ORIGINAL RESEARCH

A Clinical Reasoning-Encoded Case Library Developed through Natural Language Processing

Travis Zack, MD, PhD[1,2], Gurpreet Dhaliwal, MD[3,4], Rabih Geha, MD[3,4], Mary Margaretten, MD[5], Sara Murray, MD[6], and Julian C. Hong, MD, MS[2,7]

[1]Division of Hematology/Oncology, Department of Medicine, University of California, San Francisco, CA, USA; [2]Bakar Computational Health Sciences Institute, University of California, San Francisco, CA, USA; [3]San Francisco VA Medical Center, San Francisco, CA, USA; [4]Department of Medicine, University of California, San Francisco, CA, USA; [5]Division of Rheumatology, Department of Medicine, University of California, San Francisco, CA, USA; [6]Division of Hospital Medicine, Department of Medicine, University of California, San Francisco, CA, USA; [7]Department of Radiation Oncology, University of California, San Francisco, CA, USA.

IMPORTANCE: Case reports that externalize expert diagnostic reasoning are utilized for clinical reasoning instruction but are difficult to search based on symptoms, final diagnosis, or differential diagnosis construction. Computational approaches that uncover how experienced diagnosticians analyze the medical information in a case as they formulate a differential diagnosis can guide educational uses of case reports.

OBJECTIVE: To develop a "reasoning-encoded" case database for advanced clinical reasoning instruction by applying natural language processing (NLP), a sub-field of artificial intelligence, to a large case report library.

DESIGN: We collected 2525 cases from the *New England Journal of Medicine* (NEJM) Clinical Pathological Conference (CPC) from 1965 to 2020 and used NLP to analyze the medical terminology in each case to derive unbiased (not prespecified) categories of analysis used by the clinical discussant. We then analyzed and mapped the degree of category overlap between cases.

RESULTS: Our NLP algorithms identified clinically relevant categories that reflected the relationships between medical terms (which included symptoms, signs, test results, pathophysiology, and diagnoses). NLP extracted 43,291 symptoms across 2525 cases and physician-annotated 6532 diagnoses (both primary and related diagnoses). Our unsupervised learning computational approach identified 12 categories of medical terms that characterized the differential diagnosis discussions within individual cases. We used these categories to derive a measure of differential diagnosis similarity between cases and developed a website (universeofcpc.com) to allow visualization and exploration of 55 years of NEJM CPC case series.

CONCLUSIONS: Applying NLP to curated instances of diagnostic reasoning can provide insight into how expert clinicians correlate and coordinate disease categories and processes when creating a differential diagnosis. Our reasoning-encoded CPC case database can be used by clinician-educators to design a case-based curriculum and by physicians to direct their lifelong learning efforts.

Received January 20, 2022
Accepted July 29, 2022
Published online September 7, 2022

KEY WORDS: case-based learning; artificial intelligence; natural language processing; clinical reasoning; medical education.

J Gen Intern Med 38(1):5–11
DOI: 10.1007/s11606-022-07758-0

INTRODUCTION

Clinical reasoning is a core skill of a physician, and its mastery is a central goal in medical training.[1] One approach to teaching the complex cognitive task of reasoning is through a cognitive apprenticeship model where experienced professionals externalize their thought processes.[2,3] This structured and intentional thinking-out-loud conveys the facts and considerations that drive decision-making. One method of augmenting this cognitive training is using published cases which present clinical challenges along with the reasoning of clinical experts. The case-based learning (CBL)[1,4,5] method can be particularly effective in training clinicians when multiple, similar cases are presented and learners are prompted to compare and contrast presentations and management.[6–8]

Teachers who design such learning exercises or curricula must either recall their own cases or find suitable cases in the medical literature. Standard search engines can be used to locate cases based on the final diagnosis, but this process can often be inefficient and incomplete. There are no databases which allow a teacher or learner to search for cases based on the juxtaposition of competing diagnoses or competing categories of illness (e.g., infection vs autoimmune disease). This limits the clinician educator's ability to search for cases with specific diagnostic dilemmas and use such cases for advanced clinical reasoning instruction at the graduate medical education level (residency and fellowship).

