Machine Learning for Automated Classification of Patient Cases

Cole Deisseroth - cdeisser, Jon Wang - jonwang1, James Bai - jamesbai

Abstract

Background: Much of high-quality practice and decision-making depends on one physician. Sending patients to the right department is important to save physicians time and help patients find treatment. This is an initial study toward the development of an intelligent patient-allocation system. This serves to save medical personnel valuable time, and help patients find the care they need more efficiently by automatically categorizing cases into specific departments. We develop an algorithm which predicts the categories of patient cases from the American Board of Internal Medicine Examinations—a certification that all physicians must go through to practice general medicine. *Methods:* Our ontology breaks questions into their components (Case, AnswerChoice, Explanation). We then run an automatic concept extractor (ClinPhen) on the passage (description of the case) to compile a list of concepts (words, phenotypes, and phenotype closures). We then use a Naïve Bayes classifier to take the concepts and predict the category of the case. *Results:* We have developed a classifier that predicts the category of a patient case correctly 80.5% of the time, and has over 80% precision and recall. *Discussion:* Future work will include developing more-sophisticated techniques of leveraging up-to-date knowledge graphs, and building our own graphs to categorize these cases. Ultimately, this classifier should become applicable in clinical settings (and not just for medical board cases), and be able to accurately suggest a department to send a patient to.

Background and Motivation

Despite the rapid and widely successful incorporation of artificial intelligence into a plethora of different industries, when it comes to medicine, much of high-quality practice and decision-making depends almost entirely upon a single physician. Consistency and variability are a part of much of the current practice, especially when determining which department to send a patient to from the emergency department [1]. Many clinical decisions in the hospital lack evidence-based support, due in part to the difficulty of performing randomized controlled patient experiments, as well as the high variability in compliance to evidence-based guidelines [2], [3]. Moreover, only about 11% of recommendation guidelines are backed by high-quality evidence [4]. This has detrimental effects on healthcare, with thousands of people dying annually due to medical errors made in hospitals [5]. Providing high-quality medical care consistently involves a combination of clinical experience and knowledge derived from literature [6], [7], and can be best achieved by a clinician in the appropriate department, with the most-relevant expertise. However, with the progressively growing number of possible medications, diagnoses, and procedures, selecting the correct department for these patients becomes less tractable, requiring larger amounts of time and resources [2], [8]. Hence, when allocating patients to a department, physicians tend to rely on personal intuition rather than data from robust scientific studies [9].

Studies have demonstrated the potential of automated algorithms to produce decision support better than manually derived support systems [10]–[18]. With the increasing amount of clinical data available, machine learning algorithms have been shown to be very effective in many supervised learning tasks, such as medical image segmentation and predicting diagnoses, readmission, length of stay, and death [19]–[22].

The current study proposes an initial foray into the development of an intelligent allocation system for medical patients. This serves to save medical personnel valuable time and help patients find the care they need more efficiently by automatically categorizing cases into specific departments. Our interest is to take a first pass at this through the development of an algorithm which automatically categorizes the cases in the American Board of Internal Medicine Examinations—a certification that all physicians must go through to practice general medicine. Though a number of companies claim to have had success in developing algorithms similar to this [23]–[25], to our knowledge, none of them have published on their findings. Without published findings, it is difficult to reproduce and further progress on the development of these algorithms [26]. Thus, this remains a potentially impactful problem to solve, and would facilitate further development of intelligent allocation systems for medical patients visiting the emergency department.

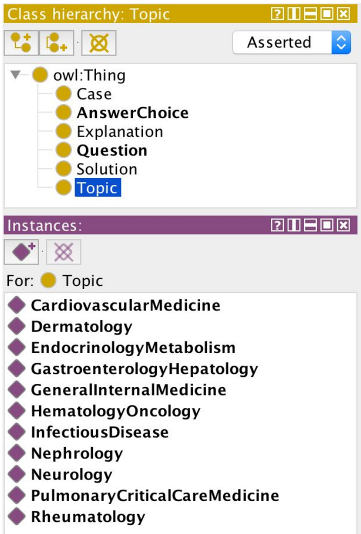
Methods

A. Data

Being aware of the page limit, please see Appendix A for details on data acquisition and data description.

