Original Paper

# AI-Driven Diagnostic Assistance in Medical Inquiry: Reinforcement Learning Algorithm Development and Validation

Xuan Zou[1*], MM; Weijie He[2,3,4*], PhD; Yu Huang[5*], ME; Yi Ouyang[5], PhD; Zhen Zhang[1], MM; Yu Wu[1], PhD; Yongsheng Wu[1], BMed; Lili Feng[6], PhD; Sheng Wu[6], BMed; Mengqi Yang[7], BE; Xuyan Chen[6], PhD; Yefeng Zheng[5], PhD; Rui Jiang[4,8], PhD; Ting Chen[2,3,4], PhD

[1]Shenzhen Center for Disease Control and Prevention, Shenzhen, China

[2]Department of Computer Science and Technology, Tsinghua University, Beijing, China

[3]Institute of Artificial Intelligence, Tsinghua University, Beijing, China

[4]Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing, China

[5]Jarvis Research Center, Tencent YouTu Lab, Shenzhen, China

[6]Beijing Tsinghua Changgung Hospital, School of Clinical Medicine, Tsinghua University, Beijing, China

[7]Tencent Healthcare, Shenzhen, China

[8]Department of Automation, Tsinghua University, Beijing, China

[*]these authors contributed equally

**Corresponding Author:**
Ting Chen, PhD
Department of Computer Science and Technology
Tsinghua University
Room 3-609, Future Internet Technology Research Center
Tsinghua University
Beijing, 100084
China
Phone: 86 010 62797101
Email: tingchen@tsinghua.edu.cn

## Abstract

**Background:** For medical diagnosis, clinicians typically begin with a patient's chief concerns, followed by questions about symptoms and medical history, physical examinations, and requests for necessary auxiliary examinations to gather comprehensive medical information. This complex medical investigation process has yet to be modeled by existing artificial intelligence (AI) methodologies.

**Objective:** The aim of this study was to develop an AI-driven medical inquiry assistant for clinical diagnosis that provides inquiry recommendations by simulating clinicians' medical investigating logic via reinforcement learning.

**Methods:** We compiled multicenter, deidentified outpatient electronic health records from 76 hospitals in Shenzhen, China, spanning the period from July to November 2021. These records consisted of both unstructured textual information and structured laboratory test results. We first performed feature extraction and standardization using natural language processing techniques and then used a reinforcement learning actor-critic framework to explore the rational and effective inquiry logic. To align the inquiry process with actual clinical practice, we segmented the inquiry into 4 stages: inquiring about symptoms and medical history, conducting physical examinations, requesting auxiliary examinations, and terminating the inquiry with a diagnosis. External validation was conducted to validate the inquiry logic of the AI model.

**Results:** This study focused on 2 retrospective inquiry-and-diagnosis tasks in the emergency and pediatrics departments. The emergency departments provided records of 339,020 consultations including mainly children (median age 5.2, IQR 2.6-26.1 years) with various types of upper respiratory tract infections (250,638/339,020, 73.93%). The pediatrics department provided records of 561,659 consultations, mainly of children (median age 3.8, IQR 2.0-5.7 years) with various types of upper respiratory tract infections (498,408/561,659, 88.73%). When conducting its own inquiries in both scenarios, the AI model demonstrated high diagnostic performance, with areas under the receiver operating characteristic curve of 0.955 (95% CI 0.953-0.956) and 0.943 (95% CI 0.941-0.944), respectively. When the AI model was used in a simulated collaboration with physicians, it notably reduced the average number of physicians' inquiries to 46% (6.037/13.26; 95% CI 6.009-6.064) and 43% (6.245/14.364; 95%

CI 6.225-6.269) while achieving areas under the receiver operating characteristic curve of 0.972 (95% CI 0.970-0.973) and 0.968 (95% CI 0.967-0.969) in the scenarios. External validation revealed a normalized Kendall τ distance of 0.323 (95% CI 0.301-0.346), indicating the inquiry consistency of the AI model with physicians.

**Conclusions:** This retrospective analysis of predominantly respiratory pediatric presentations in emergency and pediatrics departments demonstrated that an AI-driven diagnostic assistant had high diagnostic performance both in stand-alone use and in simulated collaboration with clinicians. Its investigation process was found to be consistent with the clinicians' medical investigation logic. These findings highlight the diagnostic assistant's promise in assisting the decision-making processes of health care professionals.

***(J Med Internet Res 2024;26:e54616)*** doi: 10.2196/54616

**KEYWORDS**

inquiry and diagnosis; electronic health record; reinforcement learning; natural language processing; artificial intelligence

## Introduction

### Background

The growing demand for intelligent clinical decision support systems (CDSSs) has become increasingly evident in the health care landscape today [1,2]. The surge in demand can be attributed to advances in medical knowledge, the accumulation of health care data, and the rapid progression of artificial intelligence (AI) and machine learning technologies. Traditionally, health care professionals have relied heavily on their individual experiences and medical expertise to make clinical decisions, which are often susceptible to subjective biases and information gaps [3,4]. Furthermore, the world faces challenges related to insufficient medical resources, particularly in regions where a shortage of health care professionals impedes timely access to care. Consequently, intelligent CDSSs have been developed with the potential to enhance patient care quality, reduce medical costs, and minimize diagnostic errors. These systems were developed by analyzing and training data from electronic health records (EHRs) and medical literature [5,6] using the power of extensive data analysis, machine learning algorithms, and natural language processing techniques.

Our research is focused on advancing CDSSs with a primary emphasis on aiding the diagnostic process before the diagnosis is made, specifically during the information-gathering phase. Existing diagnostic support systems often rely on comprehensive patient information, such as EHRs and medical images [7-13], to provide diagnostic predictions only at the final step of the medical investigation process. Thus, the vital need for decision support during the intermediate steps of the diagnostic process is largely overlooked. The clinical diagnostic process is a complex procedure that involves a series of inquiries and examinations. In this dynamic process, health care professionals continuously gather information, adjust hypotheses, and refine diagnostic reasoning until the optimal diagnosis and treatment are determined. The process is not merely a sequence of isolated steps but a dynamic and evolving interaction between the clinician and the patient. Conventional diagnostic support systems tend to provide little to no guidance during these intermediate steps, falling short of providing meaningful support when it is most needed.

A variety of automatic disease diagnosis techniques have emerged recently as a result of the advancement of AI with the goal of assisting in the middle stages of clinical decision-making [14-21]. However, a common limitation among these methods is their exclusive focus on online disease diagnosis. Online health care platforms have undoubtedly revolutionized medical consultations by facilitating remote access to health care professionals. Patients can seek advice and preliminary assessments for various ailments without visiting a health care facility physically. Unfortunately, there are inherent constraints to these platforms. Online consultations are primarily reliant on textual descriptions, making it challenging to gather essential information that requires palpation, auscultation, or specialized laboratory investigations, which collectively form the foundation of a comprehensive and accurate diagnosis.

Large language models (LLMs) such as ChatGPT [22] have demonstrated remarkable proficiency in following instructions and generating humanlike responses across various domains. LLMs have already attracted a lot of attention regarding their potential applications in health care settings, which include facilitating clinical documentation, summarizing research papers, assisting with medical education [23], or acting as a chatbot to respond to questions from patients about their specific concerns [24,25]. While LLMs have access to extremely large corpora containing medical dialogues, knowledge bases, and the literature, there is a significant gap in access to large-scale real patient records [26]. This lack of actual clinical data significantly impacts their ability to provide tailored inquiry recommendations for patients in varying clinical scenarios. In addition, ethical [24,25] and security [27] concerns further complicate the incorporation of sensitive health records into the training of LLMs.

### Objectives

To address the gap in diagnostic decision support, we present MedRIA, a medical reinforcement learning inquiry assistant. MedRIA is designed to facilitate the medical investigation process by intelligently guiding inquiries; symptom assessments; examination recommendations; and, in particular, the transitions between them. To the best of our knowledge, no previous system has successfully managed the entire diagnostic process. MedRIA learns its inquiry capabilities from a large number of processed EHRs. MedRIA uses a reinforcement learning actor-critic framework [28] to explore rational and effective inquiry logic. Following the standard inquiry process of physicians [29], we segmented the inquiry into 4 stages: inquiring about symptoms and medical history, conducting physical examinations (PEs),

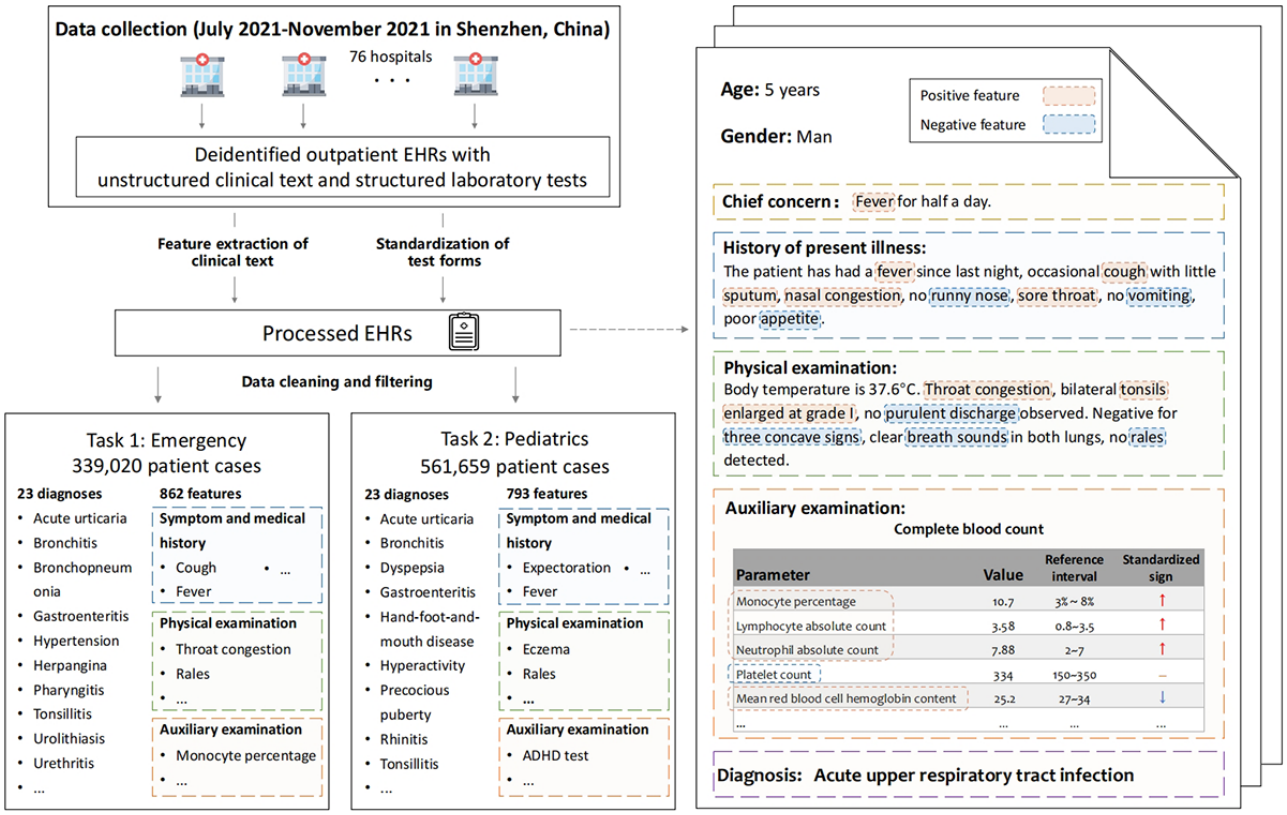
requesting auxiliary examinations (AEs), and terminating the inquiry with a diagnosis. MedRIA's ability to tailor inquiry strategies according to patient-specific conditions positions it as a promising solution for regions or populations facing constraints in medical resources.