Natural language processing (NLP) is a field of computer science that derives meaning and identifies patterns from texts.[9] NLP has the potential to analyze the reasoning process in case reports and identify cases where specific advanced dilemmas (e.g., "find all cases of Systemic Lupus

Erythematosus where cryptococcus was a consideration but not the final diagnosis") are highlighted.[10,11] We hypothesized that computational techniques such as unsupervised machine learning and NLP would allow for meaningful categorization of the NEJM CPC library in a manner that is useful to clinicians and medical educators.[12]

METHODS

The computational methods are described in general terms in the main text of the manuscript. The supplementary appendix contains a more detailed technical description.

Case Library Formation

We collected 2525 cases from NEJM CPC from 1965 to 2020. In this case series, a case is presented to a clinician, who is typically invited because of their clinical expertise. This clinician is asked to explain their problem representation and differential diagnosis construction and to generate a final diagnostic prediction. The final diagnosis derived through additional testing is then revealed.

Text from each case was divided into three sections: Presentation of Case (PoC) which contained the description of the patient's medical data, Differential Diagnosis (DDx), which encapsulated the diagnostic reasoning of the expert clinician, and Final Diagnoses (FD) which contained both the expert's leading diagnostic hypothesis and the final diagnosis determined through additional testing (Fig. 1). Patient data was derived from the PoC section, all medical terms were extracted from the DDx section, and verified diagnoses were ascertained through the FD section.

Category Discovery

Our goal was to identify the different categories of discussion within each DDx section. We used a machine learning approach called unsupervised learning, which uses data-driven (not prespecified) approaches to analyze the natural structure of the data and discover groups of terms that tend to occur together. This differs from having human experts manual create categories a priori, such as cardiac, pulmonary, or hepatic, and opens the possibility of the algorithm perceiving groupings that occur in the data but are unconventional to clinicians.

To focus the NLP algorithm on the medically relevant terms within the NEJM CPC DDx section, we selected a set of 270,666 medical terms from three published term libraries to create one reference library. We identified all terms that occurred within this library in each DDx section.

We used an unsupervised learning approach called Latent Dirichlet Allocation which identifies "categories" which are groups of terms that statistically co-occur across a collection of documents. Each identified category was established by the strength of association with each word in our medical term library (Appendix Figure 2B). For example, category 2 arose based on strong associations with terms *lung*, *heart*, *pleura*, *air*, as well as weaker associations with terms like *tuberculosis* and *hypoxemia*. After these categories were established by the algorithm, a physician assigned a label based on common medical terminology to facilitate recognition (e.g., category 2 ➔ *pulmonary*). After identification of categories, the proportion of each case DDx section attributable to each category was determined. Similarity between cases was measured by overlap of these proportions (Euclidean distance) across these 12 categories.

Final Diagnosis Search Development

Medical diagnoses can be described with differing levels of specificity and can span multiple subjects. For example, "Lupus Nephritis" is a subset of "Systemic Lupus Erythematosus" but also part of the general concepts "Autoimmune Conditions" and "Kidney Disease." To create a search engine that recognizes this nested character of medical diagnoses, we utilized the MeSH Library[13] which is a comprehensive list of medical diagnoses with a hierarchical and redundant format.

Final diagnoses from the sections titled "Final Diagnosis", "Anatomic Diagnosis", or "Pathologic Diagnosis" in each NEJM CPC case were manually assigned (TZ) to the most relevant MeSH Library term. If these sections describing a confirmed pathological or anatomic diagnosis were not present (7% of cases), we used the diagnoses within a section labeled "Expert's Diagnosis" or "Clinical Diagnosis", with verification through manual case review.

