Published in final edited form as:
J Biomed Inform. 2009 October ; 42(5): 781–789. doi:10.1016/j.jbi.2009.02.004.

Tracking medical students' clinical experiences using natural language processing

Joshua C. Denny[a,b,*], **Lisa Bastarache**[a], **Elizabeth Ann Sastre**[b], and **Anderson Spickard III**[a,b]

[a]Department of Biomedical Informatics, Vanderbilt University Medical Center, Eskind Biomedical Library, Room 442, 2209 Garland Ave., Nashville, TN 37232, USA

[b]Division of General Internal Medicine, Department of Medicine, Vanderbilt University Medical Center, Nashville, TN, USA

Abstract

Graduate medical students must demonstrate competency in clinical skills. Current tracking methods rely either on manual efforts or on simple electronic entry to record clinical experience. We evaluated automated methods to locate 10 institution-defined core clinical problems from three medical students' clinical notes ($n = 290$). Each note was processed with section header identification algorithms and the KnowledgeMap concept identifier to locate Unified Medical Language System (UMLS) concepts. The best performing automated search strategies accurately classified documents containing primary discussions to the core clinical problems with area under receiver operator characteristic curve of 0.90–0.94. Recall and precision for UMLS concept identification was 0.91 and 0.92, respectively. Of the individual note section, concepts found within the chief complaint, history of present illness, and assessment and plan were the strongest predictors of relevance. This automated method of tracking can provide detailed, pertinent reports of clinical experience that does not require additional work from medical trainees. The coupling of section header identification and concept identification holds promise for other natural language processing tasks, such as clinical research or phenotype identification.

Keywords

Natural language processing; Medical education; Concept identification; Education portfolios; Competency assessment; Experience tracking; UMLS

1. Introduction

Medical educators have long recognized the need for more rigorous accounting of what trainees are learning during their clinical years [1,2]. To meet this need, medical schools are developing education portfolios to monitor learners' progress. An important component of the portfolio is the trainee-patient encounter, typically assessed through manual case logs. Learners use handwritten log books, score sheets of clinical data, or personal digital assistants (PDA) to create case logs of encounters with patients including location and

[*]Corresponding author. Address: Division of General Internal Medicine, Department of Medicine, Vanderbilt University Medical Center, Nashville, TN, USA. Fax: +1615 936 1427. josh.denny@vanderbilt.edu (J.C. Denny).

demographic data, diagnosis, severity of illness, and/or procedures performed [3–6]. Program directors may compile the information to assess learner performance and case mix to guide personal learning and curriculum revision.

Manual tracking systems are limited for several reasons. In the context of the fast paced clinical setting, trainees find the process of uploading information to the portfolio as “intrusive busywork” [7]. Indeed captures rates of clinical information by manual means are poor [3,5,6,8]. In addition, manual logging represents an incomplete record of the learners’ experience. The typical system allows one to five diagnostic per patient [4,6,9,10]. Finally, teachers often disagree with learners about the primary diagnosis of the case [4,9]. What is needed is a system that automatically captures all concepts that a learner covers on each case and organizes the data to provide learners and teachers with meaningful reports of the learner’s experience. We hypothesized that we could more accurately identify core clinical problems using natural language processing tools than simple string searching in a corpus of clinical notes. We believe this method could serve as a valid alternative to manual tracking of trainee-patient encounters.

2. Background

The Learning Portfolio system is a standalone web application that receives data from the electronic medical record (EMR). All clinical documentation (e.g., inpatient history and physical examinations, discharge summaries, outpatient clinic notes, and procedure notes) generated in the EMR by housestaff physicians or medical students is automatically forwarded to the Portfolio system. These notes are sent in an XML-like structure containing the note itself (generally in free-text “natural language” format) with structured metadata such as the medical record number, author information, and patient information. Trainees can also create notes in Portfolio without using the EMR, primarily for use in external clinics and hospitals. Portfolio indexes notes by patient and trainee. Teachers (attending and housestaff physicians) review these notes to give feedback to learners [11]. Learners review their prior write-ups to monitor progress. Administrators review case mix to guide program revision and generate program reports, such as housestaff procedure logs that are automatically generated as physicians document procedures in the EMR.

Most clinical documentation is expressed in natural language content. Many authors have used natural language processing tools to derive computable interpretations from unstructured text [12–16]. The Portfolio system uses the KnowledgeMap concept identifier (KMCI) to identify biomedical concepts from these natural language clinical notes [13]. The KMCI algorithm has been described previously [13,17,18]. It employs the National Library of Medicine’s Unified Medical Language System (UMLS) knowledge resources, which provides, for each concept, semantic information, synonymy, and relationships to other concepts. The KMCI system bears some similarities to other current medical concept indexing systems such as MetaMap [19], the Mayo Vocabulary Processor (MVP) [16], and HITEx [20]. It uses part-of-speech information to develop a shallow sentence parse, and, similar to MetaMap, performs variant generation and normalization using the SPECIALIST Lexicon and related tools. The KMCI system was designed particularly for poorly-formatted documents containing ad hoc abbreviations and underspecified concepts often found in clinical notes (e.g., the document phrase “ST” implying the “ST segment” of an electrocardiogram instead of abnormal finding “ST elevation”) using syntactic and semantic rules in combination with automatically-derived corpus-specific prior probabilities [13]. Using probabilistic information and concept co-occurrence data derived from PubMed, KMCI can map ambiguous strings such as “CHF” to the UMLS concept C0018802 “Congestive heart failure” in an echocardiogram report but to the concept C0009714 “Congenital hepatic fibrosis” in a document discussing infantile polycystic kidney disease (a genetically related condition to congenital hepatic fibrosis). The KMCI system uses semantic and syntactic rules to map phrases such as “small and large intestine” to concepts C0021852 “Small Intestine” (instead of C0700321 “Small”) and C0021851 “Large Intestine.” The KMCI system is currently used to index medical curricular documents and clinical documentation [13,17,18]. Prior analysis has shown that it performs favorably compared to Meta-Map [13].

