DDXPlus: A new Dataset for Medical Automatic Diagnosis

Arsene Fansi Tchango, Zhi Wen, Rishab Goel, Joumana Ghosn

May 19, 2022

Abstract

There has been rapidly growing interests in Automatic Diagnosis (AD) and Automatic Symptom Detection (ASD) systems in the machine learning research literature, aiming to assist doctors in telemedicine services. These systems are designed to interact with patients, collect evidence relevant to their concerns, and make predictions about the underlying diseases. Doctors would review the interaction, including the evidence and the predictions, before making their final decisions. Despite the recent progress, an important piece of doctors' interactions with patients is missing in the design of AD and ASD systems, namely the differential diagnosis. Its absence is largely due to the lack of datasets that include such information for models to train on. In this work, we present a large-scale synthetic dataset that includes a differential diagnosis, along with the ground truth pathology, for each patient. In addition, this dataset includes more pathologies, as well as types of symtoms and antecedents. As a proof-of-concept, we extend several existing AD and ASD systems to incorporate differential diagnosis, and provide empirical evidence that using differentials in training signals is essential for such systems to learn to predict differentials[1].

1 Introduction

In a clinical conversation between a doctor and a patient, the patient usually initiates the discussion by specifying an initial set of symptoms they are experiencing. The doctor then iteratively inquires about additional informative symptoms and antecedents (describing the patient's history and potential risk factors), and produces a differential diagnosis, i.e. a set of plausible diseases, at the end of the interaction. During this multi-step process, the doctor tries to collect all relevant information to narrow down the list of the differentials. Once the differential diagnosis is established, the doctor can ask the patient to undergo medical exams to eliminate most pathologies included in the differential and confirm the one(s) the patient is suffering from, or can decide to directly prescribe a treatment to the patient.

[1]Dataset available at https://github.com/bruzwen/ddxplus.

Aiming to aid doctors in such clinical interactions, there has been significant recent progress on building Automatic Diagnosis (AD) systems and Automatic Symptom Detection (ASD) systems, using recent machine learning and Reinforcement Learning (RL) techniques [1, 2, 3, 4, 5, 6, 7]. They are meant to collect all symptoms and antecedents relevant to the patient's concern, while minimizing the length of the interaction to improve efficiency. They can potentially also predict the underlying disease to further aid the doctors in deciding appropriate next steps in the patient diagnoses.

However, this setting differs from realistic patients' interactions in an important way, namely the absence of the differential diagnosis. Based on the conversation alone, without further evidence such as physical exams, doctors tend to consider the differentials rather than a single pathology [8]. Doing so accounts for the uncertainty in the diagnosis and presents a more comprehensive view of the doctor's opinions on the underlying disease. Considering differentials is especially important for AD and ASD systems to account for the potential errors in predictions, and therefore allowing better acceptability by doctors. The absence of differential diagnosis in recent AD/ASD systems is mainly due to the lack of dataset that includes such information. The most commonly used public datasets, for example DX [1], Muzhi [2] and SymCAT [9], all are designed for predicting the ground truth pathology and lack differentials.

To close this gap and encourage future research that focuses on differential diagnosis, we present a large-scale synthetic dataset for building AD and ASD systems. This dataset is similar in format to other public datasets such as DX [1] and Muzhi [2], but differs in several important ways. First, it is larger in scale, in terms of the number of patients, as well as the number of pathologies, symptoms and antecedents. Second, it includes not only binary evidence, as existing datasets do, but also categorical, multi-choice and numerical types. Finally, each patient has a corresponding set of differential diagnosis in addition to the ground truth pathology. To the best of our knowledge, this is the first large scale dataset that includes both the ground truth pathology, a differential diagnosis and non-binary symptoms. To summarize, we make the following contributions:

- Release a large scale synthetic benchmark dataset of 1 million patients. The dataset is generated using a proprietary medical knowledge base and contains a mixture of multi-choice, categorical and binary symptoms and antecedents. It also contains a differential diagnosis for each patient.
- We extend several existing AD and ASD systems to incorporate differential diagnosis. We then show that using the differentials in training signals is essential for such systems to be able to predict differentials.

2 Existing datasets and their limitations

The agent's training requires having access to the symptoms experienced by each patient, the relevant antecedents, and the differential diagnosis. There

is unfortunately no such public data set. Existing public datasets, such as the MIMIC-III dataset [10], often lack symptom-related data and are therefore inappropriate. Other datasets, such as Munzhi [2], are of small scale, lack the differential diagnosis, and don't necessarily keep track of all the evidences experienced by patients. Moreover, due to privacy laws and security concerns, medical data is difficult to obtain from clinics and hospitals.

To tackle these limitations, previous works [9, 11] relied on the SymCAT database [12] for data synthesis. Unfortunately, SymCAT doesn't provide a differential diagnosis. Moreover, SymCAT is limited to binary evidences, which can lead to unnecessarily long interactions with patients (compared to categorical or multi-choice questions that allow the collection of information with a smaller number of dialog turns). Finally, the symptoms listed in SymCAT are not always defined in an understandable way by patients and would require to be made more explicit (e.g., "are you experiencing flu-like syndrome?").

We choose not to use DX [1] and Muzhi [2] datasets, both of which are commonly used as public benchmarks in prior works [3, 4, 5, 6]. The main reason is that they have a small number of samples (527 and 710 respectively), diseases (5 and 4 respectively), and symptoms (41 and 67 respectively), as noted by prior works [7]. Additionally, we also notice that the reliability of DX dataset was questioned by reviewers in previous submissions to ICLR[2], further undermining the suitability of DX dataset in this work.

3 Proposed dataset

3.1 Source

The dataset we propose in this work heavily relies on a proprietary knowledge base (KB) extracted from medical literature and which was used to design OTA, a rule-based system that has been deployed in a real-world telemedicine platform. In total, the knowledge base covers a set of 440 pathologies and 802 symptoms and antecedents. The pathologies are regrouped in overlapping subgroups based on common characteristics refer to as *chief complaints* [13, 14]. In this work, as a first step, we focus on pathologies belonging to the chief complaint related to cough, sore throat, or breathing issues. This subgroup is of medium size and it contains a set of 49 pathologies covering 110 symptoms and 113 antecedents. Extending the dataset to all pathologies is left for future work.

Each pathology d in the knowledge base is characterized by either an incidence rate, a prevalence rate, or both values. Both rates are conditioned on the age, the sex, and the geographical region of the patient. Additionally, for each pathology, a set of symptoms and antecedents describing the pathology is provided together with their related probabilities. These probabilities are conditioned on the age and sex of the patient. Thus, the values $p(s|d, age, sex)$ and $p(a|d, age, sex)$ are provided for each symptom s and each antecedent a.

[2] https://openreview.net/forum?id=TCAmP8zKZ6k&noteId=SCxQNdf67Rw

Unlike existing datasets mentioned above, evidences (i.e., symptoms and antecedents) within this knowledge base are of several types. They can be binary (e.g., cough?), categorical (e.g., pain intensity from 0 to 10?), or multi-choice (e.g., pain location?). Sometimes, an evidence f_s (e.g., pain location) may be dependent on another evidence s (e.g., pain), in which case the knowledge base provides means to extract the corresponding probability $p(f_s|s, d, age, sex)$. Finally, each pathology is characterized by its level of severity ranging from 1 to 5 with the lowest values describing the most severe pathologies from an emergency perspective.