## Methods

### Processed Outpatient EHRs

The Shenzhen Center for Disease Control and Prevention in Guangdong Province, China, collected deidentified outpatient EHRs from 76 hospitals in Shenzhen. We selected records of the emergency and pediatrics departments from July 2021 to November 2021. As depicted in Figure 1, these records consist of both unstructured textual information and structured laboratory test results, which underwent feature extraction and standardization using natural language processing techniques. An example of a processed EHR is provided on the right side of Figure 1.

**Figure 1.** Schematic illustration of the data collection, data processing, and task building process for the development of MedRIA. ADHD: attention-deficit/hyperactivity disorder; EHR: electronic health record.




By combining established medical terminologies with domain-specific knowledge from health care professionals, we constructed a comprehensive feature set for the feature extraction process. Initially, we used existing Chinese versions of general medical terminologies, including the *International Classification of Diseases, 10th Revision*, and Systematized Nomenclature of Medicine–Clinical Terms. In addition, we incorporated domain expertise by consulting with health care experts to identify relevant clinical concepts and terminology to the context of our study. The feature sets of the emergency and pediatrics EHRs are presented in Multimedia Appendices 1 and 2, respectively.

As depicted in Figure 1, after extracting features from clinical text and standardizing test forms, each record was converted into a feature set, where each feature was categorized into 1 of 3 types based on its source—symptom and medical history (SMH), PE, and AE—and assigned a value based on its data type. Symptom features (eg, *cough*) and some features observed from examinations (eg, *throat congestion* and *proteinuria*) were assigned binary values indicating whether the patient exhibited that feature. Numerical laboratory test features with specific reference ranges, such as *platelet count*, were categorized as high, low, or within the reference range. The values of descriptive features depend on their meanings. For example, *urine color* was given values such as *yellow*, *colorless*, *white*, *red*, and so on.

In our study, we used the Medical Bidirectional Encoder Representations From Transformers (MedBERT) model for feature extraction on clinical text, which was based on Bidirectional Encoder Representations From Transformers [30]. Using a vast corpus of Chinese clinical text data, including medical textbooks, online consultations, journal article abstracts, and deidentified EHRs, we pretrained the Bidirectional Encoder Representations From Transformers model to obtain MedBERT. We manually constructed 2 labeled data sets to fine-tune MedBERT to perform medical named entity recognition and entity normalization. In our pipeline, we first used fine-tuned MedBERT for medical named entity recognition on clinical

text. Next, we proceeded to normalize the identified entities to our predefined feature set using the fine-tuned MedBERT for entity normalization. Finally, we used a string-matching negation detection mechanism to determine the values of binary features based on the contents of the text. For structured laboratory test results, we used a string-matching algorithm to standardize the test items to our predefined feature set. The feature values could be directly obtained from the structured results.

Following the feature extraction process, we conducted data cleaning and filtering. Details are presented in Multimedia Appendix 3 [31-37]. Finally, we filtered the data sets by department, creating 2 retrospective inquiry-and-diagnosis evaluation tasks for emergency and pediatrics. Emergency and pediatrics are the 2 busiest departments in hospitals and often put tremendous pressure on health care personnel. These departments require effective decision support tools to assist health care professionals in managing the workload efficiently while providing timely and accurate patient care.

## Reinforcement Learning Formulation

MedRIA takes as initial input the patient's basic demographic information, including gender and age, along with features extracted from the patient's chief concerns. Subsequently, MedRIA provides recommendations for the physician to inquire about specific features. As the physician conducts the inquiry and gathers patient information, MedRIA updates the patient's status and continues to output the next relevant feature or suggests terminating the inquiry and providing a predicted diagnosis in terms of the probability distribution over potential disease diagnoses. This can be described as a Markov decision process [38], represented by the tuple $M(S, A, T, R, \gamma)$. Here, $S$ (state) is a set of vectors representing the inquiry states that incorporate both observed and unobserved features so far, denoted as $x_O$ and $x_U$, respectively. $A$ (action) involves the actions available to MedRIA, including recommending specific features or suggesting the termination of the inquiry process. The goal is to determine the optimal inquiry strategy composed of a series of actions, denoted as $\pi_\theta$. At each step t, the policy $\pi_\theta(a_t|s_t)$ determines the action $a_t$ based on the current state $s_t$. The transition function $T$ represents the probability distribution of the next state $s_{t+1}$ given the current state $s_t$ and action $a_t$, denoted as $p(s_{t+1}|s_t, a_t)$. $R$ is a reward function used to assess the benefit of the transition $(s_t, a_t, s_{t+1})$. $\gamma$ is the discount factor for accumulating rewards at each step. The state-value function that reinforcement learning maximizes is the expected sum of discounted rewards given the policy $\pi_\theta$ and the state $s_t$.

## Actor-Critic Framework

MedRIA uses the classic reinforcement learning actor-critic framework [28] involving 2 neural networks: the actor and the critic. The actor network determines the next action to take, whereas the critic network is responsible for providing feedback to the actor, evaluating the quality of actions and suggesting adjustments. On the basis of our previous work [39], we used a pretrained variational autoencoder (VAE) [31] as the backbone of the actor network. We used a supervised diagnostic prediction model, denoted as $D$, to provide the current disease prediction probability distribution $D(s_t)$ based on the state $s_t$ to help decide on each inquiry action.

The actor of MedRIA makes decisions for the next action based on the observed features $x_O$. It is intuitive to prioritize asking about feature x that has a high conditional probability $p(x|x_O)$. We leveraged the VAE to incorporate the conditional probability distribution between observed features $x_O$ and unobserved features $x_U$. A VAE defines a generative model of the form $p(x, z)=\prod_i p_\Theta(x_i|z)p(z)$, where features x are generated from latent variables z. The number of dimensions of z is 64. p(z) represents a prior, often a spherical Gaussian distribution. $p_\Theta(x|z)$ is represented by a 4-layer multilayer perceptron (MLP) decoder with parameters $\Theta$. The VAE uses another neural network encoder with parameters $\Phi$ to generate the variational approximation of the posterior $q_\Phi(z|x)$. To obtain $p(x_U|x_O)$, we sampled from the VAE encoder as $\sim q_\Phi(z|x)$ and then sampled $x_U$ from the VAE decoder as $p_\Theta(x_U|\ )$. Let $e_i$ represent the embedding vector for the ith observed feature $x_i$, and let $c_i=[x_i, e_i]$ denote the concatenated input carrying information for $x_i$. The number of dimensions of $e_i$ is 64. Then, we used a 4-layer MLP to map the input $c_i$ to a Gaussian distribution in latent space, with mean vector $\mu_i$ and variance vector $V_i$. To address arbitrary partial observations of features during the inquiry process, we used a product-of-experts mechanism [40] to calculate the approximate posterior. The VAE was incorporated within the actor network to leverage the conditional probability distribution $p(x_U|x_O)$. To obtain the action $a_t$, $x_O$ were first fed into the nested VAE, yielding decoded features that contain predictive information regarding $x_U$. The decoded features were then concatenated with the current diagnostic confidence $D(s_t)$ and the representation of the current timestep $n_t$. Here, $n_t=t / N_T$ represents the ratio of the number of completed inquiries t over a predefined maximum action count $N_T$. $N_T$ was set to 21 and 26 for the emergency and pediatrics tasks, respectively. Finally, a 2-layer MLP with softmax activation was used to map the concatenated vector to the action space.

The objective of the critic is to estimate the state-value function to optimize the policy $\pi_\theta$. Similar to the approach used in the actor, we concatenated the current state $s_t$ with $D(s_t)$ and $n_t$ to form an input vector to aid in estimation. In addition, we obtained the informative latent variable $z_t$ by feeding observed features $x_O$ into the VAE encoder and appended $z_t$ to the input vector. Ultimately, a 5-layer MLP was used to map the input vector to predict the state value.

## Reward Shaping

The reward function for evaluating the gain in state transition plays a crucial role in the reinforcement learning process. We start by defining the short-term reward function $R_{short}$ when the action is to inquire about a specific feature x. For MedRIA, the inquiry states can be categorized into 4 ordered stages: SMH ($S_{SMH}$), PE ($S_{PE}$), AE ($S_{AE}$), and termination with diagnosis ($S_{TD}$). During the inquiry process, there can be multiple states that fall under each stage. For example, the initial state of

MedRIA falls under $S_{SMH}$, followed by a sequence of states that fall under $S_{PE}$, and so on. Correspondingly, all features are categorized according to the 3 inquiry stages $S_{SMH}$, $S_{PE}$, and $S_{AE}$. When MedRIA selects a feature that falls under a particular inquiry stage, it signifies a transition from the inquiry stage to that stage. We ensure that MedRIA's inquiry process adheres to the sequential nature of clinical inquiries, where features from previous stages are not considered for selection once the inquiry progresses to a subsequent stage. This is achieved by manually setting the probabilities of actions associated with features from previous stages to 0 in the action probability distribution of the actor.