By indexing notes to UMLS concepts, Learning Portfolio allows teachers and trainees to quickly locate experiences and concepts relevant to education goals. Led by a team of associate deans and master clinical educators, the Vanderbilt School of Medicine has prioritized 28 core clinical problems (CCP) to be mastered by graduating medical students as part of the Association of American Medical Colleges (AAMC) Clinical Transaction Project [1]. The CCPs address common patient presentations that range from serious illnesses to every-day complaints (see Table 1 for a list of 10 of these topics). For each of the 28 problems, the team developed a set of learning objectives that included 30–60 descriptive elements for each. The representative objectives include specific history items, physical exam findings, differential diagnoses and appropriate diagnostic evaluation that a finishing medical student should have learned. For example, the topic of back pain includes the representative concepts “history of cancer,” “straight leg raise exam,” and “spinal cord compression.”

3. Methods

We used the set of clinical notes from three finishing fourth year students to evaluate the ability of an automated algorithm to identify 10 CCPs. We have developed a system to analyze the students’ clinical experiences using extracted UMLS concepts from clinical notes. This study applies the system to clinical notes derived from inpatient and outpatient clerkships in internal medicine and pediatrics during third and fourth years of medical school. The Institutional Review Board approved the study.

3.1. Identifying core clinical problems from clinical notes

Clinical notes are highly structured documents that include many commonly recognized sections, such as “Chief Complaint,” History of Present Illness,” and “Physical Exam,” among others. To take advantage of the prose structure of clinical notes, we developed segmentation algorithms to categorize clinical notes by major section headings. This study used an early version of the SecTag algorithm, which recognizes both explicitly labeled and unlabeled (implied) sections in the clinical text [21]. The SecTag algorithm uses a locally-developed section header terminology, which is concept-oriented and hierarchical [21].

Thus, SecTag recognizes that synonymous section labels such as "history of present illness" and "HPI" both indicate the "history of present illness" section. After identifying note sections, we processed notes with KMCI to encode these documents with UMLS concepts. For each note, KMCI outputs a list of UMLS concepts, their location and section heading, and semantic information about each concept (i.e., identifying a concept as a disease or pharmaceutical, for instance).

At the time of the study, Portfolio was deployed in internal medicine and pediatric clerkships. Consequently, we selected 10 of the 28 CCPs that are addressed primarily in these clerkships (listed in Table 1). We converted the available educator-created list of learning objectives into a list of UMLS concepts using KMCI. Two authors (JD and AS, both physician educators) manually reviewed the list of UMLS concepts, removing unimportant concepts, and augmented with select expansions sets using selected semantic relationships defined with the UMLS (i.e., those defined in the MRREL file) [22]. Content experts, including the original authors of the CCP learning objectives, assisted in revising the concept lists. Each list was used as a query to retrieve matching clinical notes, using the set of concepts found in all notes. Each concept query contained a median of 115 (range 72–467) individual UMLS concepts (see Fig. 1). There were a total of 1463 unique concepts in the 10 CCP concept lists.

3.2. Evaluating core clinical problem rankings

For each CCP, clinician reviewers (authors AS, BS, and JD) independently scored each document as either a primary reference to, relevant reference to, or irrelevant reference to the CCP being scored using a web interface (see Fig. 2). Each reviewer was board certified in internal medicine and actively involved in medical education. These rankings served as the "gold standard" rankings, and were the primary outcome of the study. Documents did not need to specifically mention the CCP by name or concept; documents could be considered a primary reference to the topic if the document discussed key differentials, related presentations, or treatment and evaluation plans for the clinical topic. For example, key diagnostic considerations for the "chest pain" CCP include myocardial infarction, pneumonia, pneumothorax, and cardiac tamponade, among others. Each of these conditions can present primarily with dyspnea (and the absence of chest pain). Thus, a note discussing an admission for dyspnea in a patient with active coronary disease is classified as a primary reference to the "chest pain" topic.

