The USMLE® Step 2 Clinical Skills Patient Note Corpus

Victoria Yaneva[1] Janet Mee[1] Le An Ha[2]

Polina Harik[1] Michael Jodoin[1] Alex J. Mechaber[1]

[1]National Board of Medical Examiners, Philadelphia, USA
{vyaneva, jmee, pharik, mjodoin, amechaber}@nbme.org
[2]University of Wolverhampton, UK
ha.l.a@wlv.ac.uk

Abstract

This paper presents a corpus of 43,985 clinical patient notes (PNs) written by 35,156 examinees during the high-stakes USMLE® Step 2 Clinical Skills examination. In this exam, examinees interact with *standardized patients* - people trained to portray simulated scenarios called *clinical cases*. For each encounter, an examinee writes a PN, which is then scored by physician raters using a rubric of clinical concepts, expressions of which should be present in the PN. The corpus features PNs from 10 clinical cases, as well as the clinical concepts from the case rubrics. A subset of 2,840 PNs were annotated by 10 physician experts such that all 143 concepts from the case rubrics (e.g., *shortness of breath*) were mapped to 34,660 PN phrases (*e.g., dyspnea, difficulty breathing*). The corpus is available via a data sharing agreement with NBME and can be requested at https://www.nbme.org/services/data-sharing.

1 Introduction

Large clinical text corpora are both one of the most needed and one of the least available resources in biomedical NLP, largely due to patient confidentiality considerations and expert annotation cost. This has been identified as a main reason for lagging progress in biomedical NLP compared to the general NLP domain (Chapman et al., 2011), and is evidenced by the fact that MIMIC-III (Johnson et al., 2016) is the only freely available large corpus of clinical notes to date (Section 2). As a result, biomedical NLP is heavily reliant on corpora of PubMed scientific abstracts,[1] whose academic language is in stark contrast to the often ungrammatical and telegraphic text constructions found in clinical notes.

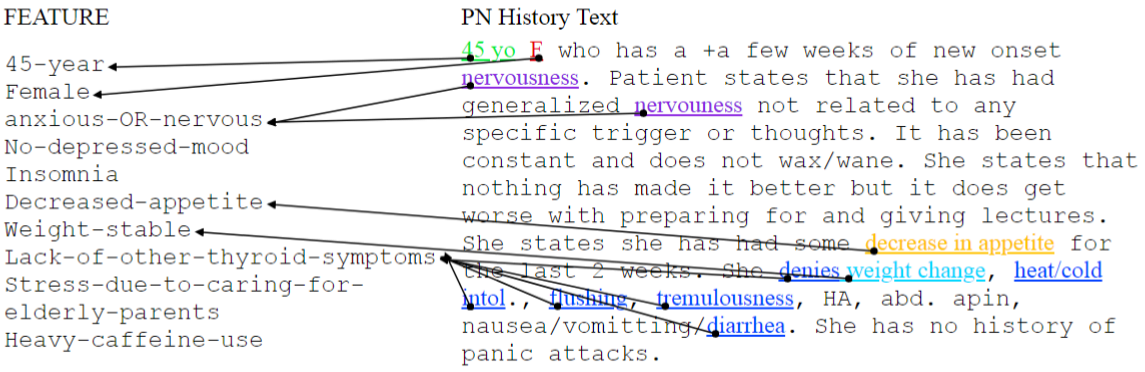
A known example of how the lack of shared clinical note corpora affects application development is the task of NLP-assisted scoring of clinical patient notes (PNs) written during exams. In medical education, students are often assessed through encounters with *standardized patients* - people trained to portray simulated scenarios called *clinical cases*. For each such encounter, the student is expected to perform a history and physical examination, determine differential diagnoses, and then document their findings in a PN. This assessment format is ubiquitous in medical education due to the important clinical skills it measures (van der Vleuten and Swanson, 1990; Wang et al., 2021), however, there is a significant cost associated with the manual scoring of the produced PNs by expert physician raters, as well as potential for human error and bias (Engelhard Jr et al., 2018).