To generate the synthetic patients, we rely on *Synthea*^{TM} [15], a synthetic patient generator that generates high-quality patient data along with the associated health records covering every aspect of healthcare, as well as the statistics extracted from the knowledge base. In the next section, we describe in details the generation process as well as the assumptions made to this end.

3.2 Generation Process

To generate the synthetic dataset, we made some assumptions and developed several rules to exploit the knowledge base.

Assumptions on Socio-Demographic Data

As mentioned above, the pathology's statistics from the knowledge base are conditioned on the age, the sex, and the geographical region of the patient. In this work, we assume that the age, sex, and geographical region are independent. In other words, we have

$$p(age, sex, geo) = p(age) \times p(sex) \times p(geo)$$

, where *geo* is the random variable representing the geographical region. The distribution on the age and the one on the sex can both be obtained from Census data. For this dataset, we used the 2010-2015 US Census data from the State of New York [16]. For more details, see Section 5.1. Regarding the geographical region, one needs to embed the notion of patient location, or at least the notion of recent travel while synthesizing a patient. In this project, we opt for the second case: each synthesized patient is generated by simulating the fact that he recently travelled or not, and if he travelled, in which geographical region. This choice is motivated by the fact that we are synthesizing patients from the state of New York population statistics and there are some pathologies of interest that can be contracted only if the patient is from a different geographical region. We thus assume the availability of a prior distribution $p(travel)$ representing the proportion of the population travelling each month and we consider that the distribution regarding the geographical regions of destination is uniform. Finally, we assume that the default geographical region is "North America" for any person who has not recently travelled. Based on these assumption, we derive the following prior distribution $p(geo)$ over the geographical regions:

- Sample $u \sim \mathcal{U}(0, 1)$.
- If $u < p(travel)$, then randomly select a geographical region from the available set of geographical regions. We used $p(travel) = 0.25$ for this dataset.
- If $u \geq p(travel)$, then set the geographical region to be "North America".

Assumptions on Pathologies

In this work, the incidence rate, when available, is used as the pathology prior distribution. However, we fall back to the prevalence rate when the incidence rate is not available. This is one of the major limitations of the data generation process which needs to be addressed in future work. From several discussions with a medical expert, it seems that one can approximate incidence rate with the prevalence by multiplying it with a constant factor (representing the average duration of the disease) which can be different for each pathology. Out of the 49 pathologies present in this dataset, 8 are affected by this shortcut.

When the resulting rate is greater than 100% (e.g., an incidence rate of 200% means that an individual will likely develop the pathology on average twice a year), we simply cap it at 100%. We first explored the strategy consisting of capping the rate at 100% and generating as much as patients to comply with the original rate (e.g., we generate two patients for an incidence rate of 200%). However this strategy led us to a dataset that was dominated by only a few pathologies (more than half of the patients within the dataset were suffering from one of the three pathologies whose incidence rate was greater than 100%).

The knowledge base also contains some diseases that have extremely low incidence rates, and therefore patients suffering from those pathologies were barely generated. To increase the chance of those pathologies to be represented within the dataset, we decided to cap the rate at a minimum of 10%. In other words, the rates used to generate the dataset were capped to lie between 10% and 100%. This simple alteration of the original rates from the knowledge base leads to the generation of a significant number of patients for all the diseases.

Assumptions on Symptoms and Antecedents

At this point, we are able to sample a pathology d from its prior distribution $p(d|age, sex, geo)$. The next step is to generate all the evidences (symptoms and antecedents) the synthesized patient will be experiencing. However, the knowledge base doesn't contain the joint distribution of symptoms and antecedents given the disease, sex and age. It only contains marginal distributions for each symptom and antecedent. So, a simplifying assumption is made according to which given the disease, age and sex, all the evidences are conditionally independent of each other. In other words, we have:

$$p(E|d, age, sex) = \prod_{e \in E} p(e|d, age, sex)$$

, where E is the set of evidences experienced by the patient. This simplifying assumption is yet another limitation of our dataset.

Some evidences, such as the pain intensity, are described as integer values on a scale of 0 to 10. However, the knowledge base only provides the average value of this evidence given the disease, the age, and the sex of the patient. To inject some randomness in the patient generation process, the values of those evidences are uniformly sampled from the interval $[\max(0, v-3), \min(10, v+3)]$ where v is the average value present in the knowledge base.

Additionally, for realistic purposes, we limit to 5 the maximum number of choices associated with multi-choices evidences such as pain location.

3.2.1 Tools

As mentioned above, we rely on *Synthea*^{TM} along with the described assumptions on conditional probabilities from the knowledge base to generate the patients. However, *Synthea*^{TM} relies on static graphs refereed to as modules to synthesize the patients. Because those graphs are static, the order in which the possible values of a categorical and a multi-choice evidence are traversed during the generation process is fixed and deterministic. Consequently, *Synthea*^{TM} will stop exploring the remaining values as soon as a value is synthesized for a categorical evidence or 5 values are synthesized for a multi-choice evidence. In this work, the possible values of an evidences are ordered, within the *Synthea*^{TM} modules, in ascending order based on their conditional probability of occurrence $p(e|d, age, sex)$.

3.3 Differential Diagnosis Generation

From the assumptions made above, we are able to synthesize a patient, that is someone suffering from a pathology and experiencing the related symptoms and antecedents. In this section, we focus on the generation of the differential diagnosis associated with the set of symptoms and antecedents experienced by a given patient.

As mentioned above, the knowledge base we rely on has been used to build a rule-based system which deployed in a real-world telemedicine platform. We leverage this platform to compute the differential diagnosis. More specifically, we proceed by using the platform in a *batch*[3] mode according to the following high-level steps:

- We provide the age and the sex of the patient, the appropriate chief complaint, and we answer "yes" to the question "Are you consulting for a new problem?".
- We add all the generated symptoms and antecedents experienced by the patient to the payload at the beginning of the interaction. The motivation behind this is to provide as much information to the platform so as

[3]As opposed to the *interactive* mode where the evidences are provided sequentially upon requests made by the platform.

to minimize the bias resulting from the interaction into the differential diagnosis.

- The platform may still inquire about additional questions. If that is the case, we answer "no" for those questions until we see a "QUIT" response from the platform or the maximum interaction length is reached.
- When the maximum interaction length is reached, the platform does not produce a differential diagnosis. We interpret this situation as if the synthesized patient is not as realistic as needed by the platform, and therefore, the patient is discarded from the dataset.
- When a "QUIT" response is provided by the platform, it contains a differential diagnosis. We further proceed by verifying if the synthesized disease is part of the generated differential diagnosis. If it is not the case (because the platform itself is not a perfect system or because the patient didn't have enough evidences for the rule-based system to include the simulated disease in the differential diagnosis), the patient is discarded from the dataset. Each pathology within the generated differential diagnosis has a score. Those scores are normalized to obtain a probability distribution.

The platform sometimes returns a differential diagnosis that contains pathologies which do not belong to the specified chief complaint. There are several options for handling this situation: (1) create an "other pathologies" category and assign it the cumulative mass of the corresponding pathologies, or (2) manually remove those pathologies from the differential diagnosis and re-normalize the distribution. We opt for the second option in this work. On average, we removed 1.78 ($\pm$1.68) pathologies from the generated differential diagnosis for an average cumulative probability mass of 0.10 ($\pm$0.11). Statistics regarding the rank from which those pathologies are excluded are described in Section 5.2.