Mapping each Final Diagnosis within these sections to the MeSH Library allowed us to understand relationships between diagnoses even if they are not explicitly mentioned in the Final Diagnosis section. For example, cases with a diagnosis of "Granulomatosis with Polyangiitis" or "Polyarteritis Nodosa" will be returned under a search for "Vasculitis," even if the term "Vasculitis" is not mentioned in the Final Diagnosis sections, as these diagnoses appear under "Vasculitis" within the MeSH Hierarchy.

Symptom Identification and Analyses

To generate a list of symptoms and exam findings occurring in a case presentation, we utilized the Human Phenotype Ontology,[14] which is a medical term library that focuses on these topics. We identified terms that were present within the PoC section, using NLP to exclude symptoms that were negated (for example, "the patient had fever but not chills" would identify fever and would exclude chills, which is negated by "but not").

Correlations between symptoms were measured using Pearson correlation, with significance determined by Fisher's exact test, corrected for false discovery rate.[15] Only the 50 most frequently observed symptoms were used in correlation analyses between symptoms and between symptoms and categories to avoid testing pairs with insufficient statistical power. Similarly, only symptoms present in at least 10 cases and Final Diagnoses present within at least 10 cases were included for the symptom to Final Diagnosis correlation analysis.

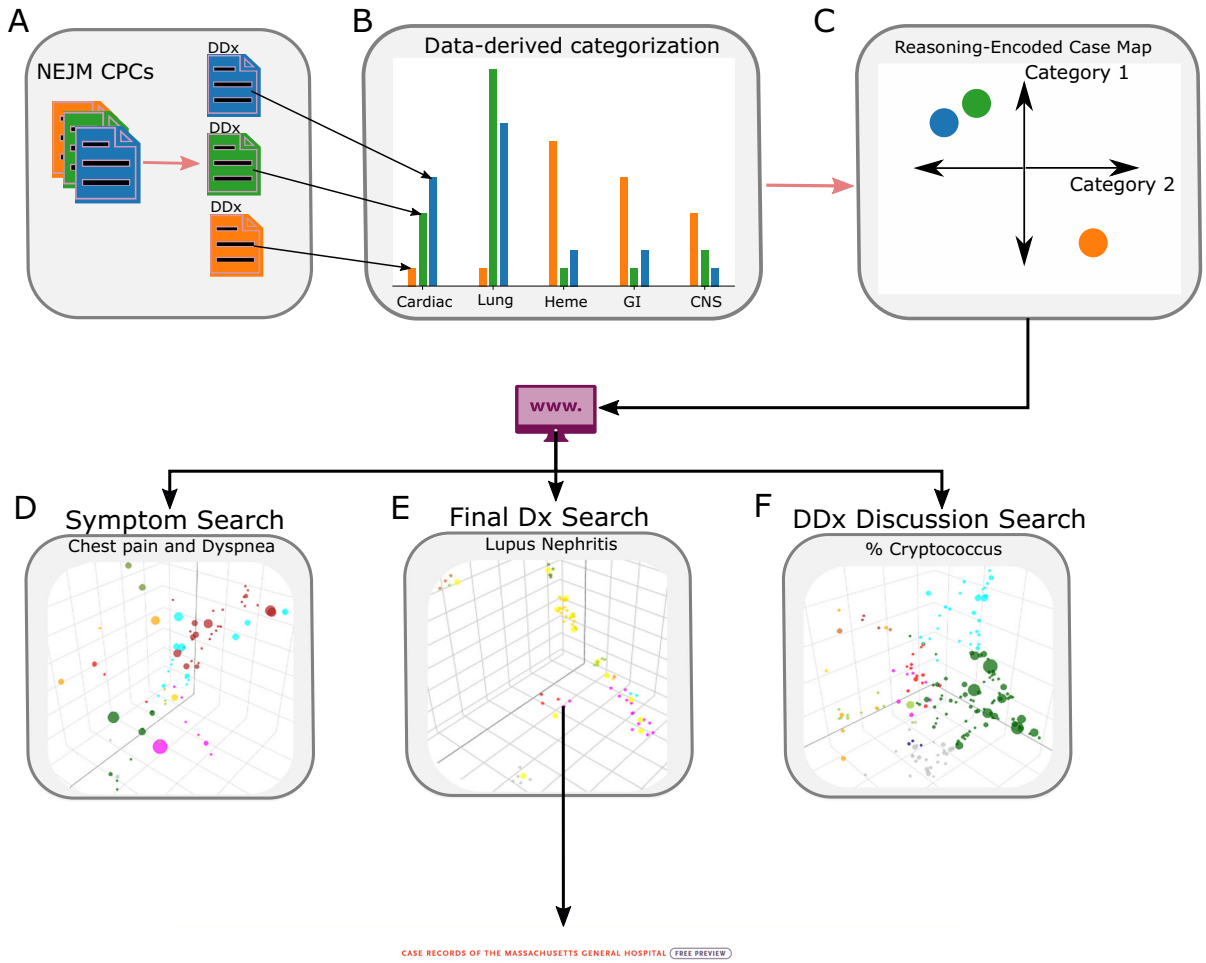

Fig. 1 Data processing concept diagram: We used a data-driven categorization process to identify "categories" within the differential diagnosis section of 2525 CPC cases. The proportions of terms assigned to each category within the DDx section were identified and were used to develop a reasoning-encoded database (top right). The database can be queried using symptoms or exam findings, by the final diagnosis, or by the proportion of the differential diagnosis that is devoted to a concept.

