

## A Additional information on the KB and rule-based AD system used to generate DDXPlus

Doctors reviewed over 2 years of relevant papers on the diseases used to create the KB. The papers along with the medical experts' knowledge and experience were used to extract the typical and atypical presentations of diseases, along with the relevant symptom and antecedent distributions to build accurate disease models based on geography, demography and baseline patient characteristics. The process was exhaustive and independently validated by the doctors, where agreement was sought for the presentation of every disease in the database. This process therefore uses the clinician experience to ensure that the diseases are accurately depicted across their usual and unusual presentations.

Some of the symptoms and antecedents in the KB were defined to be categorical or multi-choice to ensure efficient and proper coverage of important evidence. Since almost 60% of all presentations in real life will include pain, a great level of attention was given to the description of pain for each disease where this symptom can be found (Cordell et al., 2002; Mura et al., 2017). This feature includes the localization, radiation, intensity, subjective characterization of common pain description by patients, a precision feature (very small area to diffuse) and a rapidity of onset. The pain symptom encompasses all these sub-features and was created using domain experts and medical journal articles looking at disease presentations. This would be extremely hard to derive from another dataset, since we have not seen other ones cover the pain description extensively, although one of the most important symptoms a clinician will spend a good amount of time defining clearly with the patient. A similar process of specification was used for skin rashes, with the rash description including the usual dermatological lesion characterizations used by clinicians when evaluating a patient.

The rule-based AD system is a statistical engine that uses the patient's response to generate a differential diagnosis in real-time. The engine has phases, where the first phase seeks to ask questions that have the highest probability of ruling out the most diseases in the initial differential. Subsequently, the engine seeks answers to questions linked to diseases in the differential that represent the highest risk in terms of mortality and morbidity. Finally, the engine seeks to ask specific questions about the personal risk factors and antecedents for the top 5 diseases in the estimated differential.

The engine was built for primary care and acute care settings, having in mind the goal of gathering a medical history that is as close as possible to the one clinicians would gather when evaluating a patient.

The engine was tested on real patients in an acute primary care setting and the collected history was evaluated by doctors on a scale of pertinence and completeness. The differential of the engine was compared to the clinician's differential, who evaluated the patient with the usual clinical flow, blinded to the evidence collected by the engine initially.

## B Demographic statistics from census data

To synthesize patients, one needs to have access to the prior distributions of the age and sex of a population of interest. In this work, we rely on the 2010-2015 US census data US Government (2015) for the state of New York. Table 4 describes the corresponding statistics.

Table 4: The 2010-2015 census data of the state of New York.

	Category	Frequency
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

## C Geographical regions

The following geographical regions are covered in the dataset: North Africa, West Africa, South Africa, Central America, North America, South America, Asia, South East Asia, The Caribbean, Europe, Oceania.

## D Missing incidence rates

The 8 pathologies whose incidence rates were not present for all the combinations of age, sex, and geographical regions are: Anemia, Inguinal hernia, Anaphylaxis, Allergic sinusitis, Chagas, Tuberculosis, Ebola, Chronic rhinosinusitis. Some are very rare infectious diseases that follow an epidemic pattern which varies over time periods, making it hard to track incidence. Others are very mild diseases and are not tracked as they represent no public health interest.

## E Differential diagnosis post-processing

As mentioned previously, pathologies that are part of the differential diagnosis generated by the rule-based AD platform and that do not belong to the set of the 49 considered pathologies are excluded from the differential diagnosis. Table 5 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.

## F Subset data analysis

### F.1 Evidence statistics

The statistics of the evidences experienced by the synthesized patients belonging to the train, validation, and test subsets are presented in Tables 6, 7, and 8 respectively. As illustrated, the evidences are similarly distributed across the three subsets.

### F.2 Pathology statistics

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

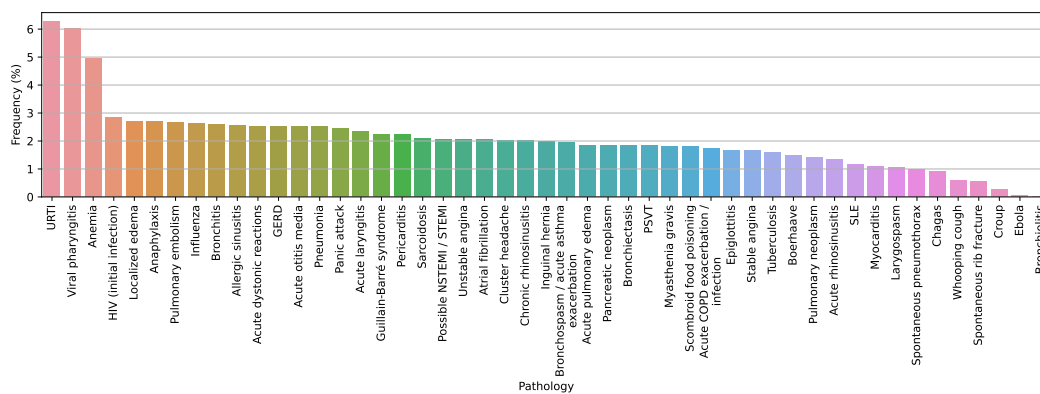


Figure 5: Histogram of the patient pathologies in the training set.

### F.3 Socio-demographic statistics

The socio-demographic statistics of the synthesized patients from the train, validation and test subsets are respectively illustrated in Figures 8, 9, and 10.

Table 5: Proportion (%) of the patients for which pathologies are excluded at each rank from the differential diagnosis returned by the rule-based AD platform.

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: Statistics describing the number of evidences of the synthesized patients for the training set.