3.4 Dataset Characteristics

With the above assumptions and limitations, we generate a large scale dataset of roughly of 1.32 Millions distinct patients where a patient is characterized by the combination of his/her age, sex, race, pathology, symptoms, antecedents, as well as the corresponding differential diagnosis. We further divide the dataset in training, validation, and test subsets using the stratified sampling strategy based on the simulated pathology as well as the classical 80-10-10 proportions.

With respect to existing datasets from the Automatic Diagnosis and Automatic Symptom Detection literature, our dataset present several advantages:

- Unlike the *SymCAT* [12] and the Muzhi [1] datasets, our dataset does not only contain binary evidences. Instead, it also includes categorical and multi-choice evidences which can be naturally match to the kind of questions a doctor can ask to a given patients.

- Our dataset makes a clear distinction between antecedents and symptoms which can be of great importance when designing automated systems. While both symptoms and antecedents are useful for characterising a pathology, antecedents are usually known from the patient medical records, and therefore one can decide to put less emphasis on retrieving them when designing such an automated system.
- Each pathology in our dataset is characterized by a severity level which makes it suitable for designing solutions dedicated to emergency scenarios.
- To the best of our knowledge, this is the first large scale dataset containing differential diagnosis which can be useful for designing fine-grained automated systems.

In the next section, we perform further analysis of the generated dataset.

3.5 Data Analysis

In this section, we conduct throughout analysis of the so generated dataset and report the resulting statistics. The results reported in this section are for the whole set of generated patients. Statistics on the train, validation, and test subsets are presented in Section 5.3.

Types of evidences: The statistics regarding the types of evidences that are part of the generated dataset are shown in Table 1. Although the dataset is mostly composed of binary evidences, it does include categorical and multi-choice evidences which are evidences that cannot be characterized by a yes-or-no option.

	Evidences	**Symptoms**	**Antecedents**
Binary	208	96	112
Categorical	10	9	1
Multi-choice	5	5	0
Total	223	110	113

Table 1: The statistics of the considered evidences from the knowledge base.

Number of evidences: Table 2 shows an overview of the synthesized patients in terms of the number of simulated evidences. On average, a synthesized patient is characterized by 10.02 symptoms as well as 3.45 antecedents.

Pathology Statistics: Figure 1 shows the histogram of the pathology the synthesized patients are suffering from in the generated dataset. Although there are three dominating pathologies (URTI, Viral pharyngitis, and Anemia), it can be observed that the other pathologies are also well represented.

Socio-Demographic Statistics: The statistics regarding the socio-demographic data of the synthesized patients are shown in Figure 2. As expected, these statistics are compliant with the 2015 US Cesus data of the state of New-York which was used as reference during the generation process.

	Evidences	Symptoms	Antecedents
Avg	13.48	10.02	3.45
Std dev	5.12	4.74	2.23
Min	1	1	0
1st quartile	10	8	2
Median	13	10	3
3rd quartile	17	12	5
Max	36	25	12

Table 2: The statistics of the synthesized patients.

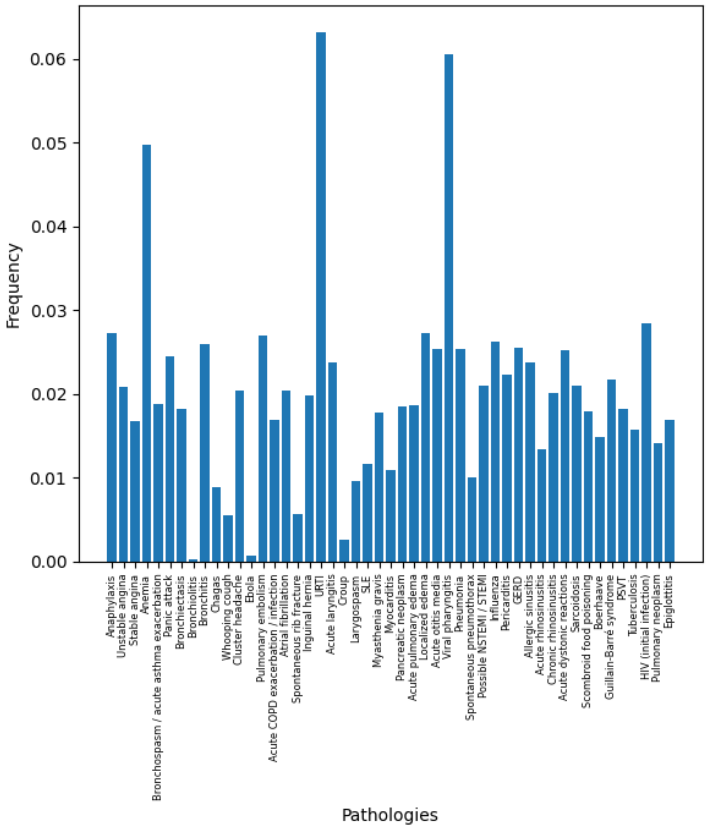
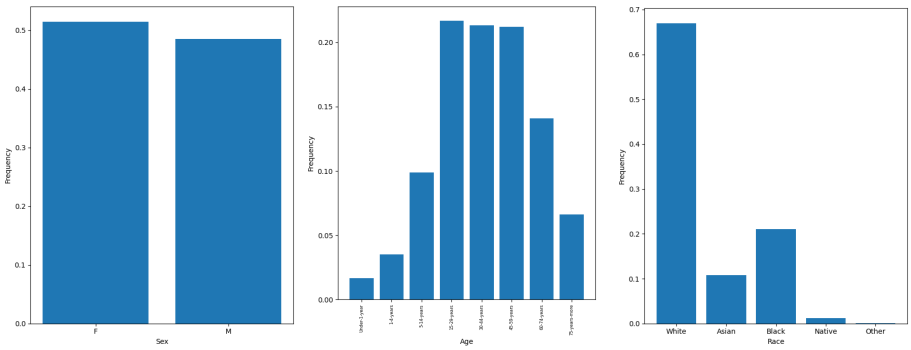


Figure 1: Histogram of the pathologies experienced by the synthesized patients.

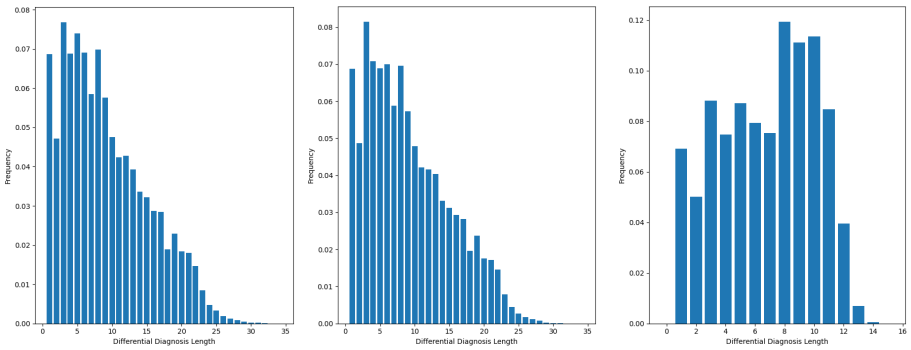
Differential Diagnosis Statistics: The histogram of the length of the differential diagnosis characterizing the generated patients is depicted in Figure 3a. Figure 3b (resp. Figure 3c) illustrates the same histogram when the pathology



(a) (b) (c)

Figure 2: The socio-demographic statistics of the synthesized patients. The sex distribution is shown in plot (a) while the age and the race distributions are respectively depicted in plots (b) and (c).