To measure the consistency of MedRIA's inquiries with those made by physicians, we introduce 3 functions: $M_{SMH}(s_t)$, $M_{PE}(s_t)$, and $M_{AE}(s_t)$. These functions quantify the number of features that physicians have inquired about in the current state $s_t$ but that MedRIA has not investigated. Specifically, $M_{SMH}(s_t)$, $M_{PE}(s_t)$, and $M_{AE}(s_t)$ quantify the number of SMH features, PE features, and AE features, respectively. When calculating $M_{AE}(s_t)$, all features within a laboratory test are considered as a unit. Essentially, $M_{AE}(s_t)$ calculates the number of AEs that physicians have recommended in the current state $s_t$ but that MedRIA has omitted. Correspondingly, when MedRIA selects a feature associated with an AE, it indicates that MedRIA has recommended the entire examination.

To assess the diagnostic quality of inquiries, we define a function $Diff(s_t, x, s_{t+1})$ to estimate the differential effect of querying the feature x in the transition from state $s_t$ to state $s_{t+1}$. If x belongs to the extracted features, meaning that the physician also inquired about that feature from the patient, then $Diff(s_t, x, s_{t+1})$ is defined as $Diff(s_t, x, s_{t+1})=|D_{KL}[y||D(s_t)]-D_{KL}[y||D(s_{t+1})]|$, where $D_{KL}$ is the Kullback-Leibler divergence between 2 distributions and y is a one-hot vector indicating the final diagnosis made by the physician. The absolute difference between these 2 Kullback-Leibler divergences represents the degree to which the feature x influences the diagnosis results, with the final diagnosis made by the physician as the reference. If x does not belong to the extracted features, then $Diff(s_t, x, s_{t+1})$ is simply set to 0. Subsequently, we define the short-term reward function $R_{short}(s_t, x, s_{t+1})$ as follows: when $s_t$ belongs to $S_{SMH}$ and $s_{t+1}$ belongs to $S_{PE}$, $R_{short}(s_t, x, s_{t+1})=Diff(s_t, x, s_{t+1})+\alpha \cdot I_{phy}(x)-m \cdot M_{SMH}(s_t)$; when $s_t$ belongs to $S_{SMH}$ and $s_{t+1}$ belongs to $S_{AE}$, $R_{short}(s_t, x, s_{t+1})=Diff(s_t, x, s_{t+1})+\alpha \cdot I_{phy}(x)-m \cdot (M_{SMH}(s_t)+M_{PE}(s_t))$; when $s_t$ belongs to $S_{PE}$ and $s_{t+1}$ belongs to $S_{AE}$, $R_{short}(s_t, x, s_{t+1})=Diff(s_t, x, s_{t+1})+\alpha \cdot I_{phy}(x)-m \cdot M_{AE}(s_t)$; otherwise, $R_{short}(s_t, x, s_{t+1})=Diff(s_t, x, s_{t+1})+\alpha \cdot I_{phy}(x)$. $I_{phy}(x)$ is an indicator representing whether feature x belongs to the extracted features. If yes, it takes the value of 1; otherwise, it is 0. $\alpha$ and m are small constant factors to balance terms in the expression. We set $\alpha$ to 0.2. m was set to 0.2 in the emergency task and 0.3 in the pediatrics task. In $R_{short}$, we penalize each hasty stage transition based on the number of features that the actual physician inquired about from the patient in the previous stage but were not inquired about by MedRIA.

Next, we elaborate on the definition of the long-term reward $R_{long}$ for the action of terminating the inquiry. When MedRIA selects the action to stop the inquiry process, if the disease predicted by the diagnostic model *D* matches the physician's diagnosis, $R_{long}$ is set to 2 in the emergency task and 3 in the pediatrics task. If it does not match, similar to the penalty of inquiry consistency in $R_{short}$, we define $R_{long}(s_t, x, s_{t+1})$ as follows: when $s_t$ belongs to $S_{SMH}$, $R_{long}(s_t, x, s_{t+1})=-m \cdot (M_{SMH}(s_t)+M_{PE}(s_t)+M_{AE}(s_t))$; when $s_t$ belongs to $S_{PE}$, $R_{long}(s_t, x, s_{t+1})=-m \cdot (M_{PE}(s_t)+M_{AE}(s_t))$; when $s_t$ belongs to $S_{AE}$, $R_{long}(s_t, x, s_{t+1})=-m \cdot M_{AE}(s_t)$.

## Training Details

During the training process of MedRIA, we first trained a VAE. In particular, we simulated the arbitrary partial observations during the inquiry process by randomly dropping a portion of input features. Next, we trained a diagnostic prediction model *D* based on a 5-layer MLP using the softmax function and cross-entropy loss function. Then, we initialized the parameters of the nested VAE in the actor network of MedRIA with the parameters of the trained VAE. Only the nested VAE decoder was fine-tuned in the follow-up training. We used the proximal policy optimization [32] algorithm to train both the actor and critic networks. To make the predicted diagnostic probability distribution more accurate when dealing with the partially observed features generated by MedRIA, we collected the observed features $x_O$ when the actor chose to terminate the inquiry process during each training epoch. At the end of each epoch, collected data were used to fine-tune *D* to better adapt to MedRIA's inquiry patterns. More details on implementation are presented in Multimedia Appendix 3.

## External Validation

We used data from the medical dialogue corpus, IMCS-21 [41], to assess MedRIA's inquiry logic. IMCS-21 contains 4116 online medical consultations between physicians and patients covering 10 pediatric diseases. After data filtering, we selected 950 consultations with clear diagnoses that overlapped with 4 diseases covered in our pediatrics task: bronchitis (286 consultations), acute upper respiratory tract infection (392 consultations), indigestion (233 consultations), and bronchopneumonia (39 consultations). This data set served as an external test set for the trained pediatric MedRIA. We constructed ordered inquiry sequences based on the dialogue contents. These inquiry sequences reflected the typical order of feature queries in the diagnostic process, and they were used to evaluate MedRIA's inquiry sequences. If MedRIA's inquiry sequences are similar to those of physicians, we can conclude that MedRIA's inquiry logic is similar to that of physicians. The similarity is measured by the Kendall $\tau$ distance, which treats an inquiry sequence as a permutation of features and quantifies the pairwise disagreements between 2 permutations. In addition, we applied the normalized Kendall $\tau$ distance, which scales the Kendall $\tau$ distance by the total number of possible feature pairs.

### Statistical Analysis

Nonparametric bootstrap sampling was used to calculate a 95% CI. Specifically, we repeatedly drew 1000 bootstrap samples from the test set. Each bootstrap sample was obtained through random sampling with replacement and has the same size as the test set.

### Ethical Considerations

All EHR data used in this study were retrospectively collected from EHR systems sourced from routine clinical practice. To ensure patient privacy and confidentiality, all data were deidentified. This study was approved by the ethics committees of the Shenzhen Center for Disease Control and Prevention (SZCDC-IRB2024055). As this study involved no direct patient intervention and was retrospective in nature, individual informed consent was waived.

## Results

### Data Characteristics

We focused on 2 retrospective inquiry-and-diagnosis evaluation tasks in the emergency and pediatrics departments. The emergency departments provided 339,020 records, and the pediatrics departments provided 561,659 records. We randomly split the data into 3 parts: 75% for training MedRIA, 10% for validation, and 15% for testing.

The emergency EHRs contained 862 features, including 42.8% (369/862) SMH features, 21.2% (183/862) PE features, and 36% (310/862) AE features. There were 18 diagnosed diseases. As some records had 2 diagnosed diseases, we classified each combination as a separate disease category, bringing the total number of distinct diagnoses to 23. Similarly, the pediatrics EHRs covered 793 features, including 44.3% (351/793) SMH features, 22.6% (179/793) PE features, and 33.2% (263/793) AE features. There were 18 diagnosed diseases. When combinations of diagnosed diseases were considered, the total number of different diagnoses increased to 23. The detailed data characteristics are presented in Tables 1 and 2.

**Table 1.** Characteristics of the data in the emergency task (N=339,020).

| Characteristics | Values |
|---|---|
| **Gender, n (%)** | |
| Male | 187,992 (55.5) |
| Female | 151,028 (44.5) |
| Age (years), median (IQR) | 5.2 (2.6-26.1) |
| **Diagnosis, n (%)** | |
| AURTI[a] | 93,594 (27.6) |
| Bronchitis | 93,564 (27.6) |
| Bronchitis with rhinitis | 2439 (0.7) |
| AURTI with bronchitis | 4923 (1.5) |
| Bronchopneumonia | 3833 (1.1) |
| Laryngopharyngitis | 6095 (1.8) |
| Pharyngitis | 24,146 (7.1) |
| AURTI with pharyngitis | 2639 (0.8) |
| Pharyngitis with bronchitis | 3564 (1.1) |
| Herpangina | 7604 (2.2) |
| Tonsillitis | 10,286 (3) |
| AURTI with tonsillitis | 1784 (0.5) |
| Gastroenteritis | 30,385 (9) |
| Gastritis | 10,142 (3) |
| Enteritis | 2690 (0.8) |
| Acute appendicitis | 1957 (0.6) |
| Urinary tract stones | 13,884 (4.1) |
| Urethritis | 4052 (1.2) |
| Dermatitis | 6239 (1.8) |
| Acute urticaria | 5597 (1.7) |
| Hand-foot-and-mouth disease | 2990 (0.9) |
| Hypertension | 4397 (1.3) |
| Lumbar disc herniation | 2216 (0.7) |

[a]AURTI: acute upper respiratory tract infection.

**Table 2.** Characteristics of the data in the pediatrics task (N=561,659).

| Characteristics | Values |
|---|---|
| **Gender, n (%)** | |
| Male | 315,074 (56.1) |
| Female | 246,585 (43.9) |
| Age (years), median (IQR) | 3.8 (2.0-5.7) |
| **Diagnosis, n (%)** | |
| Bronchitis | 209,918 (37.4) |
| AURTI[a] | 155,181 (27.6) |
| AURTI with bronchitis | 11,734 (2.1) |
| Bronchopneumonia | 8119 (1.4) |
| Bronchitis with bronchopneumonia | 2565 (0.5) |
| Pharyngitis | 41,205 (7.3) |
| Pharyngitis with bronchitis | 5232 (0.9) |
| AURTI with pharyngitis | 2859 (0.5) |
| Laryngopharyngitis | 11,909 (2.1) |
| Tonsillitis | 19,790 (3.5) |
| Herpangina | 15,647 (2.8) |
| Rhinitis | 14,824 (2.6) |
| Bronchitis with rhinitis | 10,109 (1.8) |
| Gastroenteritis | 16,261 (2.9) |
| Gastritis | 5313 (0.9) |
| Enteritis | 3479 (0.6) |
| Indigestion | 8757 (1.6) |
| Hand-foot-and-mouth disease | 5156 (0.9) |
| Dermatitis | 3151 (0.6) |
| Acute urticaria | 2595 (0.5) |
| Developmental delay | 2953 (0.5) |
| Hyperactivity | 2538 (0.5) |
| Precocious puberty | 2364 (0.4) |

[a]AURTI: acute upper respiratory tract infection.