There has been fragmented effort by individual institutions to train in-house NLP systems for clinical text scoring, with no fully transparent evaluation on public data (Luck et al., 2006; Spickard III et al., 2014; Latifi et al., 2016; Sarker et al., 2019). This has raised questions from a key stakeholder – the medical student community – about potential algorithmic bias and its implications for fairness (Spadafore and Monrad, 2019). Overall, the lack of shared data (here, mainly for exam security reasons) has slowed down innovation and limited public support, despite NLP's potential to alleviate financial burden and improve reliability.

The goal of this paper is to advance PN automated scoring specifically, and biomedical NLP in general, through the development and public release of a large corpus of examinee-written PNs. The corpus consists of 43,985 PN history portions from 10 clinical cases, where 2,840 PNs (35k phrases) were annotated with concepts from the exam scoring rubrics (Section 3). The main, but not sole, application for this data is the development of interpretable, transparent, and cost-effective systems for clinical text scoring, thus improving educational assessment in the field of medicine.

[1] See BLURB: https://microsoft.github.io/BLURB/

Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 2880 - 2886
July 10-15, 2022 ©2022 Association for Computational Linguistics



45 yo F who has a +a few weeks of new onset nervousness. Patient states that she has had generalized nervouness not related to any specific trigger or thoughts. It has been constant and does not wax/wane. She states that nothing has made it better but it does get worse with preparing for and giving lectures. She states she has had some decrease in appetite for the last 2 weeks. She denies weight change, heat/cold intol., flushing, tremulousness, HA, abd. apin, nausea/vomitting/diarrhea. She has no history of panic attacks.

Figure 1: Features from an exam rubric and their expressions within an example patient note excerpt

The two key contributions of this paper are the development of a large corpus of examinee-written PNs, made available for research purposes, and the expert annotation of a subset of 2,840 PNs to advance automated scoring of clinical PNs.

2 Related Datasets

Large corpora of clinical patient notes (e.g. > 2k) are scarcely available as shared resources. As noted in two overview articles by Savkov et al. (2016) and Campillos-Llanos et al. (2021), such large corpora include CLEF (565k notes), which is “currently restricted" , awaiting “a governance framework in which it can be made more widely available" (Roberts et al., 2007); and a corpus related to the TREC shared task, where “the University of Pittsburgh distributes the records only to track participants" (Voorhees et al., 2012). Among the larger EHR databases, the eICU database specifically excludes clinical note text: "to minimize risk of including PHI" (protected health information) [2]. These restrictions make MIMIC-III (Johnson et al., 2016) the only freely available large corpus of clinical notes to date.

As a result of patient confidentiality considerations, the use of patient notes describing fictional patients is not new in the field of biomedical NLP. This type of data has shown promise in several shared tasks: the NTCIR10[3] NTCIR11[4], NTCIR12[5], and NTCIR16[6] MedNLP tasks use descriptions of fictional patients written in Japanese. As reported in the NTCIR10 task overview paper, “we asked physicians to write down fictional medical reports of imaginary patients (...) We offered 50 collected medical reports for this task, which include 3,365 sentences in all: about 40,000 words" (Morita et al., 2013). In addition to its small size, limitations of this dataset include the lack of clarity around the procedure the physicians followed to create these patient notes. Nevertheless, given the lack of publicly available data from real patients, this dataset contributed to the field by enabling the evaluation of tasks such as patient anonymization and detection of complaint and diagnosis.

The next section describes the high-stakes clinical examination context in which the patient notes from our corpus were written.

3 Context

The United States Medical Licensing Examination® (USMLE®) is a series of examinations to support medical licensure decisions in the United States that is developed by the National Board of Medical Examiners (NBME®) and Federation of State Medical Boards (FSMB). Until 2020, one of the exams was the USMLE Step 2 Clinical Skills examination, which used standardized patients to assess examinee ability to gather information, perform physical examinations, and interpret data, as documented in the PNs examinees completed after each encounter (an example of a full PN is presented in Appendix A). Annually, the exam resulted in more than 330,000 PNs graded by more than 100 raters.