Following the creation of the gold standard, we compared several ranking schemes to identify highly relevant documents. The baseline ranking for comparison was performed with a search for the topic concept only. Each document was ranked using several different algorithms: the total concepts matched in each document (i.e., the count of matching concepts found by KMCI), the total asserted concepts (regardless of meaningfulness or correctness), the total correct concepts matched in each document, the total correct asserted concepts, and the total correct concept matches marked as meaningful by reviewers. Finally, three weighting algorithms were applied. The first was an expert weighting scheme in which two authors (JD and AS) assigned weights to each section heading and concept semantic type (e.g., "Disease or Syndrome"). The second weighting scheme was a multivariable logistic regression model using a leave-one-out cross validation training method [23]. For each CCP, the logistic regression model was trained using all problems except the CCP; then it was used to predict the CCP. The inputs to this model were the frequency of matching concepts found within individual note sections. The third weighting scheme used a term frequency–inverse document frequency (TF–IDF) model to weight concepts by their uniqueness [24]. Our hypothesis was that certain common concepts, such as “abdominal X-ray” (which could apply to several CCPs, including back pain, fever, abdominal pain, and dysuria) should carry less import than less common concepts, such as “straight leg raise” (a physical exam maneuver employed in the setting of back pain).

3.3. Evaluating recall and precision of the KnowledgeMap concept indexer

The reviewers scored each concept reference identified by KMCI as either correct and meaningful to the relevant CCP, correct but unimportant to the CCP, or an incorrect concept match (i.e., KMCI selected the wrong UMLS concept for the document phrase). The reviewers also scored each concept as asserted (e.g., “patient complains of chest pain”) or negated (e.g., “patient *denied* chest pain”). Precision was calculated as the total correct CCP concepts divided by all CCP concepts identified by KMCI. After ranking each note according to the CCPs, a concept-level recall evaluation was performed on the full text of all notes ($n = 117$) written by one student. All concepts matching CCP target concepts were color-coded and labeled by a computer program. Two reviewers (authors LB and JD) manually inspected all sentences across all of the student’s notes to identify CCP concepts not identified by KMCI. In scoring, the reviewers considered as false negatives any concept not identified or any document phrase in which a CCP concept would have been a better match than that supplied by KMCI, even if the concepts identified by KMCI were acceptable. For instance, “stone in the common bile duct” indicates a false negative for the target concept “biliary stone” if KMCI identified “Calculus, NOS” and “common bile duct.” All 117 documents (4421 sentences) were evaluated by both reviewers and a consensus approach was taken to identify missed concepts. Recall was calculated as the total correct CCP concepts identified by KMCI divided by the total number of CCP concepts identified by KMCI and human review.

3.4. Statistical analysis

The primary outcome was the number of documents of primary reference and relevant reference to the CCPs. Scoring methods were compared by calculating receiver operator characteristic curves. We compared the contribution of different note sections via multivariable ordinate logistic regression. To calculate inter-rater agreement, each reviewer independently reviewed all documents for two CCPs, using Cohen’s Kappa as a measure of agreement. All statistical analyses were performed with Stata, version 9.2 (StataCorp, College Station, TX).

4. Results

The students in this study had a total of 290 notes recorded in the Learning Portfolio system from the Internal Medicine and Pediatrics clerkships and selected fourth year electives. These notes included 81 inpatient “History of Physical Examination” admission notes, 202 outpatient clinic notes, two “Procedure Notes,” three discharge summaries, and two consult notes. The physician reviewers’ scoring of primary and relevant documents is found in Table 1. The percent agreement between the two reviewers was 96.2%; the Kappa was 0.70 ($p < 0.001$).

4.1. Core clinical problem ranking performance

Table 1 shows that each student saw between 2 and 41 cases primarily about each CCP (i.e., a primary note), but an additional one to 28 cases that were relevant to each CCP (i.e., a relevant note). The number of documents relevant to a CCP and the average number of concepts/note varied by CCP. Table 2 shows the results of the ranking algorithms. All ranking methods using broad concept queries performed superior to the topic search alone. The best performing algorithms were all asserted concepts, and all correct asserted concepts, the expert weighting scheme, and the TF–IDF algorithm with asserted concepts. Restricting ranking algorithms to asserted concepts performed better than including both asserted and negated concepts ($p = 0.02$). However, restricting the total to meaningful concepts (those selected in the scoring process as being important to the CCP) did not result in a superior ranking.

Table 3 shows the impact each individual section’s concepts had on the overall relevance of the notes. Concepts found in the “chief complaint” section (adjusted odds ratio (OR) 3.42) were the most influential, followed by those in the “history of present illness” (OR 2.00) and the “assessment and plan” sections (OR 1.93). Taken together, these three sections identified 243 of the 253 (96%) notes scored as a primary reference (AUC 0.91, 95% CI 0.89–0.93) and 506 of the 568 (89%) notes scored as relevant notes (AUC 0.87, 95% CI 0.85–0.88). Concepts found in the “review of systems” and “physical examination” sections were unlikely to predict relevance. Out of 558 note-CCP evaluations containing concepts matches in only the “review of systems” and “physical exam” sections, only 1 (0.3%) was considered a primary reference and 27 (8.6%) relevant references.

4.2. Recall and precision of concept identification

KMCI found a total of 8086 concepts in the student’s corpus of documents that matched at least one of the CCPs. Of these, 7461 (precision 92.3%) were correct concept matches (i.e., matched to the appropriate UMLS concept by KMCI). Eighty-eight percent of correct concept matches were judged meaningful to the topic being scored; approximately equal numbers of negated (47%) and asserted (53%) concepts were marked as meaningful. A review of 20 notes indicated the section segmentation algorithm appropriately identified the prose sections (e.g., “Chief Complaint,” “Past Medical History”) of the notes with 97% sensitivity and 99% specificity.