B. Ontology

We used Protegé to model an Exam Ontology (ExOn) of Board questions (Fig 1). The ABIM tells us the topics, but the question-component breakdown shown in the “Class hierarchy” is of our own design.



owl:Thing
- Case
- AnswerChoice
- Explanation
- Question
- Solution
- Topic

For: Topic

- CardiovascularMedicine
- Dermatology
- EndocrinologyMetabolism
- GastroenterologyHepatology
- GeneralInternalMedicine
- HematologyOncology
- InfectiousDisease
- Nephrology
- Neurology
- PulmonaryCriticalCareMedicine
- Rheumatology

Figure 1: Exam Ontology (ExOn)

Every Question contains exactly one Case, Explanation, and 4 or 5 answer choices. Additionally, each question hasTopic some Topic (of which there are 11 instances). Each AnswerChoice contains exactly 1 Solution.

We also tried incorporating the Human Phenotype Ontology (HPO) [27] into the program that we use to classify cases. HPO is an OBO-Foundry ontology that provides a hierarchy of human disease phenotypes. For example, Generalized tonic-clonic seizures (HP:0002069) is a subclass of Generalized seizures (HP:0002197).

Finally, we used probabilistic graph models for the purpose of training a classifier. We used two versions of the probabilistic graph model. Here, circles are nodes, and arrows represent a conditional dependency between two nodes (A -> B means B is conditionally dependent on A).

1. Each case category served as the root node. The root node points to every word in the vocabulary, where the vocabulary consists of every word ever seen in all the cases (Figure 2). In the figure below, n represents number of unique words over all cases.

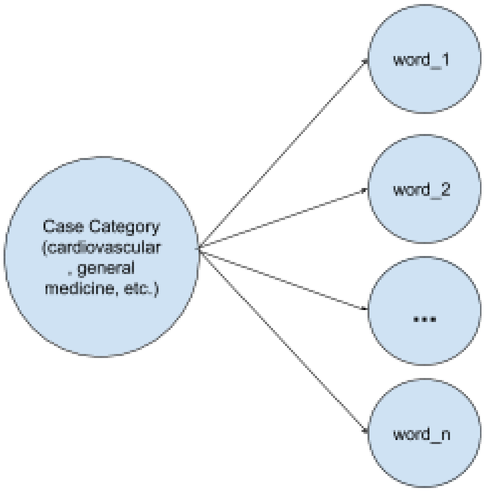


Figure 2: Probabilistic Graph Model with All Words

2. Each case category served as the root node. The root note points to every Human Phenotype Ontology term seen in all the questions (Fig 3). In the figure below, n represents the number of unique Human Phenotype Ontology terms used over all cases.

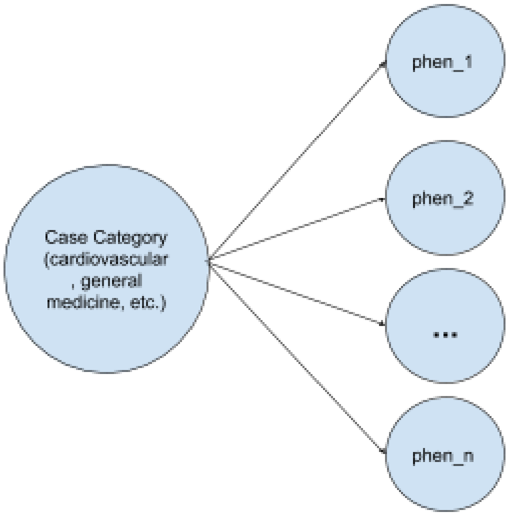


Figure 3: Probabilistic Graph Model with Human Phenotype Ontology terms

C. Problem-Solving Methods

As mentioned, every ABIM question has a passage about a patient case (in terms of our ontology, every Question contains a Case), which details the phenotypic abnormalities, such as seizures and coughing. To predict the category (or, as our ontology calls it, the Topic) of the case, we would first need to tokenize the patient's phenotypic information. We did this using ClinPhen, a recently published tool that automatically extracts HPO phenotypes from free text. We built a Naïve Bayes classifier that would take information from the patient cases and predict the category. Our classifier uses Bayes' theorem to predict the most-likely category of a case given the information provided. Let P(X) = the probability that the case is in Category "X", and P(a,b,c) = the probability that the case has phenotypes "A", "B", and "C":

$$P(X|a,b,c) = P(X)P(a|X)P(b|X)P(c|X) / P(a)P(b)P(c).$$

Using this formula, the classifier calculates the probability of each category given the phenotypes (or words) found in the patient case. One can then predict the category to be the one marked with the highest probability. The assumptions of the classifier are as follows: 1) each category is mutually exclusive/disjoint, 2) all observations are independent of one another, and 3) our categories are exhaustive or complete. Evidently our data does not fit assumption 2, as certain terms are highly correlated with each other. However, assumptions 1 and 3 are both met, as categories given from the scraping are complete and disjoint (though in real life when allocating patients, this is not true, so more sophisticated models may need to be used in the future).