	Evidences	Symptoms	Antecedents
Avg	13.52	10.03	3.49
Std dev	5.08	4.71	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 7: Statistics describing the number of evidences of the synthesized patients for the validation set.

	Evidences	Symptoms	Antecedents
Avg	13.76	10.27	3.49
Std dev	5.01	4.61	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: Statistics describing the number of evidences of the synthesized patients for the test set.

	Evidences	Symptoms	Antecedents
<b>Avg</b>	13.72	10.23	3.50
<b>Std dev</b>	5.02	4.63	2.23
<b>Min</b>	1	1	0
<b>1st quartile</b>	10	8	2
<b>Median</b>	13	10	3
<b>3rd quartile</b>	17	12	5
<b>Max</b>	35	25	12

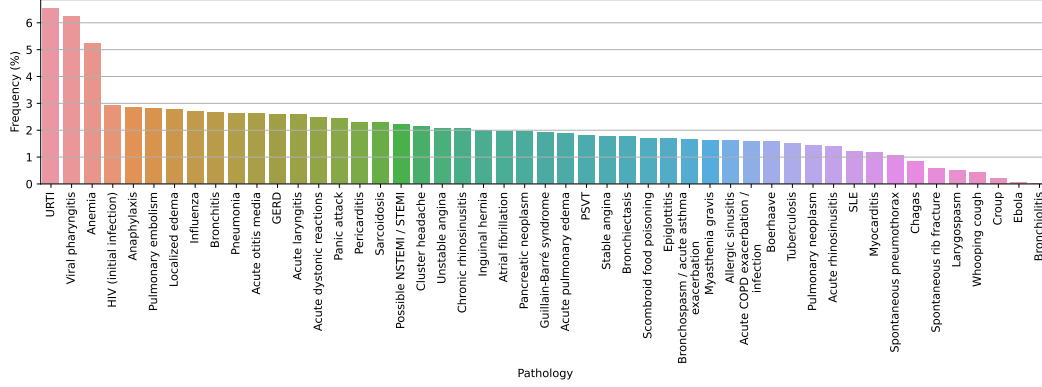


Figure 6: Histogram of the patient pathologies in the validation set.

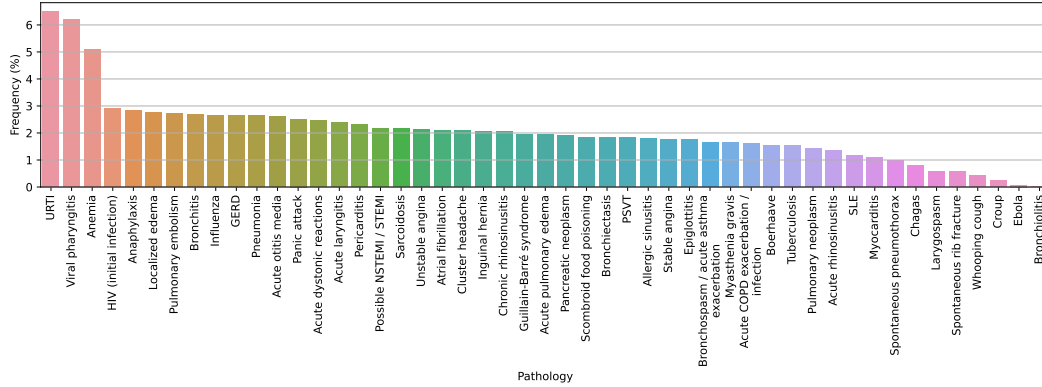


Figure 7: Histogram of the patient pathologies in the test set.

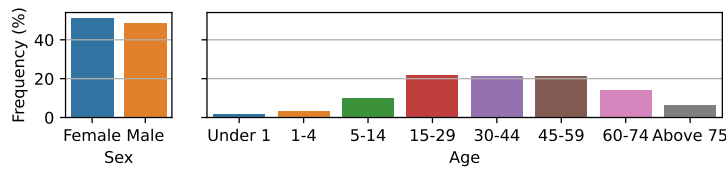


Figure 8: The socio-demographic statistics of the synthesized patients from the training set.

#### F.4 Differential diagnosis statistics

The histograms of the length of the differential diagnosis from the train, validation, and test subsets are respectively illustrated in Figures 11 (left), 12 (left), and 13 (left). Similarly, the histograms of the rank of the simulated pathology within the differential diagnosis from the train, validation, and test subsets are respectively depicted in Figures 11 (right), 12 (right), and 13 (right).

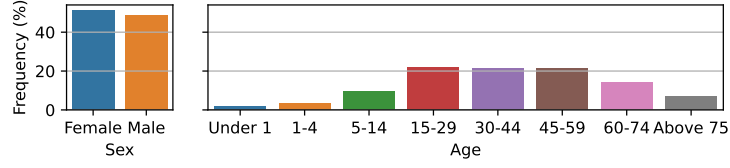


Figure 9: The socio-demographic statistics of the synthesized patients from the validation set.

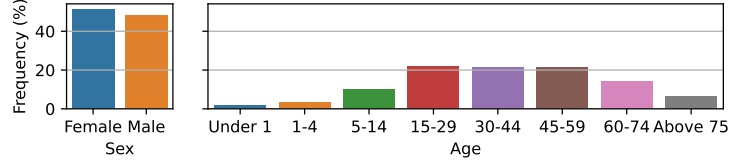


Figure 10: The socio-demographic statistics of the synthesized patients from the test set.