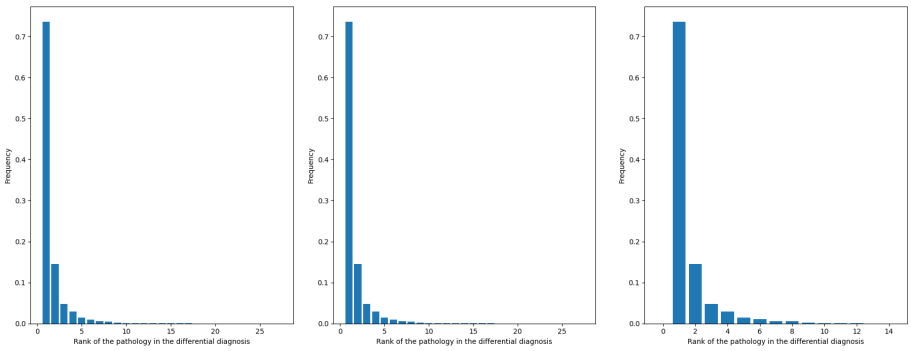
with a probability mass less than or equals to 0.01 (resp. 0.05) are filtered out from the differential diagnoses. As observed, the generated differential diagnosis can have a number of pathologies ranging from 1 to more than 10.



(a) (b) (c)

Figure 3: Statistics regarding the length of the differential diagnosis. a) Full differential as generated by the system. b) Pathologies with probability mass less than or equals to 0.01 are filtered out. c) Same as (b) but the threshold is set to 0.05.

It is also interesting to have an insight of the the rank of the simulated patient pathology within the generated differential diagnosis. Figure 4 particularly addresses this point. As it can be noticed, the simulated pathology is ranked first for more than 70% of patients. This finally validates the quality of the generated data.



(a) (b) (c)

Figure 4: Statistics regarding the rank of the simulated pathology within the differential diagnosis. a) Full differential as generated by the system. b) Pathologies with probability mass less than or equals to 0.01 are filtered out. c) Same as (b) but the threshold is set to 0.05.

3.6 Disclaimer

From the sections discussed above, it should be clear that any models trained on this dataset should not be directly used in a real-world system prior to performing rigorous evaluations to ensure proper coverage and representativity of the population that such a model will interact with.

4 Experiments

There hasn't been a lot of work on building automatic diagnosis systems that aims at producing a differential diagnosis based on the collected evidences from the patients.In this secton, we propose to adapt two of the existing approaches to our setting:

- **AARLC**: AARLC [7] is an RL-based approach consisting of two branches, an evidence acquisition branch and a classifier branch. The method proposes an adaptive approach to align the tasks performed by the two branches using entropy of the distributions predicted by the classifier branch.
- **ASD**: this supervised learning based approach builds on top of the evidence acquisition module proposed in [17] - with the exception of the knowledge graph - and adds a policy network which aims at predicting the underlying patient disease at the end of the evidence acquisition process. More details about this approach in Appendix 6.

Furthermore, we train for each approaches two versions: a version which is trained to predict the ground truth pathology and another one which is trained

Table 3

Method	IL	GTPA@1	GTPA@3	GTPA@5	PER	PEP	PEF1
ASD w/ Diff	17.46	68.23	91.35	96	87.71	74.05	76.64
AARLC w/ Diff	14.1	61.74	91.51	96.6	51.56	74.83	56.99
ASD w/o Diff	17.85	96.65	98.32	98.34	88.15	73.26	76.17
AARLC w/o Diff	7.27	98.86	99.79	99.93	35.91	79.38	45.51

Table 4

Method	DDR@1	DDP@1	DDF1@1	DDR@3	DDP@3	DDF1@3	DDR@5	DDP@5	DDF1@5
ASD w/ Diff	79.59	79.59	79.59	75.72	76.7	75.27	77.1	78.86	76.36
AARLC w/ Diff	71.59	71.59	71.59	74.86	69.58	71.14	79.33	69.79	72.53
ASD w/o Diff	71.35	71.35	71.35	37.19	92.05	51.25	28.42	96.54	41.05
AARLC w/o Diff	72.73	72.73	72.73	37.31	91.78	51.25	28.56	95.97	41.05

to predict the differential diagnosis.

4.1 Experimental setup

Each patient has one initial evidence which is provided to the model at the beginning of the interaction. The model then iteratively inquiries about various symptoms and antecedents and the patient responds with an appropriate response. The system repeats this until all the relevant symptoms and antecedents have been inquired and produces a differential diagnosis at the end of the interaction. The maximum number of turns are caped to T turns (i.e. 30 in all of our experiments).

For AARLC, we use the same setup as in [7] whereas, for the ASD approach, the agent is made of an MLPs with 2 hidden layers of size 2048.

4.2 Results

We report on the interaction length (IL). Also, to evaluate the evidence collection, we measure the recall (PER), precision (PEP), and F1 score (PEF1) of the inquired evidences. To evaluate the differential diagnosis, we calculate the recall (DDR@k), precision (DDP@k) and F1 score (DDF1@k) when considering the top k entries of the predicted distributions. We also compute the accuracy of inclusion of the gold truth pathology (i.e., the pathology a patient was simulated from) in the differential diagnosis (GTPA@k). Details of these metrics, including all formulas, can be found in Appendix 7.

Tables 3 and 4 show the results obtained for the two approaches. As it can be observed for both approaches, the differential helps collecting more evidences.

4.3 Conclusion and Future Work

In this work, we release a large scale benchmark dataset of 1 million patients suffering from pathologies that include cough, sore throat or breathing problems as symptoms. The dataset contains binary, categorical and multi-choice evidences.

We hope that this dataset will help the research community improve automatic diagnosis systems. We emphasize that this dataset should not be used to train and deploy a model prior to performing rigorous evaluations to ensure proper coverage and representativity of the populations that such a model will interact with. Based on this dataset, we extend several AD and ASD baselines. We extend two approaches (based on RL and non-RL settings) that reduces the interaction length, and improves evidence collection and differential diagnosis prediction.

In this work, we considered all diseases as equally important. But in general, when establishing a differential diagnosis, doctors ask questions to specifically explore and rule out severe pathologies. Our dataset has a severity flag associated with each pathology. We will therefore explore approaches that better handle severity. Extending our system to support uncertain answers from patients (e.g. "I don't know." or "I am not sure.") will also be an important next step.

References

[1] Zhongyu Wei, Qianlong Liu, Baolin Peng, Huaixiao Tou, Ting Chen, Xuan-Jing Huang, Kam-Fai Wong, and Xiang Dai. Task-oriented dialogue system for automatic diagnosis. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 201–207, 2018.

[2] Lin Xu, Qixian Zhou, Ke Gong, Xiaodan Liang, Jianheng Tang, and Liang Lin. End-to-end knowledge-routed relational dialogue system for automatic diagnosis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7346–7353, 2019.

[3] Junying Chen, Dongfang Li, Qingcai Chen, Wenxiu Zhou, and Xin Liu. Diaformer: Automatic diagnosis via symptoms sequence generation. *arXiv preprint arXiv:2112.10433*, 2021.