The PNs are scored by licensed physicians using case-specific rubrics that were developed by physi-

[2] https://eicu-crd.mit.edu/eicutables/note/

[3] https://sociocom.jp/mednlp/medistj-en/

[4] https://sociocom.jp/mednlp/ntcir11/#dataset

[5] https://sites.google.com/site/mednlpdoc/

[6] https://sociocom.naist.jp/real-mednlp/

cians on a test development committee. The rubrics outline each case's important concepts (henceforth called *features*) which should appear in an appropriately documented PN (Figure 1, Feature column). For example, for a clinical case about a patient with constant headaches, it may be important that the examinee asks questions leading to the information that the patient has *photophobia*. In a case like this, *photophobia* would be listed as one of the rubric features, and PNs that do not mention that specific symptom (or some expression of it such as *sensitive to light*) will receive a lower score.

A main challenge for developing an interpretable system that can identify expressions of the features in the PNs is the variety of ways in which features are expressed, with examples such as *loss of interest in activities* expressed as *no longer plays tennis*, or *shortness of breath* expressed as *dyspnea*. There is often a need to map concepts by combining multiple text segments, or resolve ambiguous negation as in *no cold intolerance, hair loss, palpitations or tremor* corresponding to the feature *lack of other thyroid symptoms*. In addition, automated scoring systems should employ a dynamic threshold to determine whether a given feature has been found in a PN, i.e., whether the F1 score for a given identified phrase is high enough for the phrase to be considered a match (Sarker et al., 2019). Finally, to be comparable to human rater performance and thus operationally usable, such systems need to be highly accurate. This requirement is crucial because of the high societal cost of passing an examinee with insufficient knowledge, and the high personal cost of failing an examinee who should have passed. As will be seen in Section 5, the average inter-rater agreement on whether a feature is mentioned in the corpus is F1 = 0.97.

4 Data

The dataset consists of the history portions[7] of 43,985 PNs from ten clinical cases (average # per case = 4,398; min = 992, max = 9,936) and the corresponding features for each case. The cases cover diverse clinical areas: Women's Health (2), Gastrointestinal (2), Neurological (1), Psychiatric (2), and Cardiovascular (3); as well as patients from diverse age groups: < 18 (2), 18-44 (6), 45-64 (1), 65+ (1). The number of tokens in the dataset is 5,958,464, with a type-token ratio of 0.022. The average length of each history portion is 135.47 tokens (SD = 24.27), and average number of history portion features per case is 14.3 (3.34).

Data were collected between 2017 and 2020 from 35,156 US or international medical students and graduates who took the exam under standardized conditions in one of five testing locations in the US. Each examinee-patient encounter resulted in a unique PN.

The dataset includes PNs only from examinees who, during registration, indicated that they agreed to have their data used in research. All PNs were assigned new IDs that cannot be linked to operational IDs used in scoring. The PNs do not include identifying information such as name, affiliation, or descriptions of personal experiences. Finally, the dataset features only the history portions of the PNs as opposed to complete PNs, and no information is given on which PNs belong to an individual examinee. This limits the inferences that can be made about the performance of individual examinees, while allowing the use of this data for advancing automated scoring and biomedical NLP research.

5 Annotation

A total of 2,840 PNs (284 per case) were annotated by 10 experienced US medical practitioners – nine with a Medical Doctorate degree and one with a degree in Nursing. The annotators were divided in five pairs of two, such that each pair would contain one experienced "senior" annotator. The annotation was performed using BRAT.[8] The annotators were instructed to first read the entire PN and then 1) identify all phrases that are expressions of a feature and link them to their corresponding feature (Figure 1), 2) mark fragmented annotations by excluding the text that is not relevant to the feature, and 3) mark each feature as a separate annotation (see detailed annotation guidelines in Appendix B). For example, if the feature was "No blood in stool", only the underlined text of the following excerpt was annotated: "No blood or mucus in stool". Unlike other features, *gender* and *age* were annotated only once for the first mention, with subsequent relevant phrases such as "she" or "his" not marked.