RESULTS

Categories in the DDx Section

By analyzing which medical terms tended to co-occur within case discussions, our approach uncovered data-derived categories in the DDx section that corresponded to recognizable organ systems or pathophysiological processes (Fig. 1). Each of these categories is defined by the medical terms most strongly associated with them (Fig. 2 and Appendix Figure 2B). These "learned" categories were then labeled by the authors using a familiar medical term based on the characteristic terms they contained. Category names were Cardiac, Central nervous system (CNS), Gastro-intestinal Lumen (GI-lumen), Liver, Neoplasia, Renal, Autoimmune, Infectious (acute), Infectious (Chronic/Opportunistic), Pulmonary, Hematology, and Ear/Nose/Throat (EENT). The latter category unexpectedly emerged featuring terms, such as *swallow*, *sinus*, and *larynx*, suggesting these anatomic structures often occur together in case discussions.

Each category was defined by the strength of its association with over 270,000 medical terms (e.g., how often dyspnea appears in the cardiac category). Using these relationships, we defined proximity between categories through a quantitative measure of the similarity in words that comprised each category and visualized this proximity in a two-dimensional projection (Appendix Figure 2A). For example, despite their anatomic relationship, GI-lumen and Liver were quite distant (suggesting pathophysiology involving these organs lead to very different diagnostic discussions), whereas Cardiac and Lung were in close proximity (suggesting that similar terms are used in discussing those conditions). Renal and Autoimmune were closely connected with many overlapping terms, likely