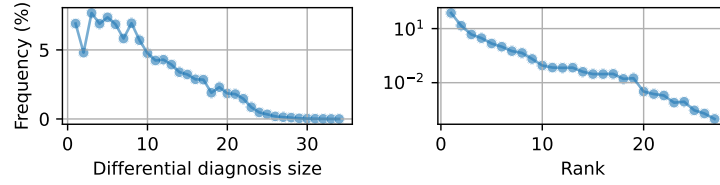


Figure 11: Statistics regarding (left) the length of the differential diagnosis, and (right) the rank of the patient's simulated pathology within the differential diagnosis (y-axis on log scale), in the training set.

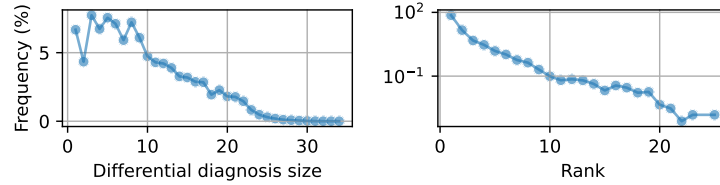


Figure 12: Statistics regarding (left) the length of the differential diagnosis, and (right) the rank of the patient's simulated pathology within the differential diagnosis (y-axis on log scale), in the validation set.

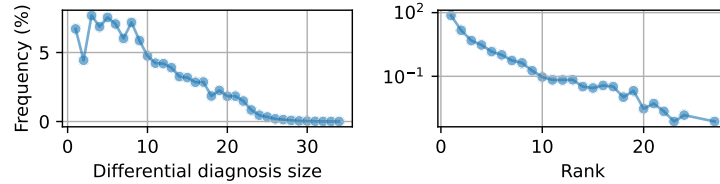


Figure 13: Statistics regarding (left) the length of the differential diagnosis, and (right) the rank of the patient's simulated pathology within the differential diagnosis (y-axis on log scale), in the test set.

## G Baselines

The code to reproduce the main results from Table 3 can be found at <https://github.com/bruzwen/ddxplus>.

## G.1 AARLC

AARLC (short for Adaptive Alignment of Reinforcement Learning and Classification) (Yuan and Yu, 2021) is a model that has two branches, an evidence acquisition branch, trained using RL, whose goal is to determine the next evidence to inquire about, and a classifier branch trained in a supervised way to predict the patient’s disease. An adaptive approach is used to align the tasks performed by the two branches using the entropy of the distributions predicted by the classifier branch.

To train AARLC with differentials, we make several changes, in addition to replacing the ground truth pathology with the ground truth differential probabilities as the classifier’s training objective:

- Instead of updating the stopping threshold  $K$  when the predicted pathology matches the ground truth pathology, we update it when the set  $D^i$  of diseases in the ground truth differential is identical to the set of top- $|D^i|$  predicted diseases. We make this change because AARLC designs the threshold to be updated when the predicted disease is correct, and therefore if the differential is replacing the ground truth pathology as the target, it should also replace it as the standard of correctness.
- Second, now that the agent does not focus on predicting one single disease, it is no longer reasonable to only update the threshold associated with one disease. Therefore, we instead use one global threshold that is not associated with any particular disease, and update it every time the aforementioned condition is met.
- Similar to the condition of updating the threshold, we change the condition under which a positive reward is given to the agent, as part of  $r_p$ , for making the correct diagnosis. We give the positive reward when the set  $D^i$  of diseases in the ground truth differential is identical to the set of top- $|D^i|$  predicted diseases.

## G.2 BASD approach

The BASD (short for Baseline ASD) agent consists of 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 a prediction regarding the underlying patient disease.

To train the network, we simulate dialogue states together with their target values. Assuming that a given patient has  $n$  evidences that they are experiencing, we simulate a dialogue state as follows:

1. Randomly select  $p \in [1, n]$  representing the number of positive evidences already inquired about. 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 number of allowed dialog turns. Sample  $q$  evidences from the ones not experienced by the patient and set them in the simulated dialog state.
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 or the ground truth differential.

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".

## H Evaluation metrics

This section describes the metrics used to evaluate the performance of trained agents. The differentials generated by the trained models as well as the ground truth differentials are post-processed to remove pathologies whose mass is less than or equal to 0.01. This threshold, approved by our collaborating physician, is selected to reduce the size of the differentials by removing highly unlikely pathologies. Let  $D$  be the number of patients,  $T^i$  be the set of evidences collected by an agent from the  $i^{th}$  patient

(including the first evidence provided by the patient),  $Y^i$  be the ground truth differential, and  $\hat{Y}^i$  be the pathology distribution generated by the agent for that patient.

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

$$IL = \frac{1}{|D|} \sum_{i=1}^{|D|} |T^i|. \quad (4)$$

**Differential diagnosis recall (DDR):** This metric measures the recall of the differential diagnosis predicted by the agent with respect to the ground truth differential:

$$DDR = \frac{1}{|D|} \sum_{i=1}^{|D|} \frac{|\hat{Y}^i \cap Y^i|}{|Y^i|}. \quad (5)$$

**Differential diagnosis precision (DDP):** This metric measures the precision of the differential diagnosis predicted by an agent with respect to the ground truth differential:

$$DDP = \frac{1}{|D|} \sum_{i=1}^{|D|} \frac{|\hat{Y}^i \cap Y^i|}{|\hat{Y}^i|}. \quad (6)$$

**Differential diagnosis F1 (DDF1):** We combine the DDR and DDP metrics to compute the F1 score of the differential diagnosis.

**Ground truth pathology accuracy (GTPA@ $k$  and GTPA):** The GTPA@ $k$  metric measures whether the differential diagnosis predicted by an agent contains the pathology  $p$  a patient was simulated from within its top- $k$  entries  $\hat{Y}_k^i$ :

$$GTPA@k = \frac{1}{|D|} \sum_{i=1}^{|D|} \mathbb{1}_{\hat{Y}_k^i}(p^i), \quad (7)$$

where  $\mathbb{1}$  is the indicator function.