For each case, 284 notes were randomly selected for annotation and each annotator pair annotated notes from two cases over a period of six weeks. Two of the notes were annotated jointly as part of an initial discussion on the specifics of each clinical

[7] The history portion is where all relevant clinical information obtained from an interview with the patient is described.

[8] https://brat.nlplab.org/

case. During this discussion, the annotators would develop consistent case-specific understanding of the requirements for phrases to be considered a match (e.g., for the feature *visual hallucinations*, is the mention of *hallucinations* sufficient or does it need to be specified as *visual*?). Next, both annotators would annotate the same set of 5 notes independently and have a follow-up meeting to discuss potential discrepancies. After these were resolved, each annotator would proceed to independent work, where 29 notes per case (18% of the data) were double-rated[9] and used to compute inter-annotator agreement and the remaining 124 notes per annotator per case were single-rated. The annotators would receive a new set of notes weekly, to ensure an even work pace and mitigate fatigue.

The produced data were cleaned by fixing instances of wrong feature attribution (81) and correcting: leading and trailing spaces (167), punctuation (533), extra characters (115), and missing characters (e.g., as in "ot flashes") (64).

F1 agreement scores were computed based on character position overlap, with a substantial agreement across all cases of F1 = .84 (SD = 0.075); Jaccard distance of 86.55% (9.89); and Cohen's κ of 0.89 (0.057) (See individual case agreement scores in Appendix C). Finally, the annotators had an even higher agreement (binary F1) on whether an expression of a given feature was found in a PN or not (mean F1 = 0.97 (0.014)).

The final corpus includes 43,985 PNs, of which 2,840 (284 per case) were annotated and contains 34,660 annotated phrases linked to 143 features.

6 Baselines

To quantify the number of phrases from the gold standard that can be matched using simple heuristics and a small amount of annotated data, we compute three baselines. First, we divide the annotated portion of the data into ten folds. Then, we apply 10-fold cross-validation such that we take the phrases from one fold and see how many of them can be found in the remaining nine folds[10] using three approaches: i) direct match between a string from the "training" fold and those from the nine "test" folds, ii) fuzzy match with a window of two characters, iii) fuzzy match with a window of two characters and synonyms from WordNet (Miller, 1995) and the Unified Medical Language System (UMLS) (Bodenreider, 2004).

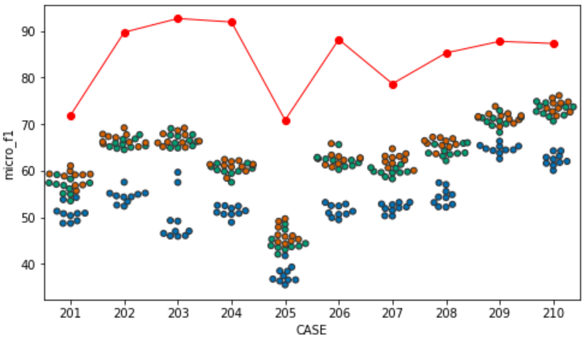


Figure 2: Comparison between inter-annotator agreement (red line) and three baselines: *exact match* (blue), *fuzzy match* (green), and *fuzzy + synonyms* (orange).

The evaluation metric is micro-averaged F1 of character span overlap between the predicted and gold-standard phrases, where a character span is a pair of indexes representing a range of characters within a text. For each instance, there is a collection of ground-truth spans (the phrases identified by the annotators) and a collection of predicted spans (the phrases identified by an automated system, in this case one of the three baselines). Each character within that span is identified as a true positive (TP) if it is within both a ground-truth and a prediction, a false negative (FN) if it is within a ground-truth but not a prediction, and a false positive (FP) if it is within a prediction but not a ground truth. The overall F1 score is computed from the TPs, FNs, and FPs aggregated across all instances.