[4] Xinyan Zhao, Liangwei Chen, and Huanhuan Chen. A weighted heterogeneous graph-based dialog system. *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[5] Hong Guan and Chitta Baral. A bayesian approach for medical inquiry and disease inference in automated differential diagnosis. *arXiv preprint arXiv:2110.08393*, 2021.

[6] Wenge Liu, Yi Cheng, Hao Wang, Jianheng Tangi, Yafei Liu, Ruihui Zhao, Wenjie Li, Yefeng Zheng, and Xiaodan Liang. "my nose is running.""are you also coughing?": Building a medical diagnosis agent with interpretable inquiry logics. *arXiv preprint arXiv:2204.13953*, 2022.

[7] Hongyi Yuan and Sheng Yu. Efficient symptom inquiring and diagnosis via adaptive alignment of reinforcement learning and classification. *arXiv preprint arXiv:2112.00733*, 2021.

[8] Mark Henderson, Lawrence M Tierney, and Gerald W Smetana. *The patient history: Evidence-based approach*. McGraw Hill Professional, 2012.

[9] Yu-Shao Peng, Kai-Fu Tang, Hsuan-Tien Lin, and Edward Chang. Refuel: Exploring sparse features in deep reinforcement learning for fast disease diagnosis. *Advances in neural information processing systems*, 31:7322–7331, 2018.

[10] Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3:160035, 2016.

[11] Hao-Cheng Kao, Kai-Fu Tang, and Edward Chang. Context-aware symptom checking for disease diagnosis using hierarchical reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[12] AHEAD Research Inc. SymCAT: Symptom-based, Computer Assisted Triage. `http://www.symcat.com/`. Accessed: 2021-01-31.

[13] Dominik Aronsky, Diane Kendall, Kathleen Merkley, Brent James, and Peter Haug. A comprehensive set of coded chief complaints for the emergency department. *Academic emergency medicine : official journal of the Society for Academic Emergency Medicine*, 8:980–9, 11 2001.

[14] David Thompson, David Eitel, Christopher Fernandes, Jesse Pines, and James Amsterdam. Coded chief complaints-automated analysis of free-text complaints. *Academic emergency medicine : official journal of the Society for Academic Emergency Medicine*, 13:774–82, 08 2006.

[15] Jason Walonoski, Mark Kramer, Joseph Nichols, Andre Quina, Chris Moesel, Dylan Hall, Carlton Duffett, Kudakwashe Dube, Thomas Gallagher, and Scott McLachlan. Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. *Journal of the American Medical Informatics Association*, 25(3):230–238, 08 2017.

[16] US Government. 2010-2015 US Census Data. `https://www2.census.gov/programs-surveys/popest/datasets/2010-2015/`. Accessed: 2021-01-31.

[17] Hongyin Luo, Shang-Wen Li, and James R. Glass. Knowledge grounded conversational symptom detection with graph memory networks. In *CLINICALNLP*, 2020.

5 Appendix

5.1 Demographics statistics from census data

To be able to synthesize patients, one needs to have access to the prior distributions regarding the age, sex and race of the population of interest. In this work, we rely on the 2010-2015 US census data [16] from the state of New York. Table 5 describes the corresponding statistics.

Category	Probability
Sex	
Male	0.4836
Female	0.5164
Age	
Less than 1-year	0.0154
1-4-years	0.0461
5-14-years	0.1146
15-29-years	0.2132
30-44-years	0.2025
45-59-years	0.2042
60-74-years	0.1399
75-years and more	0.0641
Race	
White	0.5778
Hispanic	0.1658
Black	0.1638
Asian	0.0831
Native	0.0087
Other	0.0008

Table 5: The 2010-2015 census data of the state of New York.

In our implementation, because the *Synthea*^{TM} tool uses the "Hispanic" value to characterize the ethnicity of a synthesized patient instead of the race, we remove "Hispanic" from the set of possible race values and re-normalize the probabilities of each of the remained values.

5.2 Differential Diagnosis Post-Processing

As mentioned previously, pathologies that are part of the differential diagnosis generated by the rule-base platform but that do not belong to the specified chief complaint are excluded from the differential diagnosis. Table 6 describes, for each rank in the differential diagnosis, the proportion of patients for which the pathology at that rank is excluded from the differential diagnosis.

Rank	Proportion
1	1.09
2	11.23
3	15.73
4	13.78
5	14.53
6	11.92
7	13.12
8	10.47
9	9.23
10	8.16
11	9.47
12	7.00
13	5.92
14	6.11
15	5.49
16	5.86
17	5.42
18	4.52
19	3.67
20	3.37
21	3.07
22	2.37
23	2.14
24	1.70
25	1.17
26	0.71
27	0.42
28	0.24
29	0.13
30	0.07
31	0.04
32	0.03
33	0.02
34	0.01
35	4.49e-03
36	1.76e-03
37	3.90e-04

Table 6: Proportion (%) of the patients for which pathologies are excluded at each rank from the differential returned by the platform.

5.3 Analysis of the Train, Validation, and Test Subsets

5.3.1 Evidences Statistics

The statistics of the evidences experienced by the synthesized patients belonging to the train, validation, and test subsets are depicted in Tables 7, 8, and 9 respectively. As illustrated, the evidences are uniformly distributed across the three subsets.

	Evidences	**Symptoms**	**Antecedents**
Avg	13.42	9.97	3.45
Std dev	5.14	4.76	2.23
Min	1	1	0
1st quartile	10	7	2
Median	13	10	3
3rd quartile	17	12	5
Max	36	25	12

Table 7: The statistics of the synthesized patients on the train subset.

	Evidences	**Symptoms**	**Antecedents**
Avg	13.69	10.24	3.45
Std dev	5.05	4.65	2.23
Min	1	1	0
1st quartile	10	8	2
Median	13	10	3
3rd quartile	17	12	5
Max	34	25	12

Table 8: The statistics of the synthesized patients on the validation subset.

	Evidences	**Symptoms**	**Antecedents**
Avg	13.66	10.19	3.46
Std dev	5.05	4.67	2.23
Min	1	1	0
1st quartile	10	8	2
Median	13	10	3
3rd quartile	17	12	5
Max	35	25	12

Table 9: The statistics of the synthesized patients on the test subset.

5.3.2 Pathology Statistics

Figures 5, 6, and 7 depict the histograms of the pathology experienced by the synthesized patients in the train, validation, and test subsets. As illustrated, the pathologies are evenly distributed across the three subsets.

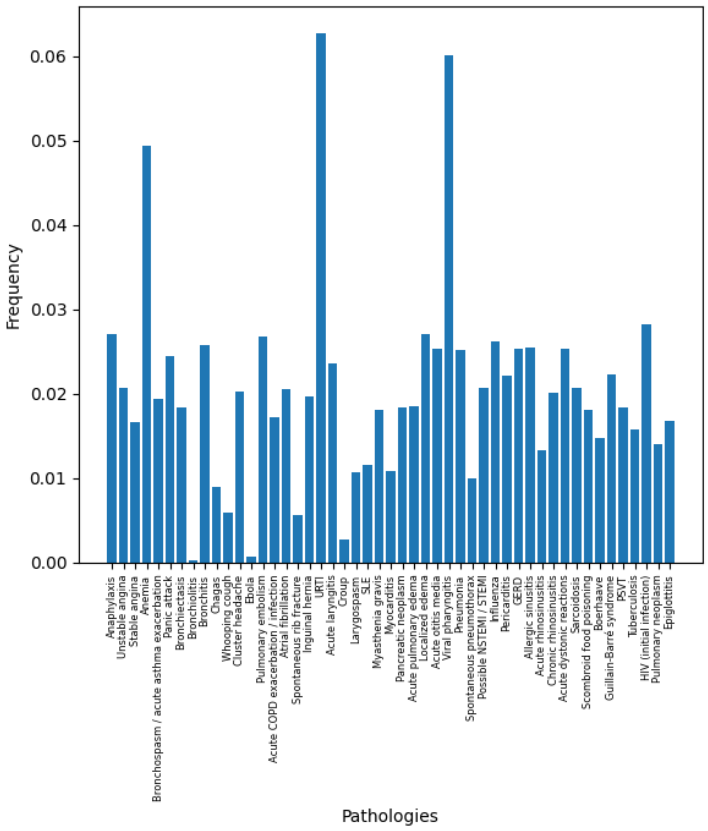


Figure 5: Histogram of the patient pathologies in the train subset.