Three errors in acronym disambiguation comprised 53% of all false positives in concept identification: misidentifying the document phrase “CN” as “constipation” instead of “Cranial Nerves” (141 occurrences, or 23% of all false positives), mapping “BP” to “hypertension” instead of “blood pressure” (105 occurrences, 17% of total), and mapping “LE” to “Lupus erythematosus” instead of “Lower extremity” (88 errors, 14% of total). In the UMLS, “BP” is listed as an equivalent synonym for the concept “hypertension”, instead of a “suppressible synonym.” Thus, when KMCI saw the string “BP” two exact-matched concepts were considered: the correct concept “blood pressure” and the incorrect concept “hypertension.” However, the high prevalence of correctly-matched “hypertension” concepts in the corpus caused KMCI to strongly favor the hypertension concept for “BP.”

The recall evaluation was performed over all notes on a single student (*n* = 117). KMCI failed to identify 198 target CCP concepts in 109 unique sentences. There were 2279 total target concepts, resulting in a recall of 91.3%. Table 4 shows a failure analysis of these false negatives. The single most common error (40% of all errors) resulted from incorrectly parsing form data inserted as review of systems data into a common note template. This form data consisted of the exact same 10 lines of text, with different concepts separated only by spaces as one large sentence. In attempting to identify separate concepts, KMCI tended to group these words in ways alternate to that intended.

4.3. Student coverage of important medical concepts

Tables 5 and 6 show the frequency of selected diseases and findings, respectively, documented in the students’ notes that were also part of the concept queries for these 10 CCPs. The students mentioned 187 unique diseases and 134 unique findings. Diseases were more commonly asserted (92%) than findings (53%). Diseases were often found in the “assessment and plan” (21%), “history of present illness” (16%), and “past medial history” (14%) sections. Findings were commonly found in the “review of systems” (38%), “physical examination” (25%) or “history of present illness” sections (17%).

5. Discussion

We have demonstrated a novel method to identify a learner’s clinical experiences by locating select biomedical concepts from clinical notes created in the normal clinical workflow. The Learning Portfolio system enables a comprehensive capture of clinical work done by a trainee, providing a rich experience log not possible with traditional manual recording methods. We found several algorithms utilizing a broad concept query performed significantly better than using only a search for the clinical problem itself. A relatively simple ranking algorithm calculated by summing all concepts found in the narrative sections of the note (chief complaint, history of present illness, and assessment and plan sections) yielded an AUC of 0.91. A potentially more accurate method, summing all asserted concepts, ignoring notes containing matches only in the physical examination or review of system sections, performed as well as more complex weighting schemes with an AUC of 0.94. This method would require adding a negation algorithm to the current system, which combines a section tagging followed by concept identification. A number of such algorithms have been published [25–27], some of which have already been used with KMCI [17]. A future competency tracking system could automatically aggregate the highest-ranking documents for each competency for review of a trainee or approval by a mentor. In addition, a concept query across a program could quickly identify learners who lack sufficient exposure to clinical problems of interest.

Concepts located in the more narrative portions of the note (e.g., the “history of present illness” and “assessment and plan” sections) conferred more relevance to a given CCP than those taken from "list" sections such as "past medical history" or "review of systems." These narrative sections offer essential information such as the trainee's diagnosis and prioritization of problems, clinical reasoning, management plans, and communication skills [28]. Concepts found in review of systems and the physical examination sections were poor predictors of note relevance, possibly due to the high presence of negated concepts found in these sections. Most electronic note templates in our EMR include these sections, and thus many matching concepts in these sections may be included by default rather than the learner's thoughtful addition. Despite the lack of relevance of these sections, a simple algorithm of identifying only the asserted algorithms performed well. Indeed, the asserted concepts in these largely negated sections represent intentionality.

Unlike other systems, Portfolio requires no additional manual entry and provides a robust report of important concepts covered by a learner across all clinical notes with 91% recall and 92% precision. Sequist et al. extracted information from the EMR to document outpatient resident experiences [10]. However, they relied on provider billing codes to find a small number of diagnoses in each document. This hindered their ability to capture "additional characteristics of patients which make important contributions to the diversity of resident education" [10]. Instead of rendering a single, high-level concept such as "abdominal pain" in a patient who presents with epigastric pain, the Portfolio system automatically locates many important concepts related to the main issue of abdominal pain while preserving the meaning of the original notation. Other concepts documented in the case may also be relevant to the trainee's learning, such as a heart murmur or a diagnosis of urinary tract infection, but may not be a primary consideration in the case.

Tracking clinical experience can reveal a trainee's progress along established learning objectives. For instance, the list of learning objectives for the back pain CCP defines key historical items and physical examination maneuvers to be performed by medical students in their approach to a patient with back pain. In this study, a concept report for "back pain" yielded a patient who presented with severe back pain. The student appropriately documented inquiry to common "red flags" of acute back pain (history of cancer, fever, or numbness), and he appropriately documented a straight leg raise exam, but the student failed to document a lower extremity neurological exam. This provides an opportunity for feedback and planning for future demonstration of a lower extremity exam in a patient with back pain to achieve training goals. This depth of analysis, allowing an improved learning experience, is not possible using a manual tracking system. Educators could identify "key concepts" associated with each competency, or derive them from peer data (e.g., students often document neurologic exams on patients with back pain). Trainees and mentors could quickly find key concepts not covered in trainee notes, potentially directing them to other educational experiences.