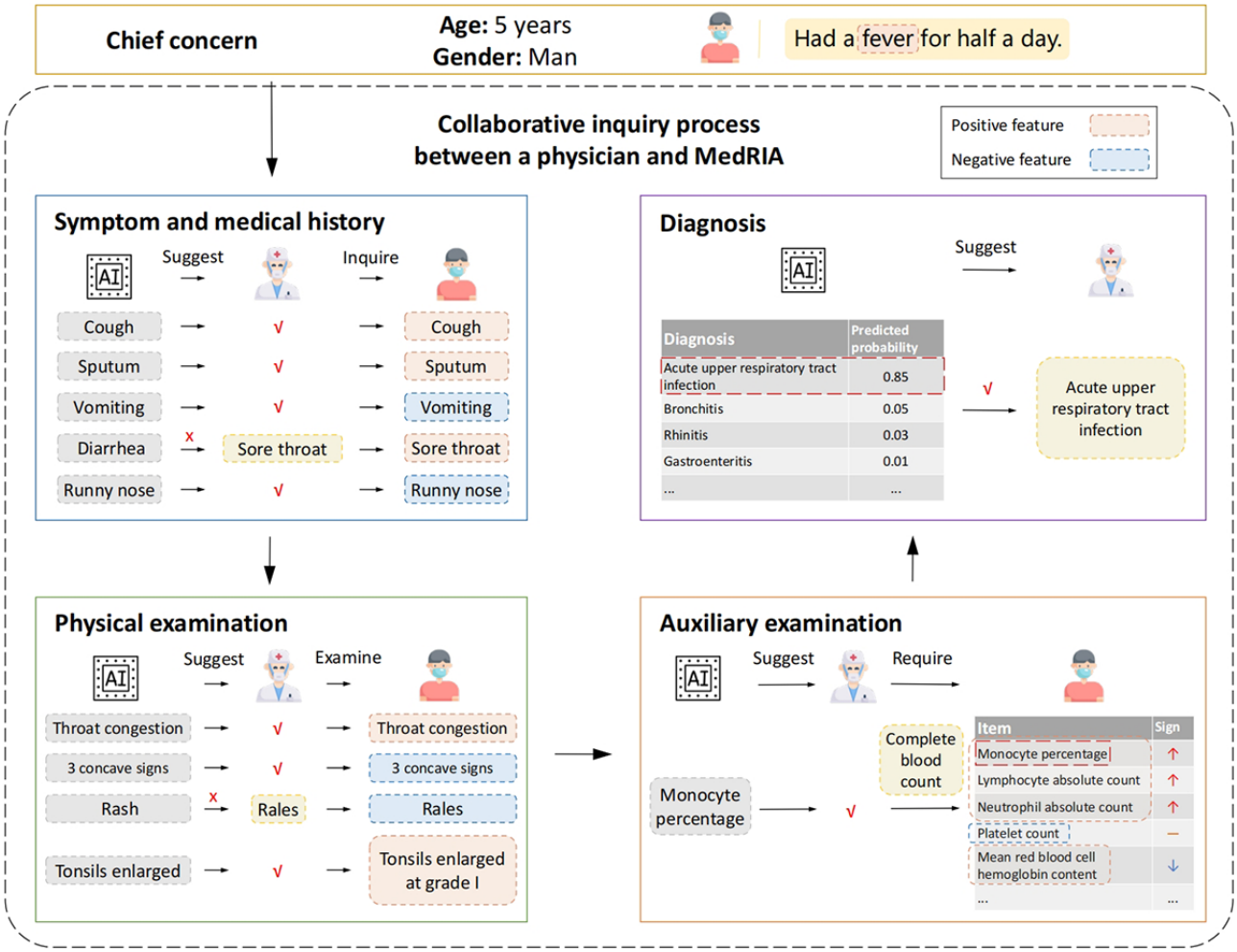
## Evaluation Criteria

We assessed the inquiry performance of MedRIA along 2 dimensions: the consistency of its inquiries with those of physicians and the contribution of its inquiries to diagnostic accuracy. It should be noted that physicians will inevitably make incorrect diagnoses, conduct inquiries based on individual previous experience, and record the EHRs in different ways. Therefore, we can establish a baseline diagnostic model by training a supervised classification model using the extracted feature sets as input and the physicians' diagnoses as output. This model simulated the physicians' diagnostic ability, and its performance approximates the actual performance of physicians. We aimed to mitigate the influence of individual physician variability and ensure a fair comparison between different inquiry methods based on the extracted feature sets. While the diagnoses in the records could serve as a gold standard, our simulated physician diagnostic model also serves as an important baseline. For training, several machine learning models were used, including logistic regression, random forest, support vector machine, MLP, and light gradient-boosting machine [42]. We chose light gradient-boosting machine for the emergency task and random forest for the pediatrics task based on their performance on the validation set.

To better evaluate the potential of MedRIA as a clinical assistant, we simulated a collaborative inquiry process between MedRIA and a physician. Specifically, MedRIA's actor network would generate probability distributions for potential inquiry features based on the currently collected medical information. These distributions suggest several high-probability features for the physician to inquire about. The physician has the flexibility to accept and incorporate some or all of these suggestions into the inquiries. In addition, the physician can

exercise their professional judgment to either ask additional questions or proceed to examination. In our simulated collaboration, we simplified this process by having MedRIA recommend only the feature with the highest probability at each step. If the inquiry features suggested by MedRIA were in agreement with part of the extracted feature set, it was assumed that the physician accepted the inquiry recommendations. Otherwise, if the suggested features were not found in the extracted feature set, the simulated physician would randomly select an unasked feature from the extracted features as the next inquiry question. This simplification represents a simplistic approximation of a less informed or junior physician's behavior. It serves as a conservative estimate of MedRIA's performance when collaborating with less experienced counterparts, reflecting a lower-bound performance of MedRIA-physician collaborative inquiry. Figure 2 illustrates a vivid example of the simulated collaborative inquiry process.

**Figure 2.** Schematic illustration of the simulated collaborative inquiry process between a physician and MedRIA. AI: artificial intelligence.




## Inquiry Quality Analysis

Table 3 illustrates the performance of MedRIA in the emergency and pediatrics tasks in terms of diagnosis accuracy and inquiry consistency. In the emergency task, physicians used an average of 13.26 (95% CI 13.202-13.317) inquiries to gather 19.488 (95% CI 19.361-19.618) features. This disparity in numbers is because a single inquiry related to a specific laboratory test can yield multiple features. For example, a complete blood count includes several features, such as lymphocyte percentage, platelet count, hemoglobin concentration, and so on. When MedRIA conducted its own inquiries, it averaged 14.24 (95% CI 14.2-14.284) inquiries, with 58% (8.253/14.24; 95% CI 8.212-8.297) of them matching the extracted features. It resulted in the acquisition of an average of 10.628 (95% CI 10.549-10.714) features, accounting for 55% (10.628/19.488) of all extracted features. In the pediatrics task, MedRIA performed, on average, 1.9 more inquiries (16.271, 95% CI 16.231-16.307) than physicians (14.364, 95% CI 14.322-14.404), with 60% (9.81/16.271; 95% CI 9.776-9.842) of them matching 64% (13.168/20.429; 95% CI 13.093-13.244) of all extracted features (20.429, 95% CI 20.338-20.513). When MedRIA collaborated with physicians, it recalled 95% (18.543/19.488; 95% CI 18.425-18.657) and 99% (20.168/20.429; 95% CI 20.080-20.251) of the extracted features in the emergency and pediatrics tasks, respectively. The simulated collaborative inquiry process required only an average of 12.648 (95% CI 12.602-12.696) and 14.146 (95% CI 14.106-14.182) inquiries, respectively, for the emergency and pediatrics tasks, both fewer than the number of inquiries made by physicians. With the assistance of MedRIA, we were able to reduce the number of inquiries made by physicians to 46% (6.037/13.26; 95% CI 6.009-6.064) and 43% (6.245/14.364;

95% CI 6.225-6.269) of those in independent inquiries while recovering nearly all the original inquiry records.

We used multiple metrics to assess final diagnoses, including the area under the receiver operating characteristic curve (AUROC), macro–$F_1$-score, and accuracy. We also introduced a new metric called general accuracy, which considers both fully and partially correct diagnoses in accuracy calculation because health care professionals from different hospitals may have varying recording habits. A diagnosis was considered fully correct when the core diagnosis was recorded and partially correct when only a constituent element of the core diagnosis was recorded [43,44]. As shown in Table 3, when MedRIA conducted its own inquiries in the emergency task, it achieved an AUROC (0.955, 95% CI 0.953-0.956) comparable to when it acquired complete extracted features (0.960, 95% CI 0.957-0.962) despite obtaining only 55% (10.628/19.488) of the extracted features. In the pediatrics task, MedRIA achieved a higher AUROC (0.943, 95% CI 0.941-0.944) with 64% (13.168/20.429) of the extracted features compared to using complete features (0.930, 95% CI 0.928-0.932). As shown in Figure 3, MedRIA's own inquiry and the collaborative inquiry in the emergency task significantly improved the AUROC of diagnostic predictions for acute appendicitis. Specifically, when compared to physicians' inquiry, both MedRIA's own inquiry and the collaborative inquiry demonstrated substantially lower false negative rates (equal to 1 minus true positive rates) at various false positive rate thresholds. These findings underscore the potential of MedRIA to significantly reduce diagnostic errors, particularly for critical conditions such as acute appendicitis in which even small increases in false negative rates may be unacceptable to patients and health system leaders.

The collaborative inquiry and the physicians' inquiry resulted in similar AUROCs for acute upper respiratory tract infections and bronchopneumonia. As shown in Figure 4, MedRIA's own inquiry resulted in a higher AUROC for rhinitis than the physicians' inquiry in the pediatrics task but a lower AUROC for bronchitis and bronchopneumonia. For these 3 diseases, the collaborative inquiry yielded the highest AUROC. The complete receiver operating characteristic curves for all diseases are shown in Figures S1 and S2 in Multimedia Appendix 4. These results demonstrate that MedRIA understands the importance of inquiring about discriminative features that have a high diagnostic contribution, especially in the pediatrics task.