We tested this classifier on different probabilistic graph models (as described above) to see what would work best. Note that the second probabilistic graph model was adapted for two forms of handling Human Phenotype Ontology terms, specific phenotypes and closures.

- All words in the cases (Fig 2): To see if ClinPhen was useful for this purpose, we also tried simply training/testing the classifier on the raw, unprocessed words found in the text.
- All Human Phenotype Ontology terms (Fig 3): We ran ClinPhen [28] on the passage and extracted a set of phenotypes for the question. We gave this to the classifier to train/test on.
- All Human Phenotype Ontology closures (Fig 3): We ran ClinPhen on the passage and got phenotypes. Then, we found the phenotype closure, which consists of the HPO phenotypes, plus all ancestors of each phenotype (For example, if a patient has Generalized tonic-clonic seizures, the closure will include that phenotype, plus its parent, Generalized seizures, and so on, all the way up to the root node, HP:0000001, All). We gave these phenotype closures to the classifier to train/test on.

D. Evaluation

Because we are designing a program that predicts the category of an ABIM case, the efficacy depends directly on how often the program gets the category right.

We have 3,421 patient cases. We split them into a Training set (3,081) and Testing set (340). We measured accuracy as the percentage of Test cases for which the classifier predicted the correct category.

This is an objectivistic summative evaluation, as we are comparing different methods for predicting categories. We evaluated our model using accuracy, precision, recall, and an ROC Curve.

Precision (P) is a measure of exactness or quality while recall (R) is a measure of completeness or quantity, they are defined below, :

$$P_i = \frac{TP_i}{TP_i + FP_i}, R_i = \frac{TP_i}{TP_i + FN_i}$$

Macro-averaged P and R, defined below, are an averaging per class:

$$P_{macro} = \frac{\sum_i^M P_i}{|M|}, R_{macro} = \frac{\sum_i^M R_i}{|M|}$$

F1, defined below, it is the harmonic mean of precision and recall:

$$F1 = 2\frac{P * R}{P + R}$$

Finally, Receiver Operating Characteristics (ROC) measures the performance of our model's classification abilities at various thresholds. Sensitivity is plotted against 1-Specificity for the macro averaged values. This allows one to see the tradeoff between the two in a graph. We include the area under ROC (AUROC) as a part of our evaluation metrics as well. We use macro averaging due to the relatively even distribution of questions.

Error analysis on the accuracy of our algorithm is based on a subjectivistic summative evaluation method. We identified the passages that were inaccurately categorized and scanned for any words or phrases that could have confounded the algorithm.

Results

We have parsed Board questions into a machine-readable format, and we have the distribution of question types recorded as well (see Appendix A, figures S1, S2).

Our all-words model outperforms on all test set evaluation metrics (Table 1). Our ROC curve shows a similar outperformance in the tradeoff between sensitivity and specificity (Fig 4).

	All Words	Phenotype Terms	Phenotype Closures
Accuracy	**80%**	56%	56%
Precision	**0.81**	0.57	0.63
Recall	**0.82**	0.58	0.58
F1	**0.80**	0.55	0.53
AUROC	**0.98**	0.89	0.88