As shown in Figure 2, the *fuzzy + synonyms* approach outperformed *exact* and *fuzzy match* with a mean F1 of .64 (.074), compared to .53 (.073), and .62 (.075). This result compares to an average inter-annotator agreement of .84 (.075) for character location overlap between phrases, showing a need for considerable improvement to match human performance. This gap varies between cases, with some having more than 20 points difference in F1 (e.g., Case 204). It is also seen that the variance in responses for certain cases (e.g., 201) is easier to capture computationally compared to others (e.g., 203). Finally, the results show that including a list of synonyms in *fuzzy + synonyms* does not lead to significant improvement, with the task requiring

[9] For the double-annotated notes, the annotations of the senior annotator are the ones included in the final dataset. As a rule, the annotations of the second annotator for the double-rated notes were only included in the final data when, for a given feature, the senior annotator did not find any matches but second annotator did. Such cases were very rare.

[10] This division is similar to those found in semi-supervised systems, which learn from the unannotated data and a small sample of annotated data.

more sophisticated semantic processing.

A binary F1 score of whether a given feature was expressed in a PN (1 if found, 0 otherwise) reveals a very strong agreement between the annotators (.97 (.014)) and a significantly worse performance for the best baseline (.86 (.048) for *fuzzy + synonyms match*). Therefore, to be comparable to human performance and thus operationally usable, automated approaches need to show a significant improvement over the baseline results presented here.

7 Discussion

The goal of this paper was to advance PN scoring and biomedical NLP through the development and annotation of a large PN corpus.

For PN scoring, this data can aid the development and evaluation of interpretable systems that identify feature expressions rather than black-box modeling of rater scores. Having a shared dataset can guarantee transparency, informing stakeholders on various aspects of system performance. It is also conceivable that the semantic mapping solutions enabled by this data could scale to scoring other constructed-response items, such as short-answer questions assessing clinical knowledge.

As noted in the Introduction, real clinical notes are scarcely available, which creates a bottleneck in the development of biomedical NLP. This corpus can help bridge this gap, since the PNs in it share many characteristics with real clinical notes – medical jargon, typos, abbreviations, and telegraphic style, among others. Moreover, having thousands of PNs written by different examinees that correspond to the same clinical case allows the development of robust NLP models exposed to a large-scale, real-life variation of clinical language. Such models would be trained to recognise the various ways in which, say, thyroid symptoms are described in clinical PNs, rather than their expressions in scientific abstracts. Beyond that, the corpus is relevant to machine reading comprehension and automated question answering, where the features are treated as yes/no questions ("Is photophobia present in this document"), and the identified phrases are supporting information.

The strengths of this data for some applications represent limitations for others. For example, all PNs in the corpus pertain to a set of ten cases, which excludes the possibility of using this data for patient cohort identification or phenotyping, typically performed with Electronic Health Record (EHR) data. In addition, the exam is a *simulation* of patient visits. Nevertheless, because of its high-stakes nature, the cases were treated as real.

Unlike EHR data, this corpus poses no risks for real patients, which is why the final data is less “sanitized" compared to deidentified EHR records; In addition, the cases were created by a team of licensed physicians ensuring that they are accurate representations of cases found in clinical practice. Including anonymized, partial data (history portions only) prevents risks for examinee identification or inferences about individual performance. Responsible use of the data for research purposes is further ensured by its distribution via a data use agreement. This is done following application to NBME’s Data Sharing and Collaboration Program at *https://www.nbme.org/services/data-sharing*.

It is our hope that the public release of this data will spur the development of interpretable and transparent solutions for PN scoring and related tasks, improving technology-assisted educational assessment in the field of medicine.

References

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.

Leonardo Campillos-Llanos, Ana Valverde-Mateos, Adrián Capllonch-Carrión, and Antonio Moreno-Sandoval. 2021. A clinical trials corpus annotated with umls entities to enhance the access to evidence-based medicine. *BMC medical informatics and decision making*, 21(1):1–19.

Wendy W Chapman, Prakash M Nadkarni, Lynette Hirschman, Leonard W D’avolio, Guergana K Savova, and Ozlem Uzuner. 2011. Overcoming barriers to nlp for clinical text: the role of shared tasks and the need for additional creative solutions.