**Table 3.** Performance of MedRIA in the emergency and pediatrics tasks.

| | AUROC[a] (95% CI) | $F_1$-score (95% CI) | General accuracy (95% CI) | Accuracy (95% CI) | Average number of inquiries (95% CI) | Average number of inquiries matching extracted features (95% CI) | Average number of inquiries conducted by physicians (95% CI) | Average number of recalled features (95% CI) |
|---|---|---|---|---|---|---|---|---|
| **Emergency** | | | | | | | | |
| Physicians | 0.96 (0.957-0.962) | 0.63 (0.623-0.638) | 0.853 (0.85-0.856) | 0.744 (0.741-0.748) | 13.26 (13.202-13.317) | 13.26 (13.202-13.317) | 13.26 (13.202-13.317) | 19.488 (19.361-19.618) |
| MedRIA | 0.955 (0.953-0.956) | 0.538 (0.531-0.545) | 0.806 (0.803-0.81) | 0.684 (0.68-0.688) | 14.24 (14.2-14.284) | 8.253 (8.212-8.297) | 0 | 10.628 (10.549-10.714) |
| Collaboration | 0.972 (0.97-0.973) | 0.646 (0.639-0.653) | 0.854 (0.851-0.857) | 0.746 (0.742-0.75) | 12.648 (12.602-12.696) | 12.648 (12.602-12.696) | 6.037 (6.009-6.064) | 18.543 (18.425-18.657) |
| **Pediatrics** | | | | | | | | |
| Physicians | 0.93 (0.928-0.932) | 0.613 (0.608-0.618) | 0.827 (0.825-0.83) | 0.704 (0.701-0.708) | 14.364 (14.322-14.404) | 14.364 (14.322-14.404) | 14.364 (14.322-14.404) | 20.429 (20.338-20.513) |
| MedRIA | 0.943 (0.941-0.944) | 0.521 (0.514-0.527) | 0.782 (0.779-0.784) | 0.656 (0.653-0.66) | 16.271 (16.231-16.307) | 9.81 (9.776-9.842) | 0 | 13.168 (13.093-13.244) |
| Collaboration | 0.968 (0.967-0.969) | 0.633 (0.627-0.638) | 0.842 (0.839-0.844) | 0.732 (0.729-0.735) | 14.146 (14.106-14.182) | 14.146 (14.106-14.182) | 6.245 (6.225-6.269) | 20.168 (20.08-20.251) |

[a]AUROC: area under the receiver operating characteristic curve.

**Figure 3.** Receiver operating characteristic curves for 3 diseases in the emergency task. AUC: area under the curve; AURTI: acute upper respiratory tract infection.

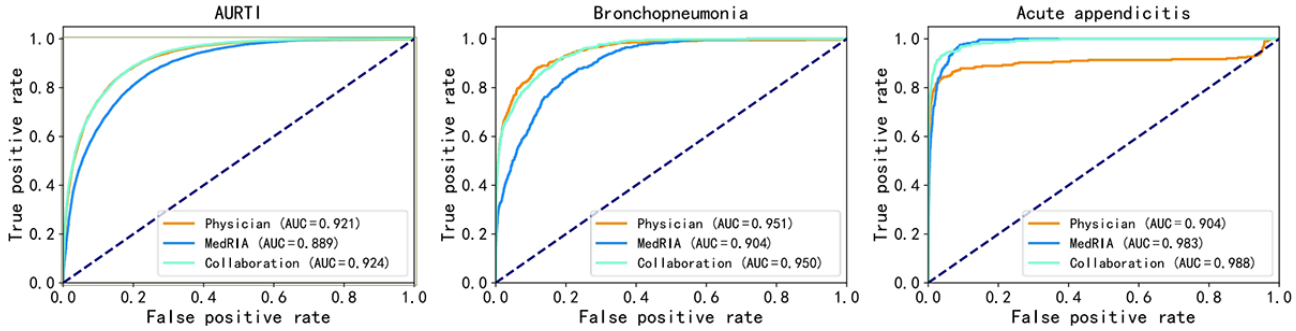
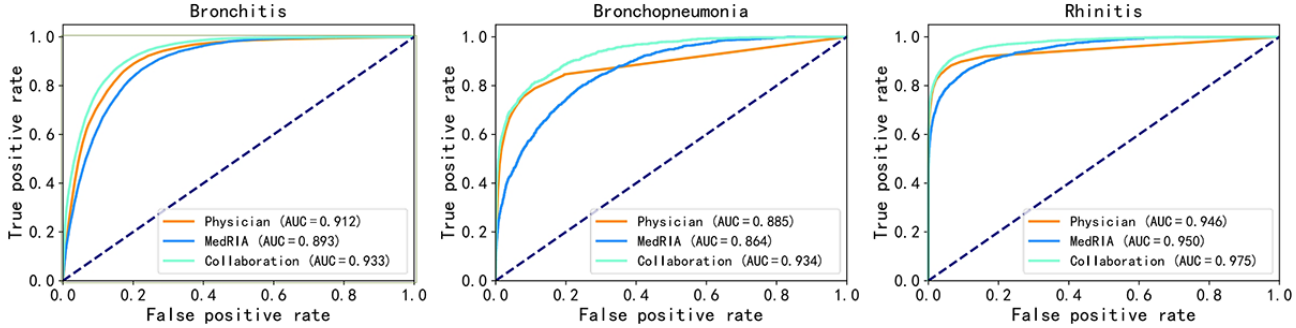



**Figure 4.** Receiver operating characteristic curves for 3 diseases in the pediatrics task. AUC: area under the curve.




MedRIA may conduct inquiries that were not asked by physicians, so they are excluded from the extracted features of patient EHRs. Hence, when MedRIA conducted its own inquiries, there was a varying degree of decline in all diagnostic metrics when compared to using complete extracted features. On the other hand, MedRIA obtained more information through the guidance of physicians when it collaborated with them. In the emergency task, the $F_1$-score (0.646, 95% CI 0.639-0.653) of diagnoses significantly improved. Both general accuracy (0.854, 95% CI 0.851-0.857) and accuracy (0.746, 95% CI 0.742-0.750) were comparable to those of physicians' diagnoses. In the pediatrics task, collaborative inquiry led to a significant increase in $F_1$-score (0.633, 95% CI 0.627-0.638), general accuracy (0.842, 95% CI 0.839-0.844), and accuracy (0.732, 95% CI 0.729-0.735).

Considering the high prevalence of respiratory diseases in our data set, we evaluated MedRIA's performance on a more representative data set to show its generalizability. Specifically, we randomly discarded two-thirds of respiratory cases from the emergency data set and three-quarters from the pediatrics data set. The results are presented in Table S1 in Multimedia Appendix 5. From these results, we can draw similar conclusions to those obtained from the experiments conducted on the complete data set.

## Inquiry Logic Analysis

Table 4 provides the number of different types of positive recalled features, which refer to exhibited symptoms or laboratory test results showing out-of-range values. Collecting more positive features under the same number of inquiries often leads to more accurate diagnosis results [16,45]. In the pediatrics task, 36% (7.326/20.429; 95% CI 7.294-7.357) of the features collected by physicians were positive, and MedRIA's own inquiries recalled 66% (4.852/7.326; 95% CI 4.825-4.877) of positive features. Specifically, it recalled 62% (1.665/2.692; 95% CI 1.654-1.675) of SMH features, 79% (2.24/2.837; 95% CI 2.23-2.25) of PE features, and 53% (0.948/1.797; 95% CI 0.931-0.965) of AE features. When MedRIA worked with physicians, the recall rates for all types of features were >95%. Overall, we observed that MedRIA was capable of accurately identifying positive features that were not mentioned in the chief concerns of patients. This ability is crucial during the investigation process because patients may not always have sufficient medical knowledge to fully describe their health conditions.

Multimedia Appendix 6 shows the top 10 physician inquiry features on patients diagnosed with pharyngitis in the emergency task and tonsillitis in the pediatrics task. We observed that MedRIA could accurately inquire about features obtained by physicians. Figures 5 and 6 illustrate when these features were investigated by MedRIA and through collaborative investigation, respectively. First, we observed a clear distinction between when SMH and PE features were acquired, which is in line with our settings for MedRIA's inquiry process. SMH features, including cough, runny nose, convulsions, and chills, were consistently acquired before the seventh step, whereas PE features, including pharyngeal congestion, enlarged tonsils, coarse breath sounds, rash, herpes, and hand and foot rash, were consistently acquired after the seventh step. We also noticed that consecutive steps of inquiries exhibited consistently high frequencies, allowing us to understand MedRIA's inquiry logic clearly. For example, when conducting PEs, MedRIA began with pharyngeal congestion and enlarged tonsils followed by coarse breath sounds and rash. This logical sequence reflects a structured approach to differential diagnosis, ensuring a thorough evaluation of relevant symptoms and signs. Pharyngeal congestion and enlarged tonsils are pivotal features in the

context of pharyngitis. Coarse breath sounds may suggest respiratory issues or lung conditions, whereas rash can be associated with a wide range of dermatological, allergic, or infectious disorders. As shown in Figures 5 and 6, when collaborating with physicians, the refinement of MedRIA's inquiry logic is evident in the increased stability of determining which features to inquire about at each step. Table 5 provides concrete examples of MedRIA's inquiry sequences, where the aforementioned logic can be traced back. Figure S1 in Multimedia Appendix 7 presents the tree diagram depicting the order of MedRIA's inquiry features for 4 patients diagnosed with pharyngitis in the emergency task. These evidences demonstrate MedRIA's coherent, stable, and well-reasoned inquiry logic. The characteristics of inquiry features for all diseases in both tasks are shown in Figures S1-S6 in Multimedia Appendix 8.