Similarly, the GTPA metric measures whether the differential diagnosis predicted by an agent contains the pathology  $p$  a patient was simulated from:

$$GTPA = \frac{1}{|D|} \sum_{i=1}^{|D|} \mathbb{1}_{\hat{Y}^i}(p^i), \quad (8)$$

where  $\mathbb{1}$  is still the indicator function.

**Positive evidence recall (PER):** Let us suppose that the  $i^{th}$  patient is experiencing the set  $S^i$  of symptoms and the set  $\mathcal{A}^i$  of antecedents. Also let us assume that the agent inquires the set  $\hat{S}^i$  (resp.  $\hat{\mathcal{A}}^i$ ) of symptoms (resp. antecedents) from which  $\hat{S}_+^i \subseteq S^i$  (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:

$$PER = \frac{1}{|D|} \sum_{i=1}^{|D|} PER_i, \quad \text{where } PER_i = \frac{|\hat{S}_+^i \cup \hat{\mathcal{A}}_+^i|}{|S^i \cup \mathcal{A}^i|}. \quad (9)$$

While collecting positive evidence is important, it is not sufficient as this evidence paints an incomplete picture to the clinical team. Clinicians want to make sure other pathologies were considered and thoroughly evaluated during their interaction with the patient. Those include severe pathologies, less prevalent pathologies, and pathologies which are similar to the one the patient is suffering from but are very different in their management. As such, clinicians inquire about other evidences and ensure they are not present. We therefore do not report the evidence precision as a metric. Defining metrics for evaluating questions about negative evidence is left for future work.

## I Dataset release

We are releasing under the CC-BY licence a new large-scale dataset for Automatic Symptom Detection (ASD) and Automatic Diagnosis (AD) systems in the medical domain. The dataset

contains patients synthesized using a proprietary medical knowledge base and a commercial rule-based AD system. Patients in the dataset are characterized by their socio-demographic data, a pathology they are suffering from, a set of symptoms and antecedents related to this pathology, and a differential diagnosis. The symptoms and antecedents can be binary, categorical and multi-choice, with the potential of leading to more efficient and natural interactions between ASD/AD systems and patients. Moreover, some symptoms are organized in a hierarchy, making it possible to design systems able to interact with patients in a logical way. Finally, each disease is characterized by its level of severity. To the best of our knowledge, this is the first large-scale dataset that includes the differential diagnosis, and non-binary symptoms and antecedents. The dataset can be downloaded at [https://figshare.com/articles/dataset/DDXPlus\\_Dataset/20043374](https://figshare.com/articles/dataset/DDXPlus_Dataset/20043374).

## I.1 Dataset documentation

In what follows, we use the term “evidence” as a general term to refer to a symptom or an antecedent. The dataset contains the following files:

- `release_evidences.json`: a JSON file describing all possible evidences considered in the dataset.
- `release_conditions.json`: a JSON file describing all pathologies considered in the dataset.
- `release_train_patients.zip`: a CSV file containing the patients of the training set.
- `release_validate_patients.zip`: a CSV file containing the patients of the validation set.
- `release_test_patients.zip`: a CSV file containing the patients of the test set.

**Evidence description** Each evidence in the `release_evidences.json` file is described using the following entries:

- `name`: name of the evidence.
- `code_question`: a code allowing to identify which evidences are related. Evidences having the same `code_question` form a group of related evidences. The value of the `code_question` refers to the evidence that needs to be simulated/activated for the other members of the group to be eventually simulated.
- `question_fr`: the query, in French, associated to the evidence.
- `question_en`: the query, in English, associated to the evidence.
- `is_antecedent`: a flag indicating whether the evidence is an antecedent or a symptom.
- `data_type`: the type of the evidence. We use “B” for binary, “C” for categorical, and “M” for multi-choice.
- `default_value`: the default value of the evidence. If this value is used to characterize the evidence, then it is as if the evidence was not synthesized.
- `possible-values`: the possible values for the evidence. Only valid for categorical and multi-choice evidences.
- `value_meaning`: The meaning, in French and English, of each code that is part of the `possible-values` field. Only valid for categorical and multi-choice evidences.

**Pathology description** The file `release_conditions.json` contains information about the pathologies patients in the datasets may suffer from. Each pathology has the following attributes:

- `condition_name`: name of the pathology.
- `cond-name-fr`: name of the pathology in French.
- `cond-name-eng`: name of the pathology in English.
- `icd10-id`: ICD-10 code of the pathology.
- `severity`: the severity associated with the pathology. The lower the more severe.



- **symptoms:** data structure describing the set of symptoms characterizing the pathology. Each symptom is represented by its corresponding name entry in the `release_evidences.json` file.
- **antecedents:** data structure describing the set of antecedents characterizing the pathology. Each antecedent is represented by its corresponding name entry in the `release_evidences.json` file.

**Patient description** Each patient in each of the 3 sets has the following attributes:

- **AGE:** the age of the synthesized patient.
- **SEX:** the sex of the synthesized patient.
- **PATHOLOGY:** name of the ground truth pathology (`condition_name` property in the `release_conditions.json` file) that the synthesized patient is suffering from.
- **EVIDENCES:** list of evidences experienced by the patient. An evidence can be either binary, categorical or multi-choice. A categorical or multi-choice evidence is represented in the format `[evidence-name]_@[evidence-value]` where `[evidence-name]` is the name of the evidence (name entry in the `release_evidences.json` file) and `[evidence-value]` is a value from the `possible-values` entry. Note that for a multi-choice evidence, it is possible to have several `[evidence-name]_@[evidence-value]` items in the evidence list, with each item being associated with a different evidence value. A binary evidence is simply represented as `[evidence-name]`.
- **INITIAL\_EVIDENCE:** the evidence provided by the patient to kick-start an interaction with an ASD/AD system. This is useful during model evaluation for a fair comparison of ASD/AD systems as they will all begin an interaction with a given patient from the same starting point. The initial evidence is randomly selected from the binary evidences found in the evidence list mentioned above (i.e., **EVIDENCES**) and it is part of this list.
- **DIFFERENTIAL\_DIAGNOSIS:** the ground truth differential diagnosis for the patient. It is represented as a list of pairs of the form `[[patho_1, proba_1], [patho_2, proba_2], ...]` where `patho_i` is the pathology name (`condition_name` entry in the `release_conditions.json` file) and `proba_i` is its related probability.