George Engelhard Jr, Jue Wang, and Stefanie A Wind. 2018. A tale of two models: Psychometric and cognitive perspectives on rater-mediated assessments using accuracy ratings. *Psychological Test and Assessment Modeling*, 60(1):33–52.

Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-Wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Syed Latifi, Mark J Gierl, André-Philippe Boulais, and André F De Champlain. 2016. Using automated scoring to evaluate written responses in english and french on a high-stakes clinical competency examination. *Evaluation & the health professions*, 39(1):100–113.

J Luck, JW Peabody, and BL Lewis. 2006. An automated scoring algorithm for computerized clinical vignettes: evaluating physician performance against explicit quality criteria. *International journal of medical informatics*, 75(10-11):701–707.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Mizuki Morita, Yoshinobu Kano, Tomoko Ohkuma, Mai Miyabe, and Eiji Aramaki. 2013. Overview of the ntcir-10 mednlp task. In *NTCIR*. Citeseer.

Angus Roberts, Robert Gaizauskas, Mark Hepple, Neil Davis, George Demetriou, Yikun Guo, Jay Subbarao Kola, Ian Roberts, Andrea Setzer, Archana Tapuria, et al. 2007. The clef corpus: semantic annotation of clinical text. In *AMIA Annual Symposium Proceedings*, volume 2007, page 625. American Medical Informatics Association.

Abeed Sarker, Ari Z Klein, Janet Mee, Polina Harik, and Graciela Gonzalez-Hernandez. 2019. An interpretable natural language processing system for written medical examination assessment. *Journal of biomedical informatics*, 98:103268.

Aleksandar Savkov, John Carroll, Rob Koeling, and Jackie Cassell. 2016. Annotating patient clinical records with syntactic chunks and named entities: the harvey corpus. *Language resources and evaluation*, 50(3):523–548.

Maxwell Spadafore and Seetha U Monrad. 2019. Algorithmic bias and computer-assisted scoring of patient notes in the usmle step 2 clinical skills exam. *Academic Medicine*, 94(7):926.

Anderson Spickard III, Heather Ridinger, Jesse Wrenn, Nathan O'brien, Adam Shpigel, Michael Wolf, Glenn Stein, and Joshua Denny. 2014. Automatic scoring of medical students' clinical notes to monitor learning in the workplace. *Medical teacher*, 36(1):68–72.

Cees PM van der Vleuten and David B Swanson. 1990. Assessment of clinical skills with standardized patients: state of the art. *Teaching and Learning in Medicine: An International Journal*, 2(2):58–76.

Ellen M Voorhees, William R Hersh, et al. 2012. Overview of the trec 2012 medical records track. In *TREC*.

Jianjian Wang, Shuaixiang Zhao, Dong Roman Xu, Janne Estill, Meng Lv, M Zhang, Y Cai, J Liao, Y Lu, R Wang, et al. 2021. Developing evidenced-based quality assessment checklist for real practice in primary health care using standardized patients: a systematic review. *Annals of palliative medicine*.

A Patient Note Example

See Table 1.

B Annotation Guidelines

- Identify all phrases that are expressions of a feature from the History portion of the PNs and link them to their corresponding feature.

- Include fragmented annotations by excluding the text that is not relevant to the feature (e.g., if the feature is *No relief with Imodium or Cipro*, only the underlined text of the following excerpt should be annotated: *Has tried Immodium (aggrevated condition), and Cipro 250mg BID (has taken 9 tablets) from prior episode of diarrhea in Kenya of lesser severity (no effect)*)

- Each feature should be marked up as a separate annotation, and the annotation should include all, but not more than, the text that captures the meaning of the corresponding entry in the feature (e.g., if the key essential is *No blood in stool*, only the underlined text of the following excerpt should be annotated: *No blood or mucus in stool*).

- Annotations should include quantifiers (e.g., *twice, four times, some*), intensifiers (e.g., *mild, severe*), and temporal modifiers (e.g., *two weeks, several years*) that are specified in the corresponding entry in the feature, as well as the object that is being described (e.g., *pain, cough*).