**Table 4.** Number of recalled features from the extracted features by MedRIA in the emergency and pediatrics tasks.

| | Average SMH[a] features (95% CI) | Average PE[b] features (95% CI) | Average AE[c] features (95% CI) | Average positive features[d] (95% CI) | Average positive SMH features (95% CI) | Average positive PE features (95% CI) | Average positive AE features (95% CI) |
|---|---|---|---|---|---|---|---|
| **Emergency** | | | | | | | |
| Physicians | 7.295 (7.26-7.332) | 5.291 (5.265-5.32) | 6.902 (6.793-7.011) | 6.83 (6.788-6.87) | 2.52 (2.503-2.538) | 2.527 (2.514-2.541) | 1.782 (1.749-1.813) |
| MedRIA | 4.31 (4.29-4.331) | 3.835 (3.808-3.865) | 2.483 (2.421-2.546) | 3.923 (3.897-3.954) | 1.363 (1.353-1.375) | 1.907 (1.894-1.92) | 0.653 (0.635-0.672) |
| Collaboration | 7.01 (6.978-7.043) | 5.222 (5.196-5.25) | 6.311 (6.206-6.415) | 6.543 (6.505-6.581) | 2.435 (2.418-2.452) | 2.501 (2.489-2.515) | 1.607 (1.576-1.636) |
| **Pediatrics** | | | | | | | |
| Physicians | 7.522 (7.494-7.546) | 6.29 (6.269-6.312) | 6.617 (6.546-6.687) | 7.326 (7.294-7.357) | 2.692 (2.677-2.707) | 2.837 (2.826-2.848) | 1.797 (1.776-1.82) |
| MedRIA | 5.021 (5.001-5.04) | 4.639 (4.619-4.66) | 3.508 (3.452-3.565) | 4.852 (4.825-4.877) | 1.665 (1.654-1.675) | 2.24 (2.23-2.25) | 0.948 (0.931-0.965) |
| Collaboration | 7.461 (7.435-7.484) | 6.28 (6.258-6.302) | 6.427 (6.356-6.496) | 7.241 (7.209-7.27) | 2.672 (2.658-2.687) | 2.833 (2.822-2.844) | 1.736 (1.714-1.758) |

[a]SMH: symptom and medical history.

[b]PE: physical examination.

[c]AE: auxiliary examination.

[d]Refers to exhibited symptoms or laboratory test results showing out-of-range values.

**Figure 5.** MedRIA and collaborative inquiry heat maps of the top 10 physician inquiry features in patients diagnosed with pharyngitis in the emergency task.

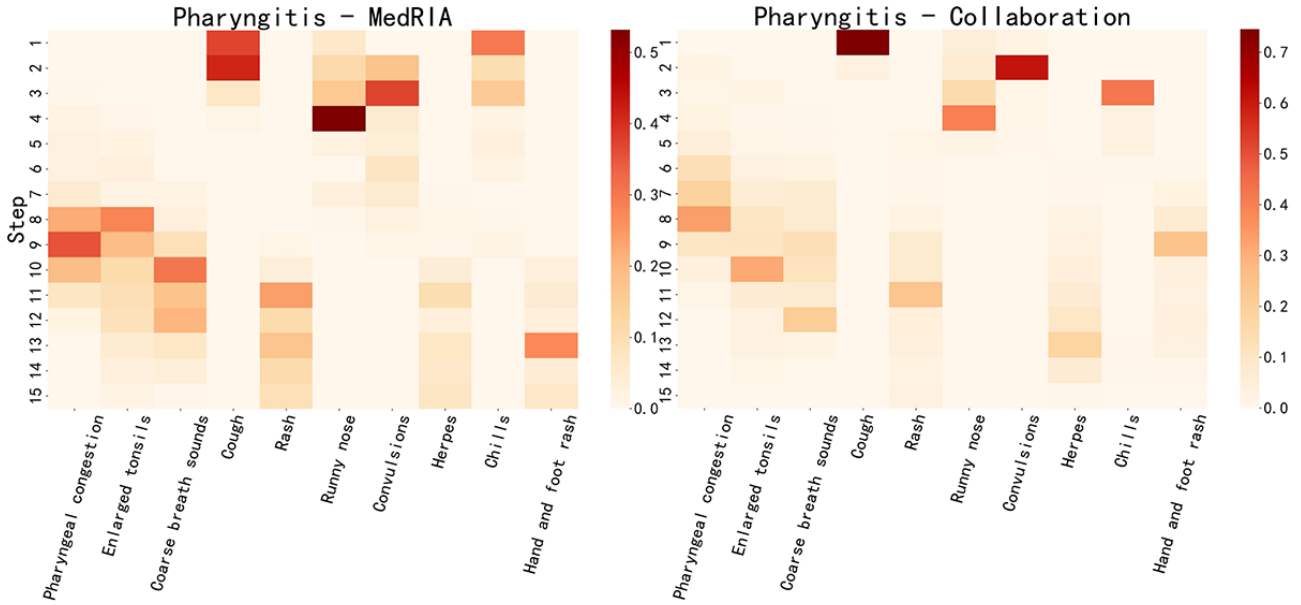


**Figure 6.** MedRIA and collaborative inquiry heat maps of the top 10 physician inquiry features in patients diagnosed with tonsillitis in the pediatrics task.

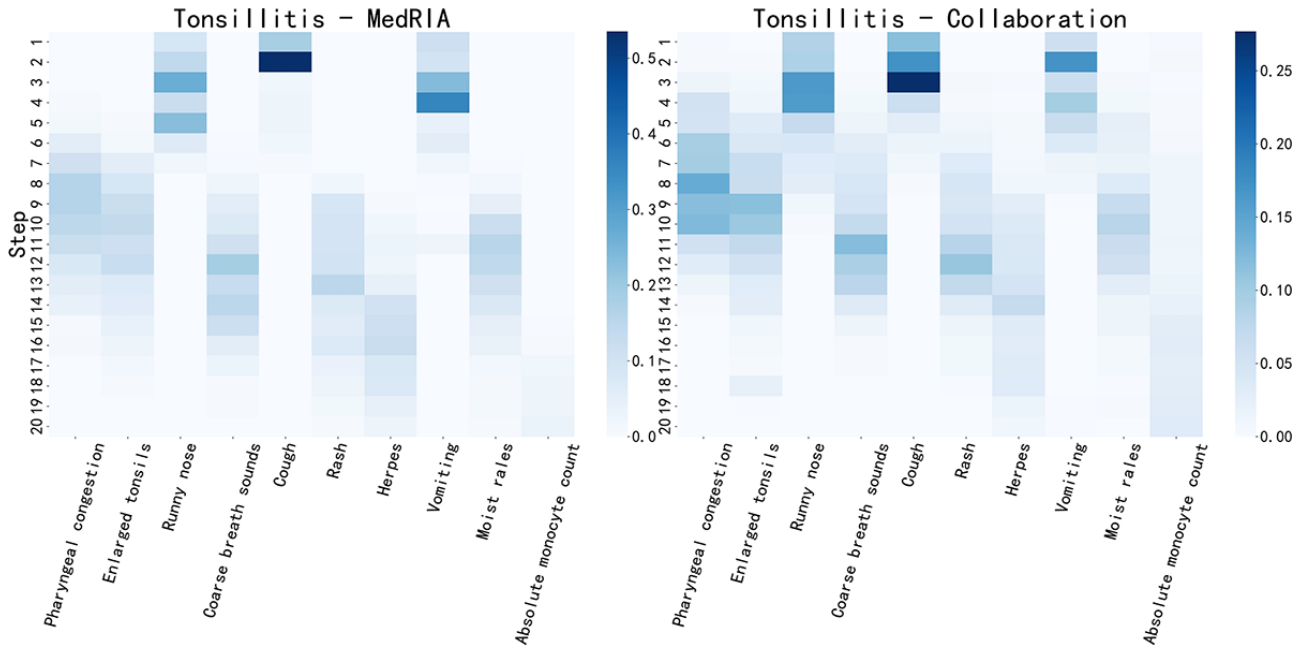



**Table 5.** Examples of the inquiry sequences of MedRIA on 3 patients diagnosed with pharyngitis.

| Gender | Age (y) | Chief concern | Inquiry sequences of SMH[a] | Inquiry sequences of PE[b] | Inquiry sequences of AE[c] |
|---|---|---|---|---|---|
| Male | 26 | *Sore throat* [d] | *Fever(0)*[e], *cough(0)*, *runny nose(0)*, *nasal congestion(0)*, *chest pain(0)*, *coughing up phlegm(0)*, abdominal pain, diarrhea | *Pharyngeal congestion(1)*, *enlarged tonsils(0)*, *moist rales(0)*, *dry rales(0)*, lower limb edema, *coarse breath sounds(0)* | *Absolute monocyte count [complete blood count]* [f] |
| Male | 50 | *Pharyngeal foreign body sensation* | Fever, *cough(1)*, *sore throat(0)*, runny nose, *coughing up phlegm(0)*, abdominal pain, diarrhea, vomiting, dizziness | *Pharyngeal congestion(1)*, *enlarged tonsils(0)*, *moist rales(0)*, *dry rales(0)*, *coarse breath sounds(0)*, rash | Absolute monocyte count |
| Male | 4 | *Fever* | *Cough(1)*, runny nose, *vomiting(0)*, coughing up phlegm, *diarrhea(0)*, *convulsions(0)*, *shortness of breath(0)* | *Pharyngeal congestion(1)*, *coarse breath sounds(1)*, *enlarged tonsils(0)*, *hand and foot Rash(0)*, *dry rales(0)*, normal muscle tone, *moist rales(0)* | *Absolute monocyte count [complete blood count]* |

[a]SMH: symptom and medical history.

[b]PE: physical examination.

[c]AE: auxiliary examination.

[d]Features belonging to the extracted features are italicized.

[e]Feature values are indicated in parentheses.

[f]Contents within square brackets represent the associated laboratory tests documented in the patient's health record.

### External Validation

We used external medical dialogues to evaluate MedRIA's inquiry logic. These dialogues consist of inquiry sequences generated by physicians during medical consultations. Limited by the online scenario, the average length of the inquiry sequences was 3.905 (95% CI 3.744-4.076), consisting of 3.043 (95% CI 2.895-3.195) inquiries about SMH and PE features and 0.862 (95% CI 0.814-0.907) recommended AEs. Diagnoses based on such few features were unreliable, so we primarily focused on evaluating MedRIA's own inquiry order. As a result, MedRIA matched 64% (1.946/3.043; 95% CI 1.835-2.058) of the inquiries in the inquiry sequences and 28% (0.244/0.862; 95% CI 0.218-0.269) of recommended AEs. We compared the order of MedRIA's inquiries to the inquiry sequences of physicians. The normalized Kendall $\tau$ distance was 0.323 (95% CI 0.301-0.346). It indicated that only 32.3% of pairs of inquiry features differed in ordering, demonstrating MedRIA's consistency with physicians and reasonable inquiry logic.

## Discussion

### Principal Findings

In this study, we proposed a reinforcement learning medical inquiry assistant, MedRIA, aiming to provide inquiry recommendations for medical investigation processes. The

variety of questions that can be asked during the medical investigation process, especially when it incorporates PEs and AEs, poses a significant challenge. In addition, patients vary in their understanding of their health status and their ability to articulate it. To address these issues, MedRIA uses an actor-critic framework [28]. The actor incorporates a pretrained VAE [31] to recommend actions based on conditional probability distributions between observed and unobserved patient information. MedRIA uses a supervised diagnostic prediction model to aid in action selection, ensuring that the actor suggests inquiries that contribute to accurate diagnoses. In addition, MedRIA places significant emphasis on crafting a reward function to guide the reinforcement learning process.