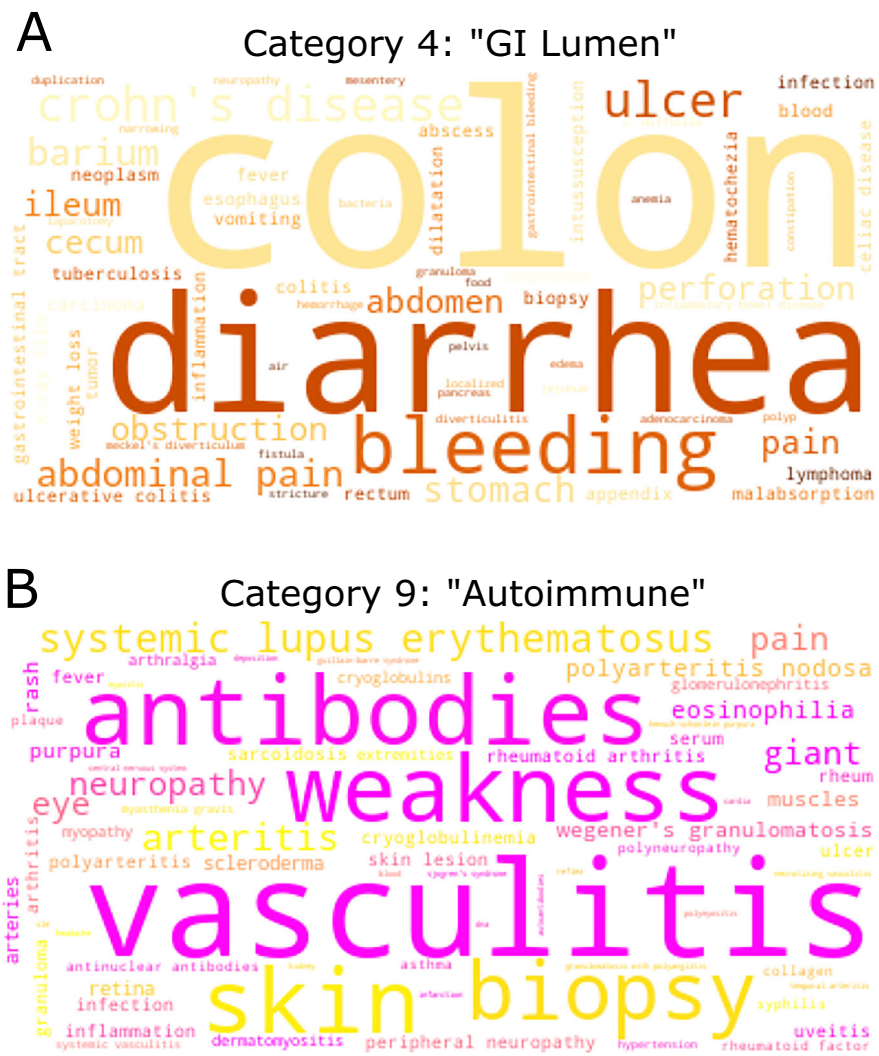

Fig. 2 Each discovered category is defined by a set of terms and the strength of association between those terms and the category. Physician review assigned a category "Label" using a familiar medical term. These labels are meant to facilitate human interpretability but are not used in any subsequent analyses. This figure illustrates two categories. Word clouds for all 12 categories available in Appendix Figure 2B.

representing the frequent involvement of the renal system in autoimmune diseases. Hematology, which encompassed both benign and malignant concepts, was located between the Neoplasia, Infectious, and Renal categories, highlighting the pathophysiologic connection between hematology and oncology, as well as the many hematological manifestations of infections and renal disease.

Category Distribution Based on Final Diagnosis Section

Clinician review assigned all diagnoses contained within the final diagnosis sections of each case within the NEJM CPC library onto 1435 unique MeSH library terms from the 2525 cases. In some instances, more than one Final Diagnosis was registered for a given case because the proximate cause of illness (e.g., acute myocardial infarction) could be accompanied by a predisposing condition found on autopsy (e.g., coronary atherosclerosis) or a downstream clinical consequence (e.g., acute kidney injury). The most frequent Final Diagnosis corresponding to a MeSH category without further subdivisions was *Mycobacterium tuberculosis* (78 cases; 3% of total), followed by *pulmonary embolism*, *myocardial infarction*, and *diffuse large B-cell lymphoma* (Appendix Table 1). This finding mirrors a previous report summarizing ten years of NEJM CPC diagnoses.[16]