Table 1: Performance Metrics

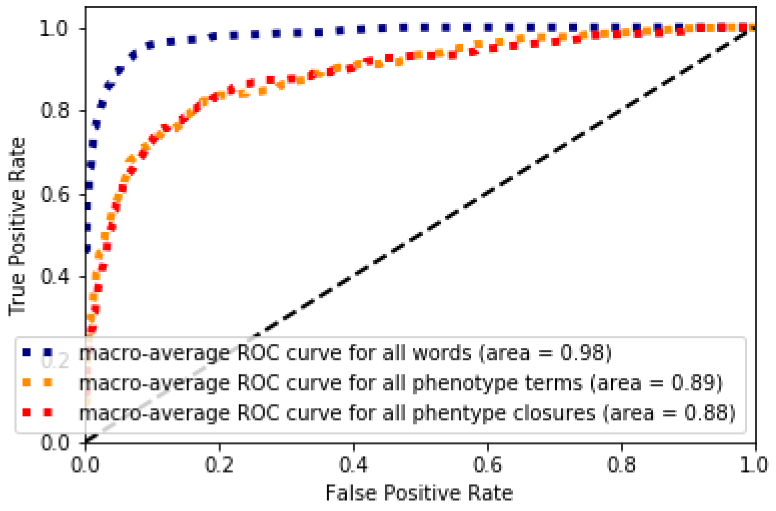


Figure 4: ROC-Curve of Model Results

Question type classification

When using ClinPhen, ClinPhen + Closures, and words to feed into the question-type classifier, we found that using words resulted in the best performance. Using ClinPhen phenotypes as features, the classifier correctly predicted the category 56.5% of the time. Using the closures of phenotypes extracted by ClinPhen resulted in 55.6% accuracy. Using just the words in the question resulted in 80.5% accuracy.

Qualitative Error Analysis

We also performed a qualitative error analysis by manual inspection. The errors seem to be attributed to 1) the high frequency of generic terms competing with the low frequency of specific terms and 2) the non-specific nature of most patient conditions.

To explain the first, as our algorithm doesn't allocate a heavier weight on specific terms that have a one-to-one relationship with the accurate category, it becomes difficult to categorize when many generic terms that have a slight preference for an inaccurate category are introduced. For example, a cardiovascular question containing heart failure as the classifying term gets inaccurately categorized as pulmonary critical care due to simultaneous appearance of more general terms like CPR and ventricular fibrillation that slightly tip the scale (Appendix B). Secondly, most patient conditions could seldom be classified under one category. To look at one example, a cardiovascular question with mentions of hypertension and respiration was "wrongly" categorized as an endocrinology/metabolism question due to simultaneous mentions of diabetes, metformin, and hemoglobin A1c (Appendix B). However, this categorization is not necessarily wrong, because all of these factors do point to more than one medical problem.

Discussion and Future Work

Our goal for this project was to take an initial step toward automating the decision of which department to send a patient to. We aimed to do this by building a tool that accurately predicts the category of patient cases given in the ABIM exam.

One advantage of our approach is that the algorithm uses real descriptions from the medical board exams that reflect accurate hypothetical patient descriptions. One drawback is that our model assumes all phenotypes are independent of one another, when in reality they are not. This becomes especially prevalent when using closures over specific phenotype terms. As shown (Table 1), the AUROC actually decreases for the closures model, likely due to the highly correlated nature of the terms (for example, if a patient has Generalized tonic-clonic seizures, then the patient always has Seizures).

As more and more patients come in with conditions that span multiple departments, it becomes harder for physicians to refer them to the most appropriate departments. Our algorithm not only identifies the best department for the relevant conditions, but also will rank the next probable departments for the physicians to pick from, to reduce the risk of erroneous assignment.

For classification, raw words were more effective than HPO phenotypes, likely because ClinPhen only extracts known phenotype terms from the text, while there are many other verbal cues (besides phenotypes) that more-strongly correlate with certain question categories.

To improve our current model, we could try and use both HPO terms and all the words in the patient cases for our classifier rather than just one of the two, and we could develop a neural network to replace the Naïve Bayes classifier. We could also enhance HPO terms by finding similar words in an embedding space trained on PubMed data (code for this submitted separately).

References

[1] J. B. McKinlay, C. L. Link, K. M. Freund, L. D. Marceau, A. B. O’Donnell, and K. L. Lutfey, “Sources of variation in physician adherence with clinical guidelines: results from a factorial experiment,” *J. Gen. Intern. Med.*, vol. 22, no. 3, pp. 289–296, 2007.

[2] S. Timmermans and A. Mauck, “The promises and pitfalls of evidence-based medicine,” *Health Aff.*, vol. 24, no. 1, pp. 18–28, 2005.

[3] F. Chollet *et al.*, “Can utilizing a computerized provider order entry (CPOE) system prevent hospital medical errors and adverse drug events?,” *AMIA Annu. Symp. Proc.*, vol. 23, no. 1, pp. 1–17, Jan. 2009.