The superiority of these designs in MedRIA was verified in our retrospective experiments. When conducting its own inquiries in both emergency and pediatrics scenarios, MedRIA demonstrated comparable diagnostic prediction performance on AUROCs with that of physicians and even achieved a higher AUROC in the pediatrics setting. This was accomplished with MedRIA obtaining only 55% (10.628/19.488) and 64% (13.168/20.429) of the information that physicians received in making the final diagnosis in the emergency and pediatrics tasks, respectively. It is important to consider that AI, similarly to MedRIA, has the capacity to identify subtle patterns within vast data sets, which physicians may overlook. This inherent capability often leads to divergence from the inquiry logic of individual clinicians, as evidenced in our experiments. Given that MedRIA has demonstrated its ability to identify important diagnostic features, such divergence should be considered acceptable as it may provide new insights. In addition, it is worth noting that our experiments were based on EHRs. Therefore, when MedRIA conducted its own inquiries, its performance might very well have been underestimated due to the unavailability of patient information that was not documented in the records.

When MedRIA worked in simulated collaboration with physicians, it notably reduced the number of inquiries made by physicians to 46% (6.037/13.26; 95% CI 6.009-6.064) and 43% (6.245/14.364; 95% CI 6.225-6.269) while maintaining feature recall rates of 95% (18.543/19.488; 95% CI 18.425-18.657) and 99% (20.168/20.429; 95% CI 20.080-20.251) in the emergency and pediatrics scenarios, respectively. This reduction suggests that MedRIA's recommendations are sufficiently informative to assist physicians in considering aspects they may have overlooked. The additional insights provided by AI recommendations have a significant impact on medical decision-making [46]. By leveraging the insights provided by MedRIA, physicians can potentially make more informed decisions and arrive at more accurate diagnoses.

The demonstration and analysis of MedRIA's inquiry feature frequencies, as well as concrete inquiry sequences, reveal its clear inquiry logic for different diseases, which resembles the approach of physicians. The ability of MedRIA to emulate the diagnostic process of human physicians suggests that it has the potential to be a useful tool in medical education, particularly in regions or populations with limited medical resources. Junior health care practitioners can leverage MedRIA to enhance their diagnostic skills, benefiting from its structured and logical inquiry process.

## Limitations

Despite these promising results, it is essential to acknowledge certain limitations. First, the feature extraction process of EHRs requires careful feature engineering, which could be further refined. Second, while MedRIA excelled in adapting its inquiry logic, it is not immune to instances of missed inquiries. For example, the reliance on AEs of MedRIA is significantly lower than that of physicians. One potential solution is to enhance MedRIA by modeling the significance of inquiries for disease treatment. This is because the purpose of examinations may not be for diagnosis but to determine the severity of the disease to formulate appropriate treatment.

Third, the composition of patient presentations can vary significantly between countries. For example, in some Western countries, only 5% to 20% of emergency [47] or pediatrics [48] outpatients are due to upper respiratory tract infections. The proportion in our data set was many times greater due to seasonal variations. We speculate that this may have arisen due to limited access to primary care, seasonal respiratory peaks during the autumn and winter months, greater COVID-19 awareness and fear, and our method of prioritizing the most frequent diagnoses. Changes in these factors, such as using summer data in China or winter data elsewhere, may influence accuracy. Furthermore, while MedRIA's workflow is language independent, the evaluation of our system has been limited to a Chinese context. Further validation in other languages and health care settings is necessary to establish its broader applicability and effectiveness.

Fourth, our AI system would be biased toward excess certainty if deployed in the real world as it lacks the humility to recognize its diagnostic limitations. Future work should include the AI system's degree of confidence and manage corresponding clinical risks. Fifth, decision system developers must consider adopting safe integration systems to mitigate potential errors, especially for life-threatening conditions such as appendicitis, where risk aversion may be greater. Sixth, future work could incorporate MedRIA with the latest developments in LLMs [49], which will assist in its future deployment and testing in clinical workflows.

## Conclusions

In conclusion, MedRIA is a significant step toward enhancing health care decision-making processes. Its proficiency in medical inquiries, ability to identify important diagnostic features, and adaptability underscore its value as an AI-driven diagnostic assistant. While there are challenges to overcome, the results of this study show a prototype for sophisticated and effective AI applied in health care, ultimately benefiting both health care professionals and patients.

## Acknowledgments

This work is supported by the National Key R&D Program of China (2021YFF1201303 and 2022YFC2703105), Guoqiang Institute of Tsinghua University, and Beijing National Research Center for Information Science and Technology. The funders had no role in the study design, data collection and analysis, the decision to publish, or the preparation of the manuscript.

## Conflicts of Interest

None declared.

## Multimedia Appendix 1

The feature set of the emergency task.
[XLSX File (Microsoft Excel File), 26 KB-Multimedia Appendix 1]

## Multimedia Appendix 2

The feature set of the pediatrics task.
[XLSX File (Microsoft Excel File), 28 KB-Multimedia Appendix 2]

## Multimedia Appendix 3

Implementation details of MedRIA.
[DOCX File , 22 KB-Multimedia Appendix 3]

## Multimedia Appendix 4

Receiver operating characteristic curves of all diseases in the emergency and pediatrics tasks.
[DOCX File , 2682 KB-Multimedia Appendix 4]

## Multimedia Appendix 5

Performance of MedRIA in more generalizable emergency and pediatrics tasks.
[DOCX File , 21 KB-Multimedia Appendix 5]

## Multimedia Appendix 6

Inquiry frequencies of the top 10 physician inquiry features in patients diagnosed with pharyngitis in the emergency task and tonsillitis in the pediatrics task.
[PNG File , 92 KB-Multimedia Appendix 6]

## Multimedia Appendix 7

Tree diagram of the inquiry features of MedRIA for 4 patients diagnosed with pharyngitis in the emergency task. These 4 patients shared similar characteristics, all being male with ages ranging from 31 to 32 years and complaining of sore throat. The numbers in brackets indicate the timesteps of inquiries.
[DOCX File , 215 KB-Multimedia Appendix 7]

## Multimedia Appendix 8

Inquiry frequencies of the physician and MedRIA and inquiry heat maps of MedRIA and the collaborative inquiry in the emergency and pediatrics tasks.
[DOCX File , 8348 KB-Multimedia Appendix 8]

## References

1. Musen MA, Middleton B, Greenes RA. Clinical decision-support systems. In: Shortliffe EH, Cimino JJ, editors. Biomedical Informatics: Computer Applications in Health Care and Biomedicine. Cham, Switzerland. Springer; 2021:795-840.
2. Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision support systems: benefits, risks, and strategies for success. NPJ Digit Med. 2020;3:17. [FREE Full text] [doi: 10.1038/s41746-020-0221-y] [Medline: 32047862]
3. Berner ES, Graber ML. Overconfidence as a cause of diagnostic error in medicine. Am J Med. May 2008;121(5 Suppl):S2-23. [doi: 10.1016/j.amjmed.2008.01.001] [Medline: 18440350]