## I.2 General notes about the dataset’s differentials

It is important to understand that the level of specificity, sensitivity and confidence that a physician will seek when evaluating a patient will be influenced by the clinical setting. The dataset was built for acute care and biased toward high mortality and morbidity pathologies. Physicians will tend to consider negative evidences as equally important in such a clinical context in order to evaluate high acuity diseases.

In the creation of the DDXPlus dataset, a small subset of the diseases was chosen to establish a baseline. Medical professionals have to consider this very important point when reviewing the results, as the differential is considerably smaller. A smaller differential means less potential evidences to collect. It is thus essential to understand this point when we look at the differential produced and the collected evidence.

## I.3 Responsibility statement

The authors declare that they bear all responsibility for violations of rights related to this dataset.

## J Dataset samples

We showcase in this section some samples from the DDXPlus dataset. For each example, we show the age, sex, the geographical region, the ground truth pathology, the symptoms, the antecedents (past medical history), as well as the corresponding differential diagnosis of a synthetic patient. The first two samples include feedback provided by a doctor to explain the differential diagnosis.

## Sample 1

Sex: M, Age: 47  
Geographical region: North America  
Pathology: PSVT  
Symptoms:

- 
- I feel pain.
- The pain is:
  - \* tugging
  - \* burning
- The pain locations are:
  - \* back of head
  - \* top of the head
  - \* temple(R)
- On a scale of 0-10, the pain intensity is 4
- On a scale of 0-10, the pain's location precision is 8
- On a scale of 0-10, the pace at which the pain appear is 5
- I feel like I am about to faint.
- I feel lightheaded and dizzy.
- I feel palpitations.

Antecedents:

- 
- I feel anxious.
- I regularly drink coffee or tea.
- I regularly consume energy drinks.
- I regularly take stimulant drugs.
- I have recently taken decongestants or substances that may have stimulant effects.

Differential diagnosis:

-----  
PSVT: 0.226, Anemia: 0.161, Panic attack: 0.142, Atrial fibrillation: 0.117,  
Anaphylaxis: 0.113, Cluster headache: 0.092, Chagas: 0.072, Scombroid food  
poisoning: 0.071, HIV (initial infection): 0.006

The differential is good, in the sense that we would need to ask additional questions on causes of palpitations, ranging from mood disorders, or search for causes for anemia. Although this is a patient with PSVT with many risk factors for it, a clinician will want to search for all the possibilities that would change the course of the care episode. It is important to note that the diagnosis space is significantly smaller than the frame of reference of a physician, hence doctors will tend to refine the medical history and seek to discriminate further in order to establish a final differential.

## Sample 2

Sex: F, Age: 55  
Geographical region: North America  
Pathology: GERD  
Symptoms:

- 
- I feel pain.
- The pain is:
  - \* haunting
  - \* tugging
  - \* sickening
- The pain locations are:
  - \* lower chest
  - \* upper chest
  - \* epigastric
- On a scale of 0-10, the pain intensity is 4
- The pain radiates to these locations:
  - \* lower chest
  - \* upper chest
- On a scale of 0-10, the pain's location precision is 6
- On a scale of 0-10, the pace at which the pain appear is 2

- I have a burning sensation that starts in my stomach then goes up into my throat, and can be associated with a bitter taste in my mouth.
- I am coughing.
- I have symptoms that get worse after eating.
- My symptoms worse when lying down and alleviated while sitting up.

Antecedents:

- I am significantly overweight compared to people of the same height as me.
- I drink alcohol excessively.
- I smoke cigarettes.
- I have a hiatal hernia.
- I have had to use a bronchodilator in the past.
- I am pregnant.

Differential diagnosis:

GERD: 0.196, Bronchitis: 0.148, Pericarditis: 0.129, Spontaneous rib fracture: 0.098, Unstable angina: 0.098, Boerhaave: 0.093, Possible NSTEMI / STEMI: 0.068, Tuberculosis: 0.060, Stable angina: 0.055, Pancreatic neoplasm: 0.054

This is a good differential based on the positive features. This differential includes diseases we would further need to inquire about as most are very relevant to the patient's demographic. In this example, we know that acute presentation of myocardial infarction in women can be atypical. Although in the differential, we would like to refine the presentation to better grasp the disease probabilities and tailor the tests to that risk level.

### Sample 3

Sex: F, Age: 66  
Geographical region: North America  
Pathology: URTI  
Symptoms:

- I have had significantly increased sweating.
- I feel pain.
- The pain is:
  - \* tedious
  - \* sensitive
- The pain locations are:
  - \* cheek(R)
  - \* cheek(L)
  - \* occiput
  - \* temple(L)
- On a scale of 0-10, the pain intensity is 8
- On a scale of 0-10, the pain's location precision is 5
- On a scale of 0-10, the pace at which the pain appeared is 1
- I have fever.
- I have a sore throat.
- I have diffuse muscle pain.
- I am coughing.