[4] P. Tricoci, “Scientific Evidence Underlying the ACC/AHA Clinical Practice Guidelines,” *Jama*, vol. 301, no. 8, p. 831, 2009.

[5] M. S. Donaldson, J. M. Corrigan, L. T. Kohn, and others, *To err is human: building a safer health system*, vol. 6. National Academies Press, 2000.

[6] D. L. Sackett, W. M. C. Rosenberg, J. A. M. Gray, R. B. Haynes, and W. S. Richardson, “Evidence based medicine: what it is and what it isn’t.” British Medical Journal Publishing Group, 1996.

[7] G. H. Guyatt *et al.*, “Users’ guides to the medical literature: IX. A method for grading health care recommendations,” *Jama*, vol. 274, no. 22, pp. 1800–1804, 1995.

[8] D. T. Durack, “The weight of medical knowledge,” *N. Engl. J. Med.*, vol. 298, no. 14, pp. 773–775, 1978.

[9] R. Madhok, “Crossing the Quality Chasm: Lessons from Health Care Quality Improvement Efforts in England,” *Baylor Univ. Med. Cent. Proc.*, vol. 15, no. 1, pp. 77–83, 2002.

[10] J. H. Chen, M. K. Goldstein, S. M. Asch, L. Mackey, and R. B. Altman, “Predicting inpatient clinical order patterns with probabilistic topic models vs conventional order sets,” *J. Am. Med. Informatics Assoc.*, vol. 24, no. 3, pp. 472–480, 2017.

[11] J. H. Chen, T. Podchiyska, and R. B. Altman, “OrderRex: clinical order decision support and outcome predictions by data-mining electronic medical records,” *J. Am. Med. Informatics Assoc.*, vol. 23, no. 2, pp. 339–348, 2015.

[12] A. Wright and D. F. Sittig, “Automated development of order sets and corollary orders by data mining in an ambulatory computerized physician order entry system,” in *AMIA annual symposium proceedings*, 2006, vol. 2006, p. 819.

[13] Y. Zhang, R. Padman, and J. E. Levin, “Paving the COWpath: data-driven design of pediatric order sets,” *J. Am. Med. Informatics Assoc.*, vol. 21, no. e2, pp. e304--e311, 2014.

[14] A. P. Wright, A. T. Wright, A. B. McCoy, and D. F. Sittig, “The use of sequential pattern mining to predict next prescribed medications,” *J. Biomed. Inform.*, vol. 53, pp. 73–80, 2015.

[15] Y. Zhang, J. E. Levin, and R. Padman, “Data-driven order set generation and evaluation in the pediatric environment,” in *AMIA Annual Symposium Proceedings*, 2012, vol. 2012, p. 1469.

[16] J. Klann, G. Schadow, and S. M. Downs, “A method to compute treatment suggestions from local order entry data,” in *AMIA Annual Symposium Proceedings*, 2010, vol. 2010, p. 387.

[17] J. G. Klann, P. Szolovits, S. M. Downs, and G. Schadow, “Decision support from local data: creating adaptive order menus from past clinician behavior,” *J. Biomed. Inform.*, vol. 48, pp. 84–93, 2014.

[18] J. Klann, G. Schadow, and J. M. McCoy, “A recommendation algorithm for automating corollary order generation,” in *AMIA Annual Symposium Proceedings*, 2009, vol. 2009, p. 333.

[19] K. Petersen, M. Nielsen, P. Diao, N. Karssemeijer, and M. Lillholm, “Breast Tissue Segmentation and Mammographic Risk Scoring Using Deep Learning,” *Breast Imaging Lect. Notes Comput. Sci.*, pp. 88–94, 2014.

[20] A. Avati, K. Jung, S. Harman, L. Downing, A. Ng, and N. H. Shah, “Improving palliative care with deep learning,” *2017 IEEE Int. Conf. Bioinforma. Biomed.*, 2017.

[21] G. Carneiro, J. Nascimento, and A. P. Bradley, “Unregistered multiview mammogram analysis with pre-trained deep learning models,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2015, pp. 652–660.

[22] C.-C. Chiu *et al.*, “Speech recognition for medical conversations,” Nov. 2017.