4. Bornstein BH, Emler AC. Rationality in medical decision making: a review of the literature on doctors' decision-making biases. J Eval Clin Pract. May 2001;7(2):97-107. [doi: 10.1046/j.1365-2753.2001.00284.x] [Medline: 11489035]
5. Shickel B, Tighe PJ, Bihorac A, Rashidi P. Deep EHR: a survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. IEEE J Biomed Health Inform. Sep 2018;22(5):1589-1604. [FREE Full text] [doi: 10.1109/JBHI.2017.2767063] [Medline: 29989977]
6. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. Nat Med. Jan 20, 2022;28(1):31-38. [doi: 10.1038/s41591-021-01614-0] [Medline: 35058619]
7. Liang H, Tsui BY, Ni H, Valentim CC, Baxter SL, Liu G, et al. Evaluation and accurate diagnoses of pediatric diseases using artificial intelligence. Nat Med. Mar 2019;25(3):433-438. [doi: 10.1038/s41591-018-0335-9] [Medline: 30742121]
8. Zhou HY, Yu Y, Wang C, Zhang S, Gao Y, Pan J, et al. A transformer-based representation-learning model with unified processing of multimodal input for clinical diagnostics. Nat Biomed Eng. Jun 12, 2023;7(6):743-755. [doi: 10.1038/s41551-023-01045-x] [Medline: 37308585]
9. Huang SC, Pareek A, Seyyedi S, Banerjee I, Lungren MP. Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. NPJ Digit Med. Oct 16, 2020;3(1):136. [FREE Full text] [doi: 10.1038/s41746-020-00341-z] [Medline: 33083571]
10. Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M, et al. Scalable and accurate deep learning with electronic health records. NPJ Digit Med. May 8, 2018;1(1):18. [FREE Full text] [doi: 10.1038/s41746-018-0029-1] [Medline: 31304302]
11. Shen D, Wu G, Suk HI. Deep learning in medical image analysis. Annu Rev Biomed Eng. Jun 21, 2017;19:221-248. [FREE Full text] [doi: 10.1146/annurev-bioeng-071516-044442] [Medline: 28301734]
12. Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJ. Artificial intelligence in radiology. Nat Rev Cancer. Dec 2018;18(8):500-510. [FREE Full text] [doi: 10.1038/s41568-018-0016-5] [Medline: 29777175]
13. Litjens G, Kooi T, Bejnordi BE, Setio AA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. Medical Image Analysis. Dec 2017;42:60-88. [doi: 10.1016/J.MEDIA.2017.07.005]
14. He W, Mao X, Ma C, Huang Y, Hernàndez-Lobato J, Chen T. BSODA: a bipartite scalable framework for online disease diagnosis. In: Proceedings of the 2022 ACM Web Conference. 2022. Presented at: WWW '22; April 25-29, 2022:2511-2521; Virtual Event. URL: https://dl.acm.org/doi/10.1145/3485447.3512123 [doi: 10.1145/3485447.3512123]
15. Chen J, Li D, Chen Q, Zhou W, Liu X. Diaformer: automatic diagnosis via symptoms sequence generation. In: Proceedings of the 36th AAAI Conference on Artificial Intelligence. 2022. Presented at: AAAI '22; February 22–March 1, 2022:4432-4440; Virtual Event. URL: https://cdn.aaai.org/ojs/20365/20365-13-24378-1-2-20220628.pdf [doi: 10.1609/aaai.v36i4.20365]
16. Liu Z, Li Y, Sun X, Wang F, Hu G, Xie G. Dialogue based disease screening through domain customized reinforcement learning. In: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. 2021. Presented at: KDD '21; August 14-18, 2021:1120-1128; Virtual Event. URL: https://dl.acm.org/doi/10.1145/3447548.3467255 [doi: 10.1145/3447548.3467255]
17. Xia Y, Zhou J, Shi Z, Lu C, Huang H. Generative adversarial regularized mutual information policy gradient framework for automatic diagnosis. Proc AAAI Conf Artif Intell. Apr 03, 2020;34(01):1062-1069. [doi: 10.1609/aaai.v34i01.5456]
18. Zhong C, Liao K, Chen W, Liu Q, Peng B, Huang X, et al. Hierarchical reinforcement learning for automatic disease diagnosis. Bioinformatics. Aug 10, 2022;38(16):3995-4001. [doi: 10.1093/bioinformatics/btac408] [Medline: 35775965]
19. Liu W, Cheng Y, Wang H, Tang J, Liu Y, Zhao R, et al. "My nose is running." "Are you also coughing?": building a medical diagnosis agent with interpretable inquiry logics. In: Proceedings of the 31st International Joint Conference on Artificial Intelligence. 2022. Presented at: IJCAI '22; July 23-29, 2022:4266-4272; Vienna, Austria. URL: https://www.ijcai.org/proceedings/2022/0592.pdf [doi: 10.24963/ijcai.2022/592]
20. Wang Z, Wen R, Chen X, Cao SL, Huang S, Qian B. Online disease diagnosis with inductive heterogeneous graph convolutional networks. In: Proceedings of the 2021 Web Conference. 2021. Presented at: WWW '21; April 19-23, 2021:3349-3358; Ljubljana, Slovenia. URL: https://dl.acm.org/doi/10.1145/3442381.3449795 [doi: 10.1145/3442381.3449795]
21. Lin J, Wang K, Chen Z, Liang X, Lin L, Lin L. Towards causality-aware inferring: a sequential discriminative approach for medical diagnosis. IEEE Trans Pattern Anal Mach Intell. 2021:1-12. [doi: 10.1109/tpami.2023.3292363]
22. Ouyang L, Wu J, Jiang X, Almeida D, Wainwright C, Mishkin P, et al. Training language models to follow instructions with human feedback. arXiv. Preprint posted online March 4, 2022. [FREE Full text]
23. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health. Feb 9, 2023;2(2):e0000198. [FREE Full text] [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
24. Meskó B, Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. NPJ Digit Med. Jul 06, 2023;6(1):120. [FREE Full text] [doi: 10.1038/s41746-023-00873-0] [Medline: 37414860]
25. Thirunavukarasu AJ, Ting DS, Elangovan K, Gutierrez L, Tan TF, Ting DS. Large language models in medicine. Nat Med. Aug 17, 2023;29(8):1930-1940. [doi: 10.1038/s41591-023-02448-8] [Medline: 37460753]
26. Wornow M, Xu Y, Thapa R, Patel B, Steinberg E, Fleming S, et al. The shaky foundations of large language models and foundation models for electronic health records. NPJ Digit Med. Jul 29, 2023;6(1):135. [FREE Full text] [doi: 10.1038/s41746-023-00879-8] [Medline: 37516790]

27. Taylor J. ChatGPT's alter ego, Dan: users jailbreak AI program to get around ethical safeguards. The Guardian. URL: https://www.theguardian.com/technology/2023/mar/08/chatgpt-alter-ego-dan-users-jailbreak-ai-program-to-get-around-ethical-safeguards [accessed 2024-04-29]
28. Mnih V, Badia A, Mirza M, Graves A, Lillicrap T, Harley TP, et al. Asynchronous methods for deep reinforcement learning. In: Proceedings of the 33rd International Conference on International Conference on Machine Learning. 2016. Presented at: ICML'16; June 19-24, 2016:1928-1937; New York, NY. URL: https://dl.acm.org/doi/10.5555/3045390.3045594
29. Balogh EP, Miller BT, Ball JR. Improving Diagnosis in Health Care. New York, NY. The National Academies Press; 2015.
30. Kenton JD, Toutanova LK. BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2019. Presented at: NAACL-HLT '19; June 2-7, 2019:4171-4186; Minneapolis, MN. URL: https://aclanthology.org/N19-1423.pdf [doi: 10.18653/v1/N19-1423]
31. Kingma DP, Welling M. Auto-encoding variational bayes. arXiv. Preprint posted online December 20, 2013. [FREE Full text] [doi: 10.61603/ceas.v2i1.33]
32. Schulman J, Wolski F, Dhariwal P, Radford A, Klimov O. Proximal policy optimization algorithms. arXiv. Preprint posted online July 20, 2017. [FREE Full text]
33. Konda VR, Tsitsiklis JN. OnActor-critic algorithms. SIAM J Control Optim. Jan 2003;42(4):1143-1166. [doi: 10.1137/S0363012901385691]
34. Schulman J, Moritz P, Levine S, Jordan M, Abbeel P. High-dimensional continuous control using generalized advantage estimation. arXiv. Preprint posted online June 8, 2015. [FREE Full text] [doi: 10.48550/arXiv.1506.02438]
35. Kingma DP, Ba J. Adam: a method for stochastic optimization. arXiv. Preprint posted online December 22, 2014. [FREE Full text]
36. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. PyTorch: an imperative style, high-performance deep learning library. arXiv. Preprint posted online December 3, 2019. [FREE Full text] [doi: 10.7551/mitpress/11474.003.0014]
37. Weng J, Chen H, Yan D, You K, Duburcq A, Zhang M. Tianshou: a highly modularized deep reinforcement learning library. J Mach Learn Res. 2022;23(267):1-6. [FREE Full text]
38. Bellman R. A markovian decision process. Indiana Univ Math J. 1957;6(4):679-684. [doi: 10.1512/iumj.1957.6.56038]
39. He W, Chen T. Scalable online disease diagnosis via multi-model-fused actor-critic reinforcement learning. In: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 2022. Presented at: KDD '22; August 14-18, 2022:4695-4703; Washington, DC. URL: https://dl.acm.org/doi/10.1145/3534678.3542672 [doi: 10.1145/3534678.3542672]
40. Wu M, Goodman N. Multimodal generative models for scalable weakly-supervised learning. arXiv. Preprint posted online February 14, 2018. [FREE Full text]
41. Chen W, Li Z, Fang H, Yao Q, Zhong C, Hao J, et al. A benchmark for automatic medical consultation system: frameworks, tasks and datasets. Bioinformatics. Jan 01, 2023;39(1):btac817. [FREE Full text] [doi: 10.1093/bioinformatics/btac817] [Medline: 36539203]
42. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W. LightGBM: a highly efficient gradient boosting decision tree. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017. Presented at: NIPS '17; December 4-9, 2017:3149-3157; Long Beach, CA. URL: https://dl.acm.org/doi/10.5555/3294996.3295074
43. Schmidt HG, Van Gog T, Schuit SC, Van den Berge K, Van Daele PL, Bueving H, et al. Do patients' disruptive behaviours influence the accuracy of a doctor's diagnosis? A randomised experiment. BMJ Qual Saf. Jan 07, 2017;26(1):19-23. [doi: 10.1136/bmjqs-2015-004109] [Medline: 26951795]
44. Mamede S, van Gog T, van den Berge K, Rikers RM, van Saase JL, van Guldener C, et al. Effect of availability bias and reflective reasoning on diagnostic accuracy among internal medicine residents. JAMA. Sep 15, 2010;304(11):1198-1203. [doi: 10.1001/jama.2010.1276] [Medline: 20841533]
45. Peng Y, Tang K, Lin H, Chang E. REFUEL: exploring sparse features in deep reinforcement learning for fast disease diagnosis. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems. 2018. Presented at: NIPS'18; December 3-8, 2018:7333-7342; Montréal, QC. URL: https://dl.acm.org/doi/10.5555/3327757.3327834 [doi: 10.7551/mitpress/11474.003.0014]
46. Nagendran M, Festor P, Komorowski M, Gordon AC, Faisal AA. Quantifying the impact of AI recommendations with explanations on prescription decision making. NPJ Digit Med. Nov 07, 2023;6(1):206. [FREE Full text] [doi: 10.1038/s41746-023-00955-z] [Medline: 37935953]
47. Nørgaard B, Mogensen CB, Teglbjærg LS, Brabrand M, Lassen AT. Diagnostic packages can be assigned accurately in emergency departments. A multi-centre cohort study. Dan Med J. Jun 2016;63(6):A5240. [FREE Full text] [Medline: 27264941]
48. Hiscock H, Roberts G, Efron D, Sewell JR, Bryson HE, Price AM, et al. Children attending paediatricians study: a national prospective audit of outpatient practice from the Australian paediatric research network. Med J Aust. Apr 18, 2011;194(8):392-397. [doi: 10.5694/j.1326-5377.2011.tb03028.x] [Medline: 21495938]
49. Moor M, Banerjee O, Abad ZS, Krumholz HM, Leskovec J, Topol EJ, et al. Foundation models for generalist medical artificial intelligence. Nature. Apr 12, 2023;616(7956):259-265. [doi: 10.1038/s41586-023-05881-4] [Medline: 37045921]

## Abbreviations

**AE:** auxiliary examination
**AI:** artificial intelligence
**AUROC:** area under the receiver operating characteristic curve
**CDSS:** clinical decision support system
**EHR:** electronic health record
**LLM:** large language model
**MedBERT:** Medical Bidirectional Encoder Representations From Transformers
**MLP:** multilayer perceptron
**PE:** physical examination
**SMH:** symptom and medical history
**VAE:** variational autoencoder

*Edited by S Ma; submitted 17.11.23; peer-reviewed by H Lee, T Tillmann; comments to author 14.03.24; revised version received 13.04.24; accepted 04.07.24; published 23.08.24*

*Please cite as:*
*Zou X, He W, Huang Y, Ouyang Y, Zhang Z, Wu Y, Wu Y, Feng L, Wu S, Yang M, Chen X, Zheng Y, Jiang R, Chen T*
*AI-Driven Diagnostic Assistance in Medical Inquiry: Reinforcement Learning Algorithm Development and Validation*
*J Med Internet Res 2024;26:e54616*
*URL: https://www.jmir.org/2024/1/e54616*
*doi: 10.2196/54616*
*PMID:*