The recall and precision of KMCI on this clinical note corpus is similar to prior results for KMCI and other available concept indexers. Previous studies on KMCI revealed recalls of 0.82–0.90 and precisions of 0.89–0.94 for general concept identification on medical curricular documents [13] and electrocardiogram impressions [18]. In an early feasibility analysis of a UMLS concept indexing, Nadkarni et al. found a true positive rate of 76% over 24 documents [29]. Friedman et al. found that MedLEE found a recall of 0.83 and a precision of 0.89 for known UMLS concepts, which was equal to or superior to that of human experts [14]. Meystre and Haug reported on use of MMTX to identify selected medical problems with a recall of 0.74 and precision of 0.76 using the entire UMLS [30]. The KMCI system's use of document-specific disambiguation techniques based on concept co-occurrence data and concept frequency is unique; prior analysis has shown its utility for improving recall and precision [13,18].

We report novel approaches that may be applied broadly to extract information from clinical documentation. Unlike some other indexing systems, the current system uses section header identification to provide section context for each concept. The weighting schemes used in this study illuminate the relevance of individual concepts to a given topic. We discovered that negated concepts and those concepts located only in the review of systems have less impact on the priority topics of an individual's note. Such findings provide contextual understanding of the clinical document that may generalize to clinical research. For example, one may assign less importance to positive findings found only in the review of systems, as these concepts are not often as thoroughly vetted by the note's author. Similarly, understanding the context of a concept holds potential for answering variety of important questions in clinical medicine (what are the most common medications associated with a disease?); in quality improvement (do patients with a given condition receive the appropriate medications?); or in genomic research (what is an individual's phenotype?). Finding the answers requires that one distinguish between an individual's medications taken and his or her allergies, past medical history, and other contexts of the recorded note, such as family history. Combining this system with sentence-based contextual methods such as employed by MedLEE [14] and ConText [31] may help address these challenges.

Some errors in concept identification discovered in this study are easily addressable (e.g., via addition of new synonyms) while others, such as errors in spelling, phrase parsing, and disambiguation, present significant and ongoing challenges to natural language processing. Use of section location could assist in concept disambiguation; for instance, the phrase “CN” is much more likely to mean “cranial nerve(s)” than “constipation” in the physical exam section of a note. Some of the most frequent causes of false positives were due to imprecise terminology synonymy. Selection of the concept “hypertension” for the document string “BP” (17% of all false positives) occurred due to the presence of the string “BP” as a synonym (i.e., a SUI) for hypertension. The string “BP” is rarely (if ever) used by clinicians to indicate “hypertension” and could be classified as a “suppressible synonym” in the UMLS. Similarly, “sick” is present in the UMLS as a synonym for a number of concepts, including “vomiting” and “influenza”; however, this also should likely be considered a “suppressible synonym” also. Inaccurate disambiguation and lack of concept inference were sources of false negatives. Interpreting phrases such as “burning pain radiating down right leg” to mean “sciatica,” or inferring the presence of implied words (e.g., mapping the document text “stone” to “biliary stone” in the appropriate circumstance) could be assisted with expanded concept co-occurrence data and the addition of clinical reasoning rules. Finally, certain poorly formed sentence structures (such as long lists of nouns and adjectives without punctuation or prepositions) caused frequent errors by combining two adjacent nouns intended to be separate. Irregular sentence structures, typing errors, ad hoc abbreviations, and poor grammatical constructions are likely to become more common as busy clinicians are increasingly typing more information directly into EMRs.

Limitations caution interpretation of this study. This study examined notes from three medical students, each rotating in two to three medical centers. It may not generalize to other students, institutions, or clerkships. Indeed, this method may not work for clerkships in which clinical notes are not a key measure of student performance and learning, such as many surgical clerkships. Furthermore, the section tagging, concept identification, and ranking algorithms may not perform as well at different institutions or with different note formats. The gold standard rankings of relevance were scored by the authors, one of whom was involved in the creation and design of some of the CCPs. While the Portfolio system does allow input of notes without an EMR, it does require notes in an electronic form. Some medical centers may not have an EMR, and inputting notes into the system would be a significant challenge in the busy clinical workflow. Given the superior performance of asserted concepts compared with negated ones, future versions of KMCI should be coupled with a negation detection algorithm, which was not done in this study. A previous study of electrocardiogram impressions revealed good performance using a modified version of the NegEx negation detection algorithm [17,27]. In this study, we used learning objectives for a subset of CCPs developed at our institution. Developing consensus learning objectives for a CCP is involved and time-intensive. National efforts, such as the renewed focus by the AAMC [1] and ACGME [2] on clinical skill competencies, can help to refine and share competency goals between institutions. The algorithms marking relevant documents may not perform as well for other CCPs. Finally, the study of concept-level recall and precision was performed on a focused set of 1463 key clinical concepts relevant to the 10 CCPs for a single student; generalized concept identification may yield different results based on different writing styles or note templates used.