5.3.3 Socio-Demographic Statistics

We depicted in this section the statistics regarding the socio-demographic data of the synthesized patients from the train (cf. Figure 8), validation (cf. Figure 9), and test (cf. Figure 10) subsets. In all the three figures, the sex distribution is shown in plot (a) while the age and the race distributions are respectively depicted in plot (b) and plot (c).

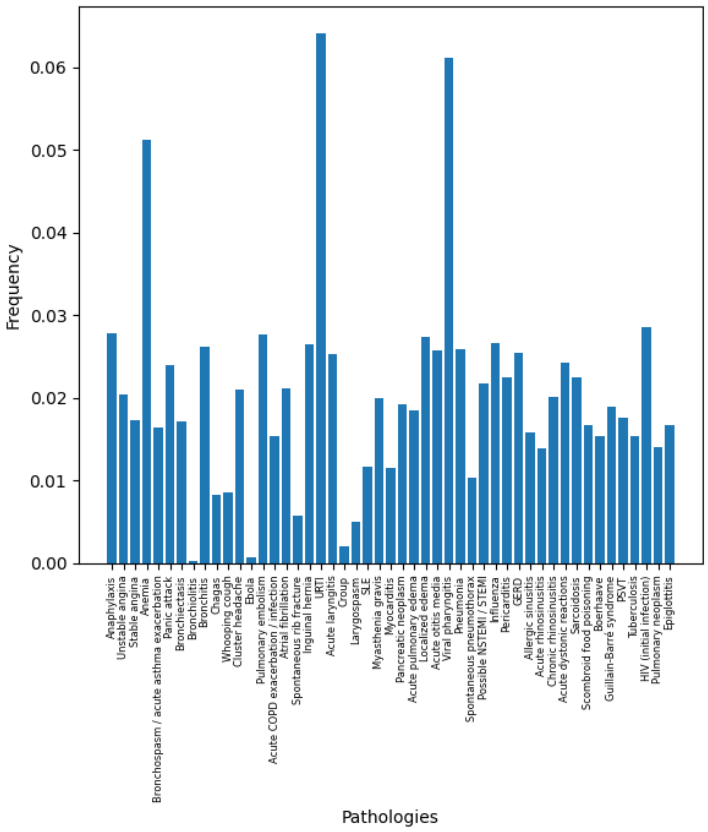


Figure 6: Histogram of the patient pathologies in the valid subset.

5.3.4 Differential Diagnosis Statistics

The histograms regarding the length of the differential diagnosis from the train, validation, and test subsets ares respectively illustrated in Figures 11, 13, and 15. Similarly, the histograms regarding the rank of the simulated pathology within the differential diagnosis from the train, validation, and test subsets ares respectively depicted in Figures 12, 14, and 16.

6 ASD Approach

The ASD agent consists on an MLP network with 2 prediction branches:

- a policy branch whose role is to predict whether to stop or continue the interaction, and if the latter, what evidence to inquire about next;
- a classifier branch to make prediction regarding the underlying patient

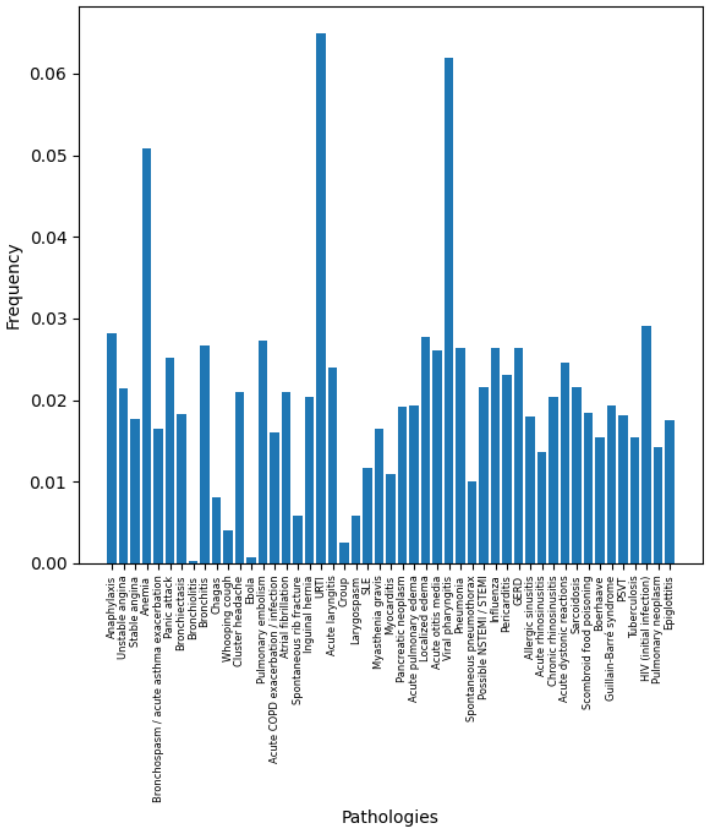
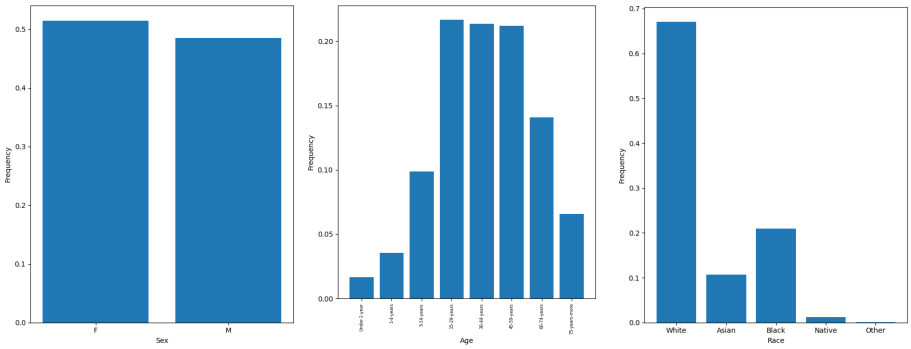
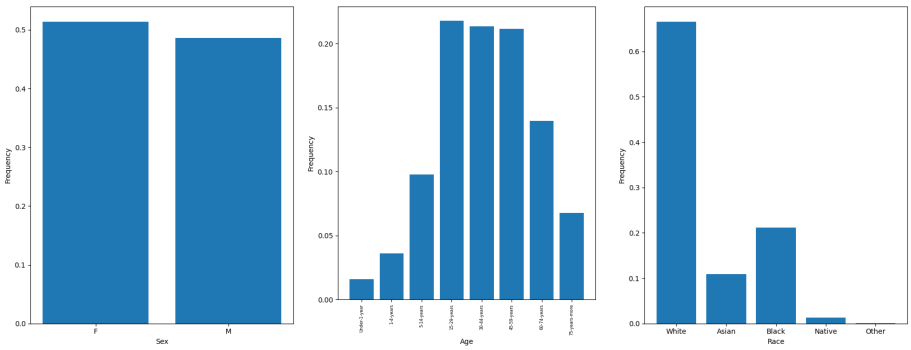


Figure 7: Histogram of the patient pathologies in the test subset.