[23] Dom Galeon, “This robot has passed a medical licensing exam with flying colours | World Economic Forum.” [Online]. Available: https://www.weforum.org/agenda/2017/11/this-robot-has-passed-a-medical-licencing-exam-with-flying-colours. [Accessed: 04-Feb-2019].

[24] “This AI Just Beat Human Doctors On A Clinical Exam.” [Online]. Available: https://www.forbes.com/sites/parmyolson/2018/06/28/ai-doctors-exam-babylon-health/#331f13a312c0. [Accessed: 04-Feb-2019].

[25] “CloudMedx Clinical AI outperforms human doctors on a US medical exam.” [Online]. Available: https://www.prnewswire.com/news-releases/cloudmedx-clinical-ai-outperforms-human-doctors-on-a-us-medical-exam-300775145.html. [Accessed: 04-Feb-2019].

[26] Sam Finnikin, “Babylon’s ‘chatbot’ claims were no more than clever PR | Article | Pulse Today.” [Online]. Available: http://www.pulsetoday.co.uk/news/gp-topics/it/babylons-chatbot-claims-were-no-more-than-clever-pr/20037041.article. [Accessed: 04-Feb-2019].

[27] Robinson, P. N., Köhler, S., Bauer, S., Seelow, D., Horn, D., & Mundlos, S. (2008). The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. American journal of human genetics, 83(5), 610-5.

[28] Deisseroth CA, Birgmeier J, Bodle EE, Kohler JN, Matalon DR, Nazarenko Y, Genetti CA, Brownstein CA, Schmitz-Abe K, Schoch K, Cope H, Signer R; Undiagnosed Diseases Network, Martinez-Agosto JA, Shashi V, Beggs AH, Wheeler MT, Bernstein JA, and Bejerano G (2018). ClinPhen extracts and prioritizes patient phenotypes directly from medical records to expedite genetic disease diagnosis. Genetics in Medicine, 2018. DOI: 10.1038/s41436-018-0381-1

Division of Labor (< 1 page): Please give a short statement describing how each team member contributed to the final project.

Jon thought of the project and got data for it, and performed quantitative analyses.

Cole wrote and ran the Naïve Bayes classifier, and used ClinPhen to parse the patient cases for phenotypes.

James did literature review and validation/labeling of data, and subjectivistic error analysis.

Appendix A: Data

Our data is pulled from 3,600 American Board of Internal Medicine Certification Exam questions. Each question is comprised of a question, accompanying context passage, and 4 or 5 answer choice selections. Once an answer choice is selected, the correct answer, explanation passage, key point, and learning objective are revealed.

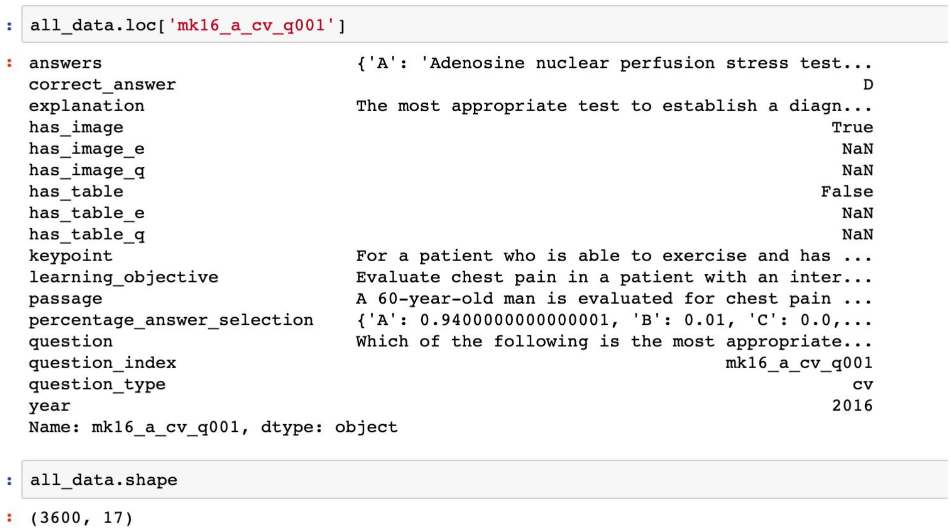
The exam website dynamically loads questions using Javascript, thus downloading and parsing HTML files directly did not provide the information we desired. We use Charles Web Debugging Proxy to identify the location the of the API that the Javascript calls to request the information.