6. Conclusion

Medical students and housestaff physicians document a wide exposure to important clinical concepts in the notes they generate through routine clinical work. Current tracking systems do not take advantage of these notes to capture and display relevant content that learners have covered. We present an automated algorithm to align a trainee’s clinical notes to core clinical problems using ranking algorithms employing a UMLS-based concept identification system and note section location. Capturing experience is the first step toward competency-based assessment. Future work will complement written documentation of clinical skills with observed performance evaluations of clinical skills by teachers who provide their assessment of learner’s live or simulated patient workups. Because the current system accurately identified section header location and the biomedical concepts within each section, this tool may also be useful for other clinical and biomedical research applications.

Acknowledgments

This work was supported by the National Library of Medicine Grant T15 LM007450, the D.W. Reynolds Foundation, and the AAMC Clinical Transaction Project.

References

1. Corbett, EC., Whitcomb, M. The AAMC project on the clinical education of medical students: clinical skills education. Washington, DC: Association of American Medical Colleges; 2004.
2. Goroll AH, Sirio C, Duffy FD, LeBlond RF, Alguire P, Blackwell TA, et al. A new model for accreditation of residency programs in internal medicine. Ann Intern Med. 2004; 140(11):902–9. [PubMed: 15172905]
3. Langdorf MI, Montague BJ, Bearie B, Sobel CS. Quantification of procedures and resuscitations in an emergency medicine residency. J Emerg Med. 1998; 16(1):121–7. [PubMed: 9472773]
4. Rattner SL, Louis DZ, Rabinowitz C, Gottlieb JE, Nasca TJ, Markham FW, et al. Documenting and comparing medical students' clinical experiences. JAMA. 2001; 286(9):1035–40. [PubMed: 11559287]
5. Alderson TS, Oswald NT. Clinical experience of medical students in primary care: use of an electronic log in monitoring experience and in guiding education in the Cambridge Community Based Clinical Course. Med Educ. 1999; 33(6):429–33. [PubMed: 10354319]
6. Bird SB, Zarum RS, Renzi FP. Emergency medicine resident patient care documentation using a hand-held computerized device. Acad Emerg Med. 2001; 8(12):1200–3. [PubMed: 11733302]
7. Bardes CL, Wenderoth S, Lemise R, Ortanez P, Storey-Johnson C. Specifying student-patient encounters, web-based case logs, and meeting standards of the liaison committee on medical education. Acad Med. 2005; 80(12):1127–32. [PubMed: 16306286]
8. Mattana J, Charitou M, Mills L, Baskin C, Steinberg H, Tu C, et al. Personal digital assistants: a review of their application in graduate medical education. Am J Med Qual. 2005; 20(5):262–7. [PubMed: 16221834]
9. Carney PA, Pipas CF, Eliassen MS, Mengshol SC, Fall LH, Schifferdecker KE, et al. An analysis of students' clinical experiences in an integrated primary care clerkship. Acad Med. 2002; 77(7):681–7. [PubMed: 12114140]
10. Sequist TD, Singh S, Pereira AG, Rusinak D, Pearson SD. Use of an electronic medical record to profile the continuity clinic experiences of primary care residents. Acad Med. 2005; 80(4):390–4. [PubMed: 15793025]
11. Spickard A 3rd, Gigante J, Stein G, Denny JC. A randomized study of feedback on student write-ups using an electronic portfolio. J Gen Int Med. 2008; 23(7):979–84.
12. Chapman WW, Cooper GF, Hanbury P, Chapman BE, Harrison LH, Wagner MM. Creating a text classifier to detect radiology reports describing mediastinal findings associated with inhalational anthrax and other disorders. J Am Med Inform Assoc. 2003; 10(5):494–503. [PubMed: 12807805]
13. Denny JC, Smithers JD, Miller RA, Spickard A 3rd. "Understanding" medical school curriculum content using KnowledgeMap. J Am Med Inform Assoc. 2003; 10(4):351–62. [PubMed: 12668688]
14. Friedman C, Shagina L, Lussier Y, Hripcsak G. Automated encoding of clinical documents based on natural language processing. J Am Med Inform Assoc. 2004; 11(5):392–402. [PubMed: 15187068]
15. Hazlehurst B, Frost HR, Sittig DF, Stevens VJ. MediClass: a system for detecting and classifying encounter-based clinical events in any electronic medical record. J Am Med Inform Assoc. 2005; 12(5):517–29. [PubMed: 15905485]
16. Brown SH, Speroff T, Fielstein EM, Bauer BA, Wahner-Roedler DL, Greevy R, et al. EQuality: electronic quality assessment from narrative clinical reports. Mayo Clin Proc. 2006; 81(11):1472–81. [PubMed: 17120403]
17. Denny JC, Miller RA, Waitman LR, Arrieta MA, Peterson JF. Identifying QT prolongation from ECG impressions using a general-purpose Natural Language Processor. Int J Med Inform. 2009; 78(1):S34–42. [PubMed: 18938105]
18. Denny, JC., Spickard, A., 3rd, Miller, RA., Schildcrout, J., Darbar, D., Rosenbloom, ST., et al. Identifying UMLS concepts from ECG Impressions using KnowledgeMap. AMIA Annu Symp Proc; 2005; p. 196-200.
19. Aronson, AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proc/AMIA Ann Symp; 2001; p. 17-21.