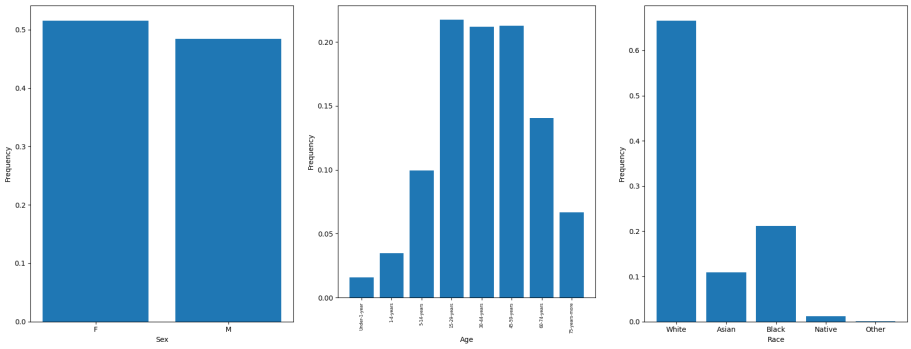
(a) (b) (c)

Figure 8: The socio-demographic statistics of the patients from the train subset.



(a) (b) (c)

Figure 9: The socio-demographic statistics of the patients from the validation subset.



(a) (b) (c)

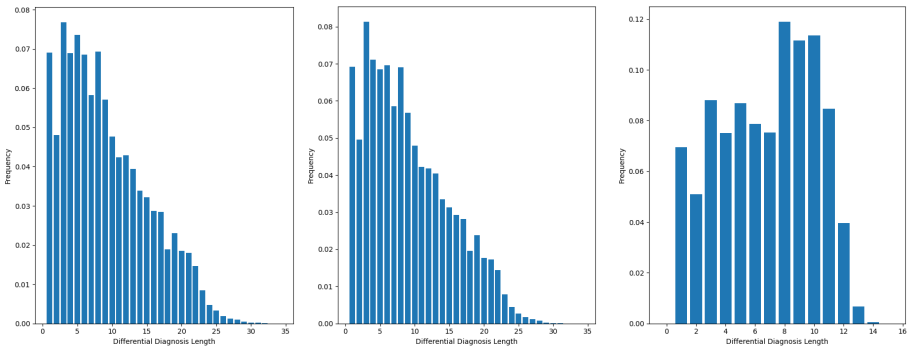
Figure 10: The socio-demographic statistics of the patients from the test subset.

disease.

To train the network, we simulate dialogue states together with their target values. In other words, let us assume a given patient has n evidences that he/she is experiencing. We simulate a dialogue state state as follows:

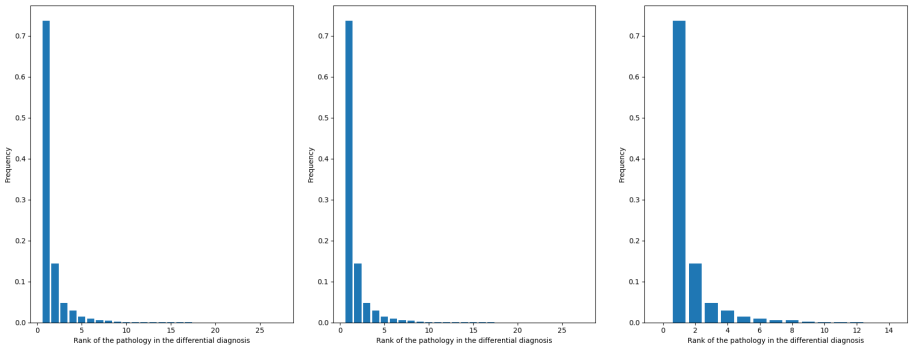
1. Randomly select $p \in [1, n]$ representing the number of positive evidences already acquired. Sample p evidences from the ones experienced by the patient and set them in the simulated dialog state.

2. Randomly select $q \in [0, T - p)$ representing the number of negative evidences already inquired where T is the maximum allowed dialog turns. Sample q evidences from the ones not experienced by the patient and set them in the simulated dialog state.



(a) (b) (c)

Figure 11: Statistics regarding the length of the differential diagnosis in the train subset. a) Full differential as generated by the system. b) Pathologies with probability mass less than or equals to 0.01 are filtered out. c) Same as (b) but the threshold is set to 0.05.



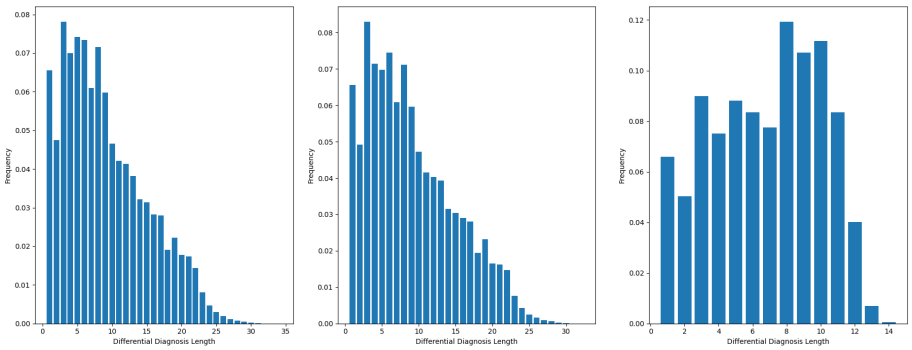
(a) (b) (c)

Figure 12: Statistics regarding the rank of the simulated pathology within the differential diagnosis in the train subset. a) Full differential as generated by the system. b) Pathologies with probability mass less than or equals to 0.01 are filtered out. c) Same as (b) but the threshold is set to 0.05.

3. If $p = n$, set the target of the policy branch to "stop"; otherwise set the target to be one of the experienced evidences that was not sampled at step 1).

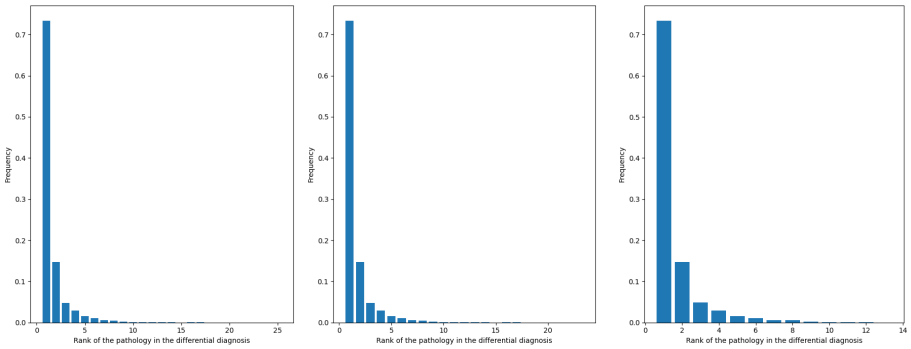
4. Set the classifier branch target to be the ground-truth pathology.