Correlations of Symptoms Across Cases

We found a total of 43,291 instances of 1930 unique symptoms within the PoC section of the NEJM CPC corpus. "Pain" was the most common symptom identified (58% of the cases), followed by fatigue/weakness (39%), edema (35%), and fever (30%) (Appendix Table 2). Correlation between each pair of the top 50 most frequent symptoms showed 281 symptom pairs that positively correlated with false discovery rate–

corrected p-value < 0.01 (Appendix Figure 3B, Appendix Table 3). We looked for paired symptoms and final diagnoses with high correlation, suggesting the presence of these symptoms may have high predictive value for these final diagnosis. We found 1815 Symptom-Diagnosis pairs with FDR q-value < 0.01 (Appendix Table 4). This includes familiar pairings such as *clubbing* and *respiratory diseases*, but also rare, more specific symptoms, such as *perseveration* and *viral encephalitis*. A few symptom pairs, such as fever and hypertension or headache and respiratory symptoms, showed a significant inverse correlation.

Relationships Between Presenting Symptoms and DDx Categories Uncovered Through Hierarchical Clustering

We were also interested in what terms in the PoC section would be most characteristic of the categories identified in the DDx section (Appendix Figure 3C). Because the categories were developed independent of the PoC section, we could use similarities between terms present in each case as an independent measure of category relationships. For example, while Cardiac and Lung categories share many symptoms (e.g., *chest pain* and *exertional dyspnea*), cases with reasoning focused on the Cardiac category more frequently involved findings of *peripheral edema* or *hypotension*. Similarly, *lymphadenopathy*, *sweats*, and *night sweats* distinguished the Chronic/Opportunistic Infection category from the Acute Infection category.

UniverseofCPC.com as a Free Resource for Concept-Directed Case-Based Learning

We created a web site (universeofcpc.com) that allows visualization of all CPC cases available in the case library. The site creates a visual display where each dot on a 3D graph represents an individual case and proximity between dots represents the reasoning-encoded similarity between cases. The term "reasoning-encoded" refers to the statistical similarity in medical terms discussed within the DDx section of each case. It does not rely on terms used in the case presentation (the PoC section) or on the eventual final diagnoses (FD section). For example, two cases may have similar presenting symptoms but lead to very different discussions about their differential diagnosis based on specific features that are selected for analysis by the discussant. Such cases will be distant from one another in our 3D graph as the reasoning within the DDx sections would be divergent. Conversely, cases that have disparate presentations but lead to similar discussions will be near each other within this visualization.

Three search functions were developed: Symptom search (where a user can search for cases that contain one or more symptoms such as hemoptysis), Final Diagnosis search, and DDx relevance search. To maintain diagnostic mystery, the Final Diagnosis search function can also identify several cases within the "reasoning-encoded" proximity of cases with this final diagnosis and mask which of the cases contains the specified diagnosis. The DDx relevance search allows users to search for cases where specific diagnosis (e.g., histoplasmosis) was discussed within the DDx section, with the size of each dot in the search results related to the proportion of the differential discussion dedicated to that diagnosis.

DISCUSSION

Case records with expert clinician analysis are a valuable repository of clinical reasoning. However, these case series are underutilized for advanced reasoning instruction because there is no method to quickly identify cases centered around specific reasoning dilemmas or clinical presentations. We demonstrated that natural language processing can uncover characteristics of the reasoning process contained within these cases and can be used to create a database that allows a clinician to search through a 55-year case library and make queries based on symptoms, diagnoses, or similarities in differential diagnosis construction.

While reviewing weekly cases that cover a broad and rotating subject matter is important for foundational learning, learners in residencies and fellowships are often faced with specific recurrent problems (e.g., is this fever caused by an infection or autoimmune disease?). Allowing easy curation of case collections where experts faced these same dilemmas has the potential to support and advance the reasoning of trainees. In Table 1, we outline 3 scenarios where universeofcpc.com could be used for case-based instruction in a rheumatology fellowship program.