We then use a Python script adapted from StackOverflow to scrape these examples from the API to obtain raw text data. Ultimately, 3564 examples were scraped from 2012, 2015, and 2018 exams.

We used Regex and BeautifulSoup to parse the following fields for each question:

1. Question ID (str): UUID of a question
2. Question (str): Question corresponding to passage
3. Passage (str): Context and background information required to answer question
4. Answer Choices (dict): key is answer choice (char), value is answer choice descriptor (str)
5. Learning Objective (str): Learning objective of the question
6. Key Point (str): Key idea needed to answer the question properly
7. Distribution of Answer Selections (list[float]): the percent distribution of answer selections made my human test takers
8. Question Type (str): Category of question (cardiovascular, neurology, etc)
9. Year (str): Year question was pub
10. Image in Explanation (bool)
11. Table in Passage (bool)
12. Image in Passage (bool)

Below is an example of a parsed question:



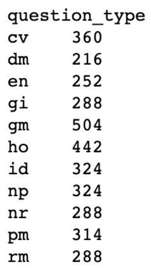
```
: all_data.loc['mk16_a_cv_q001']

: answers                          {'A': 'Adenosine nuclear perfusion stress test...
  correct_answer                                                                  D
  explanation                      The most appropriate test to establish a diagn...
  has_image                                                                    True
  has_image_e                                                                   NaN
  has_image_q                                                                   NaN
  has_table                                                                   False
  has_table_e                                                                   NaN
  has_table_q                                                                   NaN
  keypoint                         For a patient who is able to exercise and has ...
  learning_objective               Evaluate chest pain in a patient with an inter...
  passage                          A 60-year-old man is evaluated for chest pain ...
  percentage_answer_selection      {'A': 0.9400000000000001, 'B': 0.01, 'C': 0.0,...
  question                         Which of the following is the most appropriate...
  question_index                                                     mk16_a_cv_q001
  question_type                                                                  cv
  year                                                                         2016
  Name: mk16_a_cv_q001, dtype: object

: all_data.shape

: (3600, 17)
```

Figure S1: Machine-readable format of board questions snapshot



```
question_type
cv    360
dm    216
en    252
gi    288
gm    504
ho    442
id    324
np    324
nr    288
pm    314
rm    288
```

Figure S2: Number of questions per question type

Appendix B: Example Medical Board Question Description

Correct category: Cardiovascular

Predicted Category: Pulmonary Critical Care

A 74-year-old man hospitalized for a heart failure exacerbation goes into ventricular fibrillation and is administered cardiopulmonary resuscitation (CPR). An external defibrillator is attached and he receives a 200 J shock.

Correct category: Cardiovascular

Predicted category: Endocrinology

A 72-year-old woman is evaluated during a routine examination. Medical history is significant for hypertension, type 2 diabetes mellitus, and dyslipidemia. Medications are lisinopril, metformin, and pravastatin. She exercises daily; ingests a diet high in fruits, nuts, and vegetables; and does not smoke. She has no allergies. Blood pressure is 118/70 mm Hg, pulse rate is 73/min, and respiration rate is 16/min. The remainder of the physical examination, including cardiovascular, pulmonary, and neurologic examinations, is normal. Laboratory studies: Hemoglobin A1c

Appendix C: Code submission:

On Canvas, we have submitted our code, including the following scripts:

pheno_dag.py, get_pheno_closures.py: these scripts are used to relate phenotypes to their parent nodes, and get phenotype closures.

naive_bayes_prob_table.py: runs the Naïve Bayes classifier and outputs the probability of each category, for each question.

json_to_word_tables.py: converts a json with ABIM question data to a table mapping questions to the words that appear in them.

json_to_tables.py: converts a json with ABIM question data to a table mapping questions to their concepts, and to their categories.

prob_table.txt, prob_table_clinphen.txt, prob_table_clinphen_closure.txt: The probability that each question is in each category, according to our classifier (using words, HPO phenotypes, and HPO phenotype closures, respectively)

web_scrape: Gathers ABIM questions from the internet, in the form of a json

analysis.ipynb: Makes ROC curves, analyzes accuracy, etc. from the output of our classifier.

embeddings.zip: Mappings between words and similar words. Can be useful for more-advanced analysis.