Both branches are trained using the cross-entropy loss function and the classifier branch is only updated when the target of the policy branch is set to "stop".



(a) (b) (c)

Figure 13: Statistics regarding the length of the differential diagnosis in the validation subset. a) Full differential as generated by the system. b) Pathologies with probability mass less than or equals to 0.01 are filtered out. c) Same as (b) but the threshold is set to 0.05.

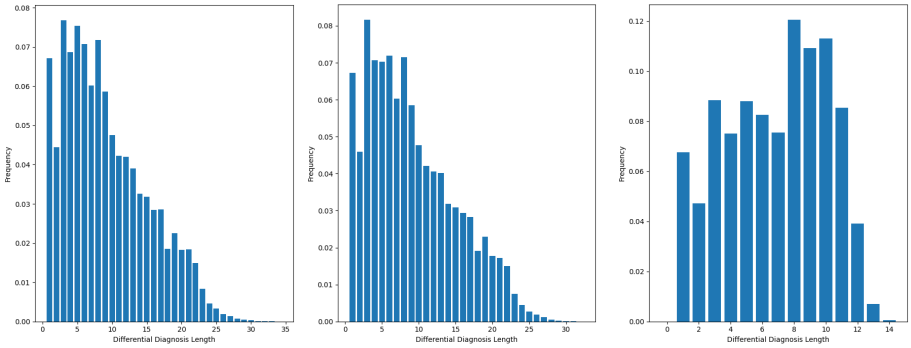


(a) (b) (c)

Figure 14: Statistics regarding the rank of the simulated pathology within the differential diagnosis in the validation subset. a) Full differential as generated by the system. b) Pathologies with probability mass less than or equals to 0.01 are filtered out. c) Same as (b) but the threshold is set to 0.05.

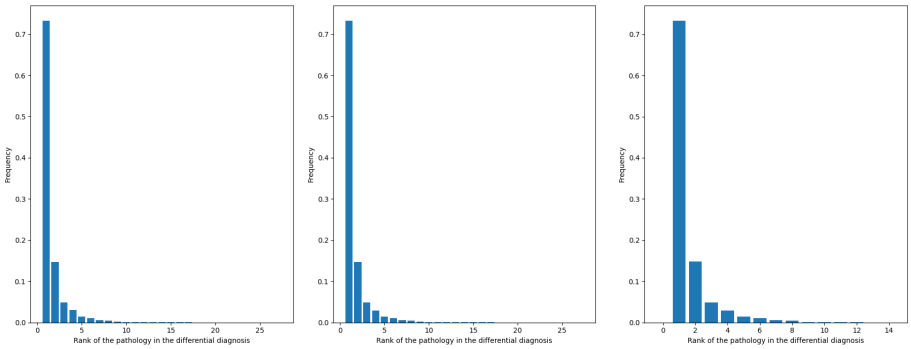
7 Evaluation Metric

This section describes the metrics used to evaluate the implemented agents. Let D be the number of patients. And let T^i be the set of evidences collected by an agent from the i^{th} patient, Y^i be the ground truth differential, and $\hat{Y}^i$ be the pathology distribution generated by the agent for that patient.



(a) (b) (c)

Figure 15: Statistics regarding the length of the differential diagnosis in the test subset. a) Full differential as generated by the system. b) Pathologies with probability mass less than or equals to 0.01 are filtered out. c) Same as (b) but the threshold is set to 0.05.



(a) (b) (c)

Figure 16: Statistics regarding the rank of the simulated pathology within the differential diagnosis in the test subset. a) Full differential as generated by the system. b) Pathologies with probability mass less than or equals to 0.01 are filtered out. c) Same as (b) but the threshold is set to 0.05.

Interaction Length (IL): The average interaction length is defined as:

$$IL = \frac{1}{|D|}\Sigma_{i=1}^{|D|}|T^i|. \tag{1}$$

Differential Diagnosis Recall for top k entries (DDR@k): At the end of an interaction, the agent produces a distribution $\hat{Y}$ over the pathologies. We measure whether the top k pathologies $\hat{Y}_k$ extracted from this distribution are also present in the top k pathologies Y_k of the ground truth differential (if

the ground truth differential has less than k pathologies, then we set k to the size of the ground truth differential). The metric is defined as:

$$DDR@k = \frac{1}{|D|}\Sigma_{i=1}^{|D|}\frac{|\hat{Y}_k^i \cap Y_k^i|}{|Y_k^i|}, \tag{2}$$

where $\hat{Y}_k^i$ and Y_k^i are respectively the predicted and gold truth top k pathologies for the i^{th} patient.

Differential Diagnosis Precision for top k entries (DDP@k): This metric measures the precision of the differential diagnoses predicted by an agent when considering the top-k entries:

$$DDP@k = \frac{1}{|D|}\Sigma_{i=1}^{|D|}\frac{|\hat{Y}_k^i \cap Y_k^i|}{|\hat{Y}_k^i|}. \tag{3}$$

Differential Diagnosis F1 for top k entries (DDF1@k): We combine the aforementioned recall and precision metrics to compute the F1 score of the differential diagnosis.

Gold Truth Pathology Accuracy (GTPA@k): This metric measures whether the differential diagnosis predicted by an agent contains the pathology p a patient was simulated from within its top-k entries:

$$GTPA@k = \frac{1}{|D|}\Sigma_{i=1}^{|D|}\mathbb{1}_{p^i \in \hat{Y}_k^i}. \tag{4}$$

Positive Evidence Recall (PER): Let us suppose that the i^{th} patient is experiencing the set $\mathcal{S}^i$ of symptoms and the set $\mathcal{A}^i$ of antecedents. Also let us assume that the agent inquires the set $\hat{\mathcal{S}}^i$ (resp. $\hat{\mathcal{A}}^i$) of symptoms (resp. antecedents) from which $\hat{\mathcal{S}}_+^i \subseteq \mathcal{S}^i$ (resp. $\hat{\mathcal{A}}_+^i \subseteq \mathcal{A}^i$) is the set of symptoms (resp. antecedents) experienced by the i^{th} patient. Then, the recall for the positive evidences is calculated as:

$$PER = \frac{1}{|D|}\Sigma_{i=1}^{|D|}PER_i, \qquad \text{where } PER_i = \frac{|\hat{\mathcal{S}}_+^i \cup \hat{\mathcal{A}}_+^i|}{|\mathcal{S}^i \cup \mathcal{A}^i|}. \tag{5}$$

Positive Evidence Precision (PEP): Similarly to the recall metric, we compute the precision of inquired evidences. We have:

$$PEP = \frac{1}{|D|}\Sigma_{i=1}^{|D|}PEP_i, \qquad \text{with } PEP_i = \frac{|\hat{\mathcal{S}}_+^i \cup \hat{\mathcal{A}}_+^i|}{|T^i|}. \tag{6}$$

Positive Evidence F1 (PEF1): We combine the aforementioned recall and precision metrics to compute the F1 score of the inquired evidences.

$$PEF1 = \frac{1}{|D|}\Sigma_{i=1}^{|D|}2\frac{PEP_i \times PER_i}{PEP_i + PER_i}. \tag{7}$$