Antecedents:

- I have been in contact with a person with similar symptoms in the past 2 weeks.
- I live with 4 or more people.
- I smoke cigarettes.
- I am exposed to secondhand cigarette smoke on a daily basis.

Differential diagnosis:

URTI: 0.208, Influenza: 0.204, Bronchitis: 0.186, Pneumonia: 0.146, Tuberculosis: 0.121, HIV (initial infection): 0.078, Chagas: 0.057

### Sample 4

Sex: M, Age: 69  
Geographical region: North America  
Pathology: Anaphylaxis  
Symptoms:

- 
- I have eaten something that I have an allergy to.
  - I feel pain.
  - The pain is:
    - \* a knife stroke
    - \* sharp
  - The pain locations are:
    - \* flank(L)
    - \* pubis
    - \* belly
    - \* epigastric
  - On a scale of 0-10, the pain intensity is 6
  - On a scale of 0-10, the pace at which the pain appeared is 4
  - I have rashes.
  - The color of the rash is pink
  - On a scale of 0-10, the rash swelling is 6
  - The regions affected by the rash are:
    - \* biceps(R)
    - \* biceps(L)
    - \* mouth
    - \* thyroid cartilage
    - \* ankle(R)
  - On a scale of 0-10, the pain intensity caused by the rash is 2
  - The rash lesions are larger than 1cm
  - On a scale of 0-10, the severity of the rash itching is 10
  - I have nausea.
  - I have swelling in one or more areas of my body.
  - The swelling locations are:
    - \* forehead
    - \* cheek(R)
    - \* cheek(L)
    - \* nose
  - I have noticed a high pitched sound when breathing in.

Antecedents:

- 
- I have a known severe food allergy.
  - I am more likely to develop common allergies than the general population.

Differential diagnosis:

-----

Anaphylaxis: 0.159, Possible NSTEMI / STEMI: 0.112, Localized edema: 0.110, GERD: 0.110, Unstable angina: 0.104, Larygospasm: 0.095, Boerhaave: 0.093, Pulmonary embolism: 0.069, Chagas: 0.057, Pericarditis: 0.046, Stable angina: 0.045

## Sample 5

Sex: F, Age: 36  
Geographical region: North America  
Pathology: Pulmonary embolism  
Symptoms:

- 
- I feel pain.
  - The pain is:
    - \* a knife stroke
  - The pain locations are:
    - \* side of the chest(L)
    - \* breast(R)
    - \* breast(L)
    - \* posterior chest wall(R)
    - \* posterior chest wall(L)
  - On a scale of 0-10, the pain intensity is 1

- The pain radiates to these locations:
  - \* side of the chest(L)
  - \* scapula(R)
  - \* scapula(L)
  - \* posterior chest wall(R)
  - \* posterior chest wall(L)
- On a scale of 0-10, the pain's location precision is 3
- On a scale of 0-10, the pace at which the pain appeared is 9
- I am experiencing shortness of breath or difficulty breathing in a significant way.
- I have swelling in one or more areas of my body.
- The swelling locations are:
  - \* calf(L)
- I have pain which increases when I breathe in deeply.

Antecedents:  
-----

Differential diagnosis:  
-----

Pulmonary embolism: 0.091, Panic attack: 0.072, Anaphylaxis: 0.071, Spontaneous pneumothorax: 0.070, Possible NSTEMI / STEMI: 0.069, Unstable angina: 0.066, Stable angina: 0.054, Pulmonary neoplasm: 0.054, Pericarditis: 0.053, Guillain-Barre syndrome: 0.051, Myocarditis: 0.050, Atrial fibrillation: 0.049, Acute pulmonary edema: 0.044, SLE: 0.041, Acute dystonic reactions: 0.036, Myasthenia gravis: 0.036, Anemia: 0.036, Chagas: 0.030, Sarcoidosis: 0.029

## Sample 6

Sex: M, Age: 43  
 Geographical region: North America  
 Pathology: Panic attack  
 Symptoms:

-----

- I feel pain.
- The pain is:
  - \* a cramp
- The pain locations are:
  - \* side of the chest(R)
  - \* flank(R)
  - \* hypochondrium(R)
  - \* hypochondrium(L)
  - \* belly
- On a scale of 0-10, the pain intensity is 6
- On a scale of 0-10, the pain's location precision is 9
- On a scale of 0-10, the pace at which the pain appeared is 7
- I am experiencing shortness of breath or difficulty breathing in a significant way.
- I feel lightheaded and dizzy.
- I feel like I am dying.
- I have nausea.
- I feel palpitations.
- I feel like I am detached from my own body or my surroundings.
- I have recently had numbness, loss of sensation or tingling on my body.

Antecedents:

-----

- Some members of my immediate family have a psychiatric illness.
- I drink alcohol excessively.
- I have been diagnosed with depression.
- I suffer from chronic anxiety.
- Some family members are known to have migraines.
- I suffer from fibromyalgia.
- I have had a head trauma.

Differential diagnosis:

-----

Panic attack: 0.094, PSVT: 0.076, Possible NSTEMI / STEMI: 0.073, Spontaneous pneumothorax: 0.066, Anaphylaxis: 0.063, Unstable angina: 0.061, Anemia: 0.059, Guillain-Barre syndrome: 0.058, Boerhaave: 0.056, Atrial fibrillation: 0.054, Acute pulmonary edema: 0.053, Pulmonary embolism: 0.049, Scombroid food poisoning: 0.044, GERD: 0.041, Stable angina: 0.038, Acute dystonic reactions: 0.033, Myasthenia gravis: 0.033, Sarcoidosis: 0.027, Chagas: 0.025

### Sample 7

Sex: F, Age: 63  
Geographical region: North America  
Pathology: Bronchitis  
Symptoms:

-----  
- I feel pain.  
- The pain is:  
    \* burning  
- The pain locations are:  
    \* lower chest  
    \* side of the chest(L)  
    \* pharynx  
- On a scale of 0-10, the pain intensity is 3  
- On a scale of 0-10, the pain's location precision is 2  
- On a scale of 0-10, the pace at which the pain appeared is 3  
- I am experiencing shortness of breath or difficulty breathing in a significant way.  
- My cough produces colored or more abundant sputum than usual.  
- I have fever.  
- I have a sore throat.  
- I am coughing.  
- I have noticed a wheezing sound when I exhale.