- Annotations should not include articles (e.g., *a, the*) or references to the patient (e.g., *her, he*) that occur at the beginning of note entries, or end punctuation (e.g., periods); however, it is not necessary to fragment annotations if words or characters, such as these, occur within relevant text and do not modify the meaning of the feature entry.

- Annotations may overlap; that is, they may share text with other annotations. For example, negations (e.g., *negative for, no, denies*) frequently will be shared among several annotations. In the phrase *Negative for fever, chills, nausea, vomiting, hematochezia*, the negated nouns refer to different features and should be annotated as Negative for fever, *Negative for chills, Negative for nausea*, etc.

- Mark up every instance of the feature whether it is identical to an existing annotation or not.

History: Describe the history you just obtained from this patient. Include only information (pertinent positives and negatives) relevant to this patient's problem(s).

Karin Moore is a 45 yo F here for nervousness. A few weeks ago she noticed that she was feeling more nervous than usual and that it has been worsening. It is exacerbated by family and work. She feels espeically nervous on Sunday night and Monday morning when she is preparing for the week. She is unable to fall asleep and doesn't want to eat anything, though she does make herself eat. Nothing helps her nervousness. She otherwise denies significant changes in appetite, weight loss, or overall wellbeing. She denies fevers, chills, nausea, constipation, diarrhea, skin changes, racing heart, shortness of breath, dizziness, headaches or rashes.
ROS: otherwise negative
PMH: None; PSH: None
Meds: Tylenol for occasional HA
FHX: Father had an MI, died at 65yo
Allergies: NKDA
SH: Lives at home with husband, mother, and youngest son. Is an english literature professor at a local college. Has 2 drinks/mo, no tobacco or drug use.

Physical Examination: Describe any positive and negative findings relevant to this patient's problem(s).
Be careful to include only those parts of examination you performed in this encounter.

VS: Blood Pressure: 130/85 mm Hg
Heart Rate: 96/min
Gen: No acute distress, conversational, thin
Neck: No thyromegaly, no lymphadeopathy
Heart: RRR, no murmurs, rubs or gallops. Radial pulses +2 bilaterally
Lungs: Clear to ascultation bilaterally, no wheezes
Psych: Well-groomed. Non-pressured speech, linear though process.

Data Interpretation: Based on what you have learned from the history and physical examination, list up to 3 diagnoses that might explain this patient's complaint(s). (...)

General anxiety disorder
Panic disorder
Hyperthyroidism

Table 1: Example of a PN. The dataset features only the history portions of the PNs.

For example, if the feature is *NSAID-use* and the examinee wrote *Uses NSAIDs* as well as *took ibuprofen*, both snippets of text should be annotated. If the exact snippet *Uses NSAIDs* appeared more than once in a note, it should be annotated every time it appears in the note.

- Gender is a special case of a feature and should only be annotated once for the first mention. Subsequent phrases that may be linked to gender such as *she* or *his* should not be annotated.

C Inter-annotator Agreement Per Case

Case	**f1**	**Jaccard**	κ	**yes_no_f1**
201	.72	77.37	.82	.96
202	.90	91.48	.93	.98
203	.93	94.93	.96	.99
204	.92	93.05	.95	.99
205	.71	72.61	.78	.94
206	.88	90.05	.93	.99
207	.79	86.09	.89	.98
208	.85	87.02	.89	.97
209	.88	86.40	.89	.97
210	.87	86.54	.89	.97
Mean	.84	86.55	.89	.97
SD	.075	6.89	.057	.014

Table 2: Inter-annotator agreement per case, where the gold standard is the annotation of the senior annotator. The columns represent (in order): Micro F1 character-position based agreement, Jaccard distance, Cohen's κ coefficient, and a binary F1 score for whether the annotators agree that a given feature expression was found in a PN (1 if found, 0 if not found). As can be seen, the annotators agree very well on whether a feature was found in a PN or not, with some differences in agreement about the exact span of characters that represent that feature.