20. Zeng QT, Goryachev S, Weiss S, Sordo M, Murphy SN, Lazarus R. Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system. BMC Med Inform Decision Making. 2006; 6:30.
21. Denny, JC., Miller, RA., Johnson, KB., Spickard, A. Development and evaluation of a clinical note section header terminology. AMIA Annu Symp Proc; 2008; p. 156-60.
22. Denny JC, Smithers JD, Armstrong B, Spickard A 3rd. "Where do we teach what?" Finding broad concepts in the medical school curriculum. J Gen Intern Med. 2005; 20(10):943–6. [PubMed: 16191143]
23. Cross-validation [cited 2008 July 3]. Available from: http://en.wikipedia.org/wiki/Cross-validation.
24. Meystre S, Haug PJ. Automation of a problem list using natural language processing. BMC Med Inform Decision Making. 2005; 5:30.
25. Elkin PL, Brown SH, Bauer BA, Husser CS, Carruth W, Bergstrom LR, et al. A controlled trial of automated classification of negation from clinical notes. BMC Med Inform Decision Making. 2005; 5:13.
26. Huang Y, Lowe HJ. A novel hybrid approach to automated negation detection in clinical radiology reports. J Am Med Inform Assoc. 2007; 14(3):304–11. [PubMed: 17329723]
27. Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. A simple algorithm for identifying negated findings and diseases in discharge summaries. J Biomed Inform. 2001; 34(5): 301–10. [PubMed: 12123149]
28. Kogan JR, Shea JA. Psychometric characteristics of a write-up assessment form in a medicine core clerkship. Teach Learn Med. 2005; 17(2):101–6. [PubMed: 15833718]
29. Nadkarni P, Chen R, Brandt C. UMLS concept indexing for production databases: a feasibility study. J Am Med Inform Assoc. 2001; 8(1):80–91. [PubMed: 11141514]
30. Meystre S, Haug PJ. Natural language processing to extract medical problems from electronic clinical documents: performance evaluation. J Biomed Inform. 2006; 39(6):589–99. [PubMed: 16359928]
31. Chapman WW, Chu D, Dowling JN. ConText an algorithm for identifying contextual features from clinical text. BioNLP 2007 Biol Trans Clin Lang Process. 2007:81–8.

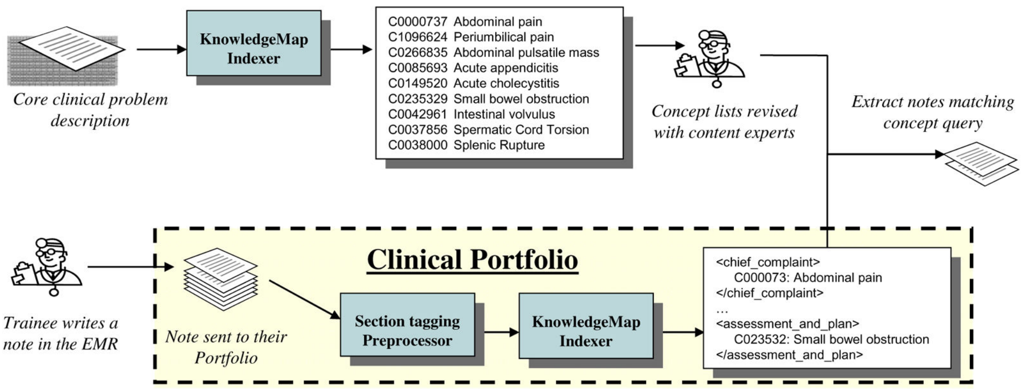
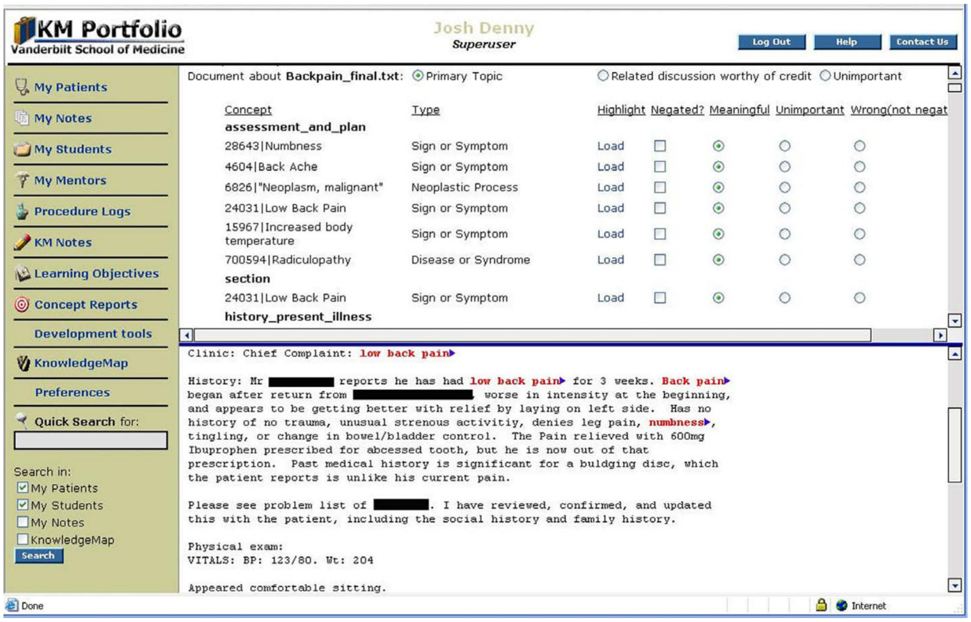