Antecedents:

-----  
- I have a chronic obstructive pulmonary disease.

Differential diagnosis:

-----  
Acute COPD exacerbation / infection: 0.075, Pneumonia: 0.070, Bronchitis: 0.070, Bronchiectasis: 0.065, Panic attack: 0.053, Pulmonary neoplasm: 0.052, Tuberculosis: 0.052, Possible NSTEMI / STEMI: 0.051, GERD: 0.051, Unstable angina: 0.049, Pericarditis: 0.049, URTI: 0.045, Boerhaave: 0.045, Stable angina: 0.041, Acute laryngitis: 0.039, Atrial fibrillation: 0.036, Viral pharyngitis: 0.035, Guillain-Barre syndrome: 0.026, Acute dystonic reactions: 0.026, Myocarditis: 0.026, Sarcoidosis: 0.021, PSVT: 0.014, Influenza: 0.005, Chagas: 0.005

### Sample 8

Sex: F, Age: 35  
Geographical region: North America  
Pathology: Acute rhinosinusitis  
Symptoms:

-----  
- I feel pain.  
- The pain is:  
    \* burning  
    \* sharp  
- The pain locations are:  
    \* forehead  
    \* cheek(R)  
    \* nose  
    \* eye(L)  
- On a scale of 0-10, the pain intensity is 2  
- The pain radiates to these locations:

- \* forehead
- \* nose
- On a scale of 0-10, the pain's location precision is 8
- I have fever.
- I have lost my sense of smell.
- I have greenish/yellowish nasal discharge.
- I am coughing.

Antecedents:

-----

- I have had a cold in the last 2 weeks.
- I have had to use a bronchodilator in the past.
- I have been diagnosed with gastroesophageal reflux.
- my vaccinations are up to date.
- I am more likely to develop common allergies than the general population.

Differential diagnosis:

-----

Acute rhinosinusitis: 0.232, Chronic rhinosinusitis: 0.204, Bronchitis: 0.198,  
Tuberculosis: 0.146, Influenza: 0.145, Chagas: 0.064, Pneumonia: 0.011

## Sample 9

Sex: M, Age: 49  
Geographical region: North America  
Pathology: Allergic sinusitis  
Symptoms:

-----

- My nose or the back of my throat is itchy.
- I have severe itching in one or both eyes.
- I have nasal congestion.
- I am coughing.

Antecedents:

-----

- Some family members suffer from allergies, hay fever or eczema.
- Some family members have asthma.

Differential diagnosis:

-----

Allergic sinusitis: 0.348, Bronchitis: 0.251, URTI: 0.201, Influenza: 0.200

## Sample 10

Sex: M, Age: 32  
Geographical region: North America  
Pathology: Pulmonary embolism  
Symptoms:

-----

- I have been coughing up blood.
- I feel pain.
- The pain is:
  - \* a knife stroke
  - \* sharp
- The pain locations are:
  - \* upper chest
  - \* breast(R)
  - \* breast(L)
  - \* posterior chest wall(R)
  - \* posterior chest wall(L)
- On a scale of 0-10, the pain intensity is 10
- The pain radiates to these locations:
  - \* side of the chest(R)
  - \* side of the chest(L)
  - \* breast(L)
  - \* posterior chest wall(R)
  - \* posterior chest wall(L)

- On a scale of 0-10, the pain's location precision is 9
- On a scale of 0-10, the pace at which the pain appeared is 8
- I am experiencing shortness of breath or difficulty breathing in a significant way.
- I lost consciousness.
- I have pain which increases when I breathe in deeply.

Antecedents:

-----

- I have had surgery within the last month.

Differential diagnosis:

-----

Pulmonary embolism: 0.098, Pericarditis: 0.080, Spontaneous pneumothorax: 0.079, Possible NSTEMI / STEMI: 0.074, Panic attack: 0.072, Unstable angina: 0.066, Boerhaave: 0.061, Acute pulmonary edema: 0.058, Myocarditis: 0.057, Stable angina: 0.054, Guillain-Barre syndrome: 0.051, Atrial fibrillation: 0.049, GERD: 0.046, Acute dystonic reactions: 0.036, Myasthenia gravis: 0.036, Anemia: 0.036, Sarcoidosis: 0.029, PSVT: 0.019

## Sample 11

Sex: M, Age: 75

Geographical region: North America

Pathology: Pericarditis

Symptoms:

-----

- I feel pain.
- The pain is:
  - \* sharp
- The pain locations are:
  - \* upper chest
  - \* breast(R)
  - \* breast(L)
- On a scale of 0-10, the pain intensity is 7
- The pain radiates to these locations:
  - \* posterior chest wall(L)
- On a scale of 0-10, the pain's location precision is 7
- On a scale of 0-10, the pace at which the pain appeared is 7
- I feel palpitations.
- My symptoms worse when lying down and alleviated while sitting up.
- I have pain which increases when I breathe in deeply.

Antecedents:

-----

- I have had a pericarditis.