Natural language processing is a branch of machine learning that can be divided into two approaches, each with different aims. In *supervised* machine learning, humans provide the algorithm with examples that define each category (e.g., cardiovascular or gastrointestinal) and the machine is asked to identify features that predict belonging to each category. In *unsupervised* machine learning, techniques are applied on ungrouped data to establish categories without any preconceived notions about what associations or signals may define a given category (e.g., the algorithm is not "taught" in advance what cardiovascular means or looks like, or even that "cardiovascular" may be a category at all). While the former approach is useful for predicting known features, information for the latter is derived from the inherent structure present within the data. The data creates these categories without us labeling them. For example, our unsupervised approach thought terms like *thrombosis*, *hemorrhage*, and *iron* were important to consider in the category which we later named Liver; these terms may not have been chosen in predefined models to specify a Liver category. Similarly, this approach created separate categories for acute infections and chronic/opportunistic infections, which highlights differences

Table 1 Applications for a Database of Reasoning-Encoded CPC Cases

Feature	Primary function	Secondary functions	Example educational applications
Symptom search	Search for any number of symptoms to find cases where ALL symptoms were present	Cases weighted by how often searched symptoms were discussed in DDx discussion	1. Create case series to practice assessment of hemoptysis and fever 2. Create case series for the symptom of polyarthritis
Final diagnosis search	Identify cases by specific diagnosis or diagnosis category	- Return most "similar" cases based on natural language processing of DDx discussion - Mask Final Dx of searched and similar cases to enhance the diagnostic challenge	1. Create a list of lupus nephritis cases to teach differences in presentation and management 2. Create a list of 5 cases with similar diagnostic reasoning but different diseases (e.g., spondyloarthritis, rheumatoid arthritis, and psoriatic arthritis), of which only one case has rheumatoid arthritis as the final diagnosis
DDx browser	Identify cases where a specific diagnosis was considered and deliberated during DDx discussion, regardless of final Dx	Weighs cases based on how "strongly" the searched diagnosis was considered during differential diagnosis discussion in the case	1. Create case series where pulmonary sarcoidosis was strongly considered to practice the clinical reasoning of including or excluding sarcoid as final diagnosis 2. Create list of cases where systemic sclerosis was considered to help create connections between systemic sclerosis and other diseases with similar presentations

in how these categories are analyzed by discussants. Many other observations can be made through study of this inherent structure, which may provide insight into how clinicians organize diagnostic discussions.

Case series designed for education purposes, such as the NEJM CPC, can represent a rich source for understanding clinical problem-solving that can be used to not only to train physicians, but also to train AI systems in decision support and diagnostics. This unbiased machine learning approach captured medical term relationships that occur within the reasoning processes outlined by invited experts during a diagnostic reasoning exercise. In future work, we hope to build on these techniques to analyze how specific information is utilized and processed by expert diagnosticians.

Limitations of this study include the NEJM CPC's emphasis on rare diagnoses and complex presentations of common conditions. Therefore, universeofcpc.com is best understood as a representation of diagnostic reasoning in the setting of challenging cases rather than more routine scenarios. Another limitation is the categories were derived based on these rare diagnoses and therefore may not generalize to diagnostic reasoning in everyday medical practice. It will be important to apply these approaches to other internal medicine case series or structured versions of real-world clinical documentation to determine how stable the categories formed by these unbiased approaches are outside of the NEJM CPC corpus. We were limited in our ability to extract accurate quantitative features from case presentations, such as vital signs or laboratory values. We are building models to incorporate such information which is critical in diagnostic reasoning. Finally, time-intensive manual annotation determined the final diagnoses for each case which could limit further expansion of this resource.

This proof of concept study demonstrates the use of NLP and unsupervised machine learning for categorization of a clinical reasoning case library. Future work should examine how clinician educators utilize reasoning-encoded case libraries to create case-based exercises or curricula. Additionally, we envision future NLP analyses on case report series that can help elucidate the connections between diseases, presenting symptoms, and pathophysiology.

