Chatbot for Clinical Vignettes with Common Chief Complaints: A Pilot Study (2023)
28. Task-oriented Dialogue System for Automatic Diagnosis (2018)
29. End-to-End Knowledge-Routed Relational Dialogue System for Automatic Diagnosis (2019)
30. Task-oriented Dialogue System for Automatic Disease Diagnosis via Hierarchical Reinforcement Learning (2020)
31. Refining clinical diagnosis with likelihood ratios (2005)
32. PheLR
33. Interpretable Clinical Genomics with a Likelihood Ratio Paradigm (2020)
34. Induction of decision trees (1986)
35. A Bayesian CART algorithm (1998)
36. The variant call format and VCFtools (2011)
37. The Monarch Initiative: an integrative data and analytic platform connecting phenotypes to genotypes across species (2017)
38. Expansion of the genotypic and phenotypic spectrum in patients with KRAS germline mutations (2007)
39. Comparison of causative variant prioritization tools using next-generation sequencing data in Japanese patients with Mendelian disorders (2019)
40. Diagnosis support systems for rare diseases: A scoping review (2020)
41. Theoretical comparison between the gini index and information gain criteria (2004)
42. Evaluating the impact of GINI index and information gain on classification using decision tree classifier algorithm (2020)
43. Investigation of a joint splitting criteria for decision tree classifier use of information gain and gini index (2018)
44. GenomeDiver: a platform for phenotype-guided medical genomic diagnosis (2021)
45. A Visual Phenotype-Based Differential Diagnosis Process for Rare Diseases (2022)
46. Conversational AI Models for ophthalmic diagnosis: Comparison of ChatGPT and the Isabel Pro Differential Diagnosis Generator (2023)
47. Isabel Pro (2019)
48. International Cooperation to Enable the Diagnosis of All Rare Genetic Diseases (2017)
49. Jannovar: a java library for exome annotation (2014)
50. ClinVar: public archive of interpretations of clinically relevant variants (2016)
51. Interactive data visualization for the web: an introduction to designing with D3 (2017)
52. ECharts: A declarative framework for rapid construction of web-based visualization (2018)
53. The GA4GH Phenopacket schema defines a computable representation of clinical data (2022)
54. Effective diagnosis of genetic disease by computational phenotype analysis of the disease-associated genome (2014)
55. Phenotype-driven strategies for exome prioritization of human Mendelian disease genes (2015)
56. Xrare: a machine learning method jointly modeling phenotypes and genetic evidence for rare disease diagnosis (2019)

# Figure Legends

**Fig. 1: Workflow overview and system architecture of RDmaster.**

**Fig. 2: Web-based user interface screenshots of RDmaster.** **a** User interface for clinical information collection. **b** User interface for dialogue diagnosis.

**Fig. 3: Brief statistics on RDmaster’s dialogue strategy.** Our proposed three dialogue strategies are marked in the top right corner of each legend. All these graphs are averages of test results from tens of thousands of simulated cases. **a** Average association scores of top 5 diseases with the interrogated phenotype under different diagnostic states. The horizontal axis from 1 to 0 represents a decrease in diagnostic uncertainty, i.e., the Gini index of the diagnostic status tends to 1 when there are multiple competing candidate diagnoses, and tends to 0 when there is only one obvious candidate diagnosis. **b** Average association scores of top 100 candidate phenotypes with top 5 diseases.

**Fig. 4: Diagnostic performance of RDmaster.** DS\_T0 represents the diagnostic effect of RDmaster before the dialogue, DS\_T10 represents after 10 turns of dialogue, DS\_T20 is after 20 turns and so on. **a** Ranking distribution of target diseases in phenotype-driven diagnosis. Note that there are two vertical coordinates, the left-hand one corresponding to the blue box plot and the right-hand one to the grey. **b** Cumulative number of target diseases ranked in the top 10 in phenotype-driven diagnosis. **c** Distribution of disease-causing gene rankings in phenotype-genotype diagnosis.

# Tables

**Table 1** **Summary of collected published cases for evaluation.**

|  |  |
| --- | --- |
| Total published case reports | 238 |
| **Patient** | |
| Male | 113 |
| Female | 119 |
| Unrecorded gender | 6 |
| **Rare diseases** | |
| Total disease number | 92 |
| Mean-median-maximum # cases per disease | 2.6-1-18 |
| Total cases | 238 |
| **Rare genetic diseases** | |
| Total disease number | 86 |
| Autosomal-dominant diseases | 34 |
| Autosomal-recessive diseases | 42 |
| X-linked dominant diseases | 1 |
| X-linked recessive diseases | 3 |
| Disease-causing genes | 86 |
| Mean-median-maximum # cases per disease | 2.6-1-18 |
| Total cases | 225 |
| **HPO terms** | |
| Total recorded HPO terms over all cases | 811 |
| Mean # HPO terms per case | 9.1 |
| Mean # excluded HPO terms per case | 1.0 |

**Table 2 Statistic results of target disease rankings of collected patients in differential diagnosis lists generated by KB-based tools and LLMs.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Methods | Target disease in | | | |
| Top 1 | Top 3 | Top 5 | Top 10 |
| Isolated Approaches: | | | | |
| GPT-4 | 33.61% | 41.60% | 44.54% | 46.22% |
| GPT-3.5 | 24.79% | 31.09% | 35.29% | 41.18% |
| RDmaster | **41.60%** | **51.26%** | **58.40%** | **68.49%** |
| LIRICAL | 37.39% | 51.26% | 57.56% | 66.81% |
| PhenoPro | 34.45% | 44.96% | 53.78% | 65.55% |
| BOQA | 32.35% | 44.96% | 50.42% | 58.82% |
| Phenomizer | 26.05% | 36.97% | 43.70% | 55.04% |
| Joint Approaches (min\_target\_disease\_rank (A, B)): | | | | |
| RDmaster & GPT-4 | **57.56%** | **67.23%** | **72.69%** | **77.31%** |
| RDmaster & GPT-3.5 | 51.68% | 60.92% | 68.49% | 76.47% |
| RDmaster & LIRICAL | 43.70% | 56.30% | 63.03% | 72.27% |
| RDmaster & PhenoPro | 44.54% | 54.20% | 61.34% | 72.69% |
| RDmaster & BOQA | 45.80% | 56.72% | 61.76% | 72.69% |
| RDmaster & Phenomizer | 44.96% | 54.20% | 61.76% | 72.27% |
| LLMs, considered the potentially correct recommendation, which is a broader concept of the target disease | | | | |
| GPT-4 | 38.66% | 48.32% | 53.78% | 56.30% |
| GPT-3.5 | 29.41% | 36.97% | 43.70% | 50.42% |