Differential diagnosis:

-----

Pericarditis: 0.180, Pulmonary embolism: 0.112, Panic attack: 0.111, Possible NSTEMI / STEMI: 0.111, Unstable angina: 0.104, Boerhaave: 0.103, PSVT: 0.098, Stable angina: 0.054, GERD: 0.050, Scombroid food poisoning: 0.042, Spontaneous pneumothorax: 0.035

## K Model Analysis

We present in this section the sequence of question-answer pairs as well as the differentials predicted by AARLC and BASD for the patient introduced in Section 3.5. Two variants are considered for each model, one trained to predict the ground truth differential and one trained to predict the ground truth pathology. At the beginning of the interaction with a model, the patient provides her age, sex, and an initial evidence. The behavior of all 4 models is evaluated by a doctor.

### AARLC trained to predict the differential diagnosis

Sex: F, Age: 79



Initial evidence:

-----

I have symptoms that increase with physical exertion but alleviate with rest

Agent inquiries:

-----

- Characterize your pain:
  - \* a knife stroke
- Do you have pain somewhere, related to your reason for consulting?
  - Y
- How fast did the pain appear?
  - \* 9
- How precisely is the pain located?
  - \* 4
- How intense is the pain?
  - \* 7
- Where is the affected region located?
  - nowhere
- Do you feel pain somewhere?
  - \* upper chest
  - \* breast(R)
  - \* breast(L)
- Does the pain radiate to another location?
  - \* nowhere
- Where is the swelling located?
  - nowhere
- Have you traveled out of the country in the last 4 weeks?
  - \* N
- Are you experiencing shortness of breath or difficulty breathing in a significant way?
  - N
- Do you have a cough?
  - N
- Do you have a chronic obstructive pulmonary disease (COPD)?
  - Y
- Do you have a fever (either felt or measured with a thermometer)?
  - N
- Do you feel lightheaded and dizzy or do you feel like you are about to faint?
  - N
- Do you smoke cigarettes?
  - Y
- Have you ever had a spontaneous pneumothorax?
  - Y

Predicted Differential:

-----

Unstable angina, Stable angina, Possible NSTEMI / STEMI, Spontaneous pneumothorax, GERD, Pericarditis, Pulmonary embolism, Atrial fibrillation

## AARLC trained to predict the ground truth pathology

Sex: F, Age: 79

Initial evidence:

-----

I have symptoms that increase with physical exertion but alleviate with rest

Agent inquiries:

-----

- Do you have pain somewhere, related to your reason for consulting?
  - Y
- Does the pain radiate to another location?
  - \* nowhere
- Where is the affected region located?
  - nowhere

- Do you feel pain somewhere?
  - \* upper chest
  - \* breast(R)
  - \* breast(L)
- How intense is the pain?
  - \* 7

Predicted Differential:

-----

Myocarditis, Spontaneous pneumothorax

### **BASD trained to predict the differential diagnosis**

Sex: F, Age: 79

Initial evidence:

-----

I have symptoms that increase with physical exertion but alleviate with rest

Agent inquiries:

-----

- Do you have pain somewhere, related to your reason for consulting?
  - Y
- Do you feel pain somewhere?
  - \* upper chest
  - \* breast(R)
  - \* breast(L)
- How intense is the pain?
  - \* 7
- Characterize your pain:
  - \* a knife stroke
- How precisely is the pain located?
  - \* 4
- How fast did the pain appear?
  - \* 9
- Do you smoke cigarettes?
  - Y
- Have you ever had a spontaneous pneumothorax?
  - Y
- Do you have a chronic obstructive pulmonary disease (COPD)?
  - Y

Predicted Differential:

-----

Unstable angina, Possible NSTEMI / STEMI, Stable angina, Pericarditis, Atrial fibrillation, Spontaneous pneumothorax, GERD

### **BASD trained to predict the ground truth pathology**

Sex: F, Age: 79

Initial evidence:

-----

I have symptoms that increase with physical exertion but alleviate with rest

Agent inquiries:

-----

- Characterize your pain:
  - \* a knife stroke
- How intense is the pain?
  - \* 7
- Do you have pain somewhere, related to your reason for consulting?
  - Y
- How precisely is the pain located?
  - \* 4

- Do you feel pain somewhere?
  - \* upper chest
  - \* breast(R)
  - \* breast(L)
- How fast did the pain appear?
  - \* 9
- Are you experiencing shortness of breath or difficulty breathing in a significant way?
  - N
- Do you have pain that is increased when you breathe in deeply?
  - N
- Are your symptoms worse when lying down and alleviated while sitting up?
  - N
- Do you smoke cigarettes?
  - Y
- Do you have a chronic obstructive pulmonary disease (COPD)?
  - Y
- Have you ever had a spontaneous pneumothorax?
  - Y
- Have any of your family members ever had a pneumothorax?
  - Y
- Do you have chest pain even at rest?
  - Y

Predicted Differential:

-----  
Spontaneous pneumothorax

Following is the evaluation performed by the doctor:

- AARLC trained on the ground truth pathology is the worst, having too few questions and an incomplete differential.
- BASD trained on the differential has a good differential, but too few collected evidences, making the collected evidence unspecific. More questions are needed to adequately cover the proposed differential.
- AARLC trained with the differential and BASD trained on the ground truth pathology produce the best results, gather a sensible amount of evidence. AARLC actively searches a wider array of diseases, searching also for infectious causes. BASD is still very good, unsurprisingly searching for a pathology confirmation and goes a bit further to ask specific questions. A clinician would like the best of both worlds, but AARLC covers a wider range of diseases, although a doctor would need to complete the evaluation with more questions. Pertinent questions will be influenced by the clinical context. In an acute care setting, a physician will try to gain specificity and include more questions centered around high acuity diseases in order to wisely choose the following medical tests to rule-in or rule-out the diseases.