Corresponding Author: *Travis Zack, MD, PhD; Bakar Computational Health Sciences Institute, University of California, San Francisco, CA 94158, USA (e-mail: travis.zack@ucsf.edu).*

Supplementary Information *The online version contains supplementary material available at https://doi.org/10.1007/s11606-022-07758-0.*

REFERENCES

1. **Trowbridge R, Rencic J, Durning S**. *Teaching Clinical Reasoning*. Philadelphia, American College of Physicians; 2015.
2. **Stalmeijer RE, Dolmans DHJM, Wolfhagen IHAP, Scherpbier AJJA**. Cognitive apprenticeship in clinical practice: can it stimulate learning in the opinion of students? Adv Health Sci Educ. 2009;14(4):535. https://doi.org/10.1007/S10459-008-9136-0
3. **Stalmeijer RE, Dolmans DHJM, Snellen-Balendong HAM, Van Santen-Hoeufft M, Wolfhagen IHAP, Scherpbier AJJA**. Clinical teaching based on principles of cognitive apprenticeship: Views of experienced clinical teachers. Acad Med. 2013;88(6):861-865. https://doi.org/10.1097/ACM.0b013e31828fff12
4. **Donner RS, Bickley H**. Problem-based learning in American medical education: An overview. Bull Med Libr Assoc. 1993;81(3):294-298.
5. **Thistlethwaite JE, Davies D, Ekeocha S, et al.** The effectiveness of case-based learning in health professional education. A BEME systematic review: BEME Guide No. 23. Med Teach. 2012;34(6):142-159. https://doi.org/10.3109/0142159X.2012.680939
6. **Eva KW, Neville AJ, Norman GR**. Exploring the etiology of content specificity: Factors influencing analogic transfer and problem solving. Acad Med. 1998;73: S1-5. https://doi.org/10.1097/00001888-199810000-00028
7. **Mylopoulos M, Steenhof N, Kaushal A, Woods NN**. Twelve tips for designing curricula that support the development of adaptive expertise.

Med Teach. 2018;40(8):850-854. https://doi.org/10.1080/0142159X.2018.1484082

8. **Lessing JN, Pierce RG, Dhaliwal G**. Teaching More About Less: Preparing Clinicians for Practice. Am J Med. 2022;135(6):673-675. https://doi.org/10.1016/J.AMJMED.2022.01.060
9. **Wu S, Roberts K, Datta S**, et al. Deep learning in clinical natural language processing: A methodical review. J Am Med Inform Assoc. 2020;27(3):457-470. https://doi.org/10.1093/jamia/ocz200
10. **Prakash A, Zhao S, Hasan SA, et al.** Condensed memory networks for clinical diagnostic inferencing. *31st AAAI Conf Artif Intell AAAI 2017*. Published online 2017:3274-3280.
11. **Lehman E, DeYoung JB, Barzilay R, Wallace BC**. Inferring which medical treatments work from reports of clinical trials. *NAACL HLT 2019 - 2019 Conf North Am Chapter Assoc Comput Linguist Hum Lang Technol - Proc Conf*. 2019;1(Figure 1):3705-3717. https://doi.org/10.18653/v1/n19-1371
12. **Hassan S**. About clinicopathological conference and its' practice in the school of medical sciences, USM. Malaysian J Med Sci. 2006;13(2):7-10.
13. Medicine NL of. Medical Subject Headings. Accessed January 5, 2020. https://www.nlm.nih.gov/databases/download/mesh.html
14. **Köhler S, Carmody L, Vasilevsky N, Jacobsen JOB et al.** Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources. Nucleic Acids Res. 2019;47(D1):D1018-D1027.
15. **Hochberg B.** Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc. 1995;61(1):1-15.
16. **Falagas M.** An analysis of the published Massachusetts General Hospital case records (1994-2004). Am J Med. 2005;118(12):1452-3. https://doi.org/10.1016/j.amjmed.2005.06.027