Core clinical problem description
KnowledgeMap Indexer
C0000737 Abdominal pain
C1096624 Periumbilical pain
C0266835 Abdominal pulsatile mass
C0085693 Acute appendicitis
C0149520 Acute cholecystitis
C0235329 Small bowel obstruction
C0042961 Intestinal volvulus
C0037856 Spermatic Cord Torsion
C0038000 Splenic Rupture
Concept lists revised with content experts
Extract notes matching concept query
Clinical Portfolio
Trainee writes a note in the EMR
Note sent to their Portfolio
Section tagging Preprocessor
KnowledgeMap Indexer
<chief_complaint>
C000073: Abdominal pain
</chief_complaint>
…
<assessment_and_plan>
C023532: Small bowel obstruction
</assessment_and_plan>

Fig. 1.

Design of Learning Portfolio system and algorithm to find core clinical problems.



Document about **Backpain_final.txt**: ● Primary Topic ○ Related discussion worthy of credit ○ Unimportant

Concept	Type	Highlight	Negated?	Meaningful	Unimportant	Wrong(not negat
assessment_and_plan						
28643|Numbness	Sign or Symptom	Load	□	●	○	○
4604|Back Ache	Sign or Symptom	Load	□	●	○	○
6826|"Neoplasm, malignant"	Neoplastic Process	Load	□	●	○	○
24031|Low Back Pain	Sign or Symptom	Load	□	●	○	○
15967|Increased body temperature	Sign or Symptom	Load	□	●	○	○
700594|Radiculopathy	Disease or Syndrome	Load	□	●	○	○
section						
24031|Low Back Pain	Sign or Symptom	Load	□	●	○	○
history_present_illness						

```
Clinic: Chief Complaint: low back pain▸

History: Mr [redacted] reports he has had low back pain▸ for 3 weeks. Back pain▸
began after return from [redacted], worse in intensity at the beginning,
and appears to be getting better with relief by laying on left side.  Has no
history of no trauma, unusual strenous activitiy, denies leg pain, numbness▸,
tingling, or change in bowel/bladder control.  The Pain relieved with 600mg
Ibuprophen prescribed for abcessed tooth, but he is now out of that
prescription.  Past medical history is significant for a buldging disc, which
the patient reports is unlike his current pain.

Please see problem list of [redacted]. I have reviewed, confirmed, and updated
this with the patient, including the social history and family history.

Physical exam:
VITALS: BP: 123/80. Wt: 204

Appeared comfortable sitting.
```

Fig. 2.

Screenshot of scoring interface. The scoring interface shows concepts matching the query, the UMLS semantic type, and the document section (e.g., "assessment and plan") in which the concept was located. The section label "section" refers to a concept found outside a known section, as identified by the section tagger.

Table 1

Count of clinical notes discussing each core clinical problem and the number of relevant concepts in each group of notes. Primary notes were those judged as containing a primary reference to the core clinical problem (CCP). Relevant notes contained related discussions to the CCP. The average concepts/note refers to the average number of pertinent concepts to each CCP found in a document.

	Core clinical problems										**Total**
	Abdominal pain	**Back pain**	**Chest pain**	**Cough**	**Dezression**	**Dysuria**	**Fever**	**GI bleeding**	**Jaundice**	**Rash**	
Primary notes count											
Student 1	8	7	6	15	7	3	41	3	2	4	96
Student 2	6	3	13	11	6	9	26	6	2	11	93
Student 3	4	3	8	8	2	6	14	4	6	9	64
Average concepts/note	9.5	5.4	24.6	9.9	4.9	5.1	11.3	15.1	8.0	5.7	9.9
Relevant notes count											
Student 1	8	23	15	12	11	1	31	28	3	6	138
Student 2	16	1	13	14	4	22	19	10	9	11	119
Student 3	13	1	5	3	5	4	13	6	4	4	58
Average concepts/note	6.5	2.5	10.1	5.0	4.1	2.8	6.5	8.4	3.8	3.9	5.4

Table 2

Ranking method accuracy as measured by area under receiver operator characteristic curves (AUC). The “topic search only” restricts the search to the concept corresponding to the core problem only (e.g., “abdominal pain” or “back pain”). “Asserted” refers to concepts that are instantiated (i.e., not negated) in the text (e.g., “she has chest pain”). *p*-Values compare each ranking scheme to the topic search only. The best performing algorithms are bolded.

	Abdominal pain	Back pain	Chest pain	Cough	Depression	Dysuria	Fever	GI bleeding	Jaundice	Rash	Mean (95% CI)	*p*
Primary documents												
Topic concept only	0.72	0.64	0.90	0.87	0.87	0.81	0.52	0.50	0.81	0.65	0.73 (0.64–0.82)	
Total matching concepts	0.87	0.92	0.91	0.87	0.90	0.97	0.80	0.88	0.98	0.90	0.90 (0.87–0.93)	0.002
Total matching asserted concepts	0.90	0.94	0.95	0.94	0.92	0.98	0.88	0.94	0.98	0.95	**0.94 (0.92–0.96)**	0.001
Matching correct concepts	0.87	0.91	0.90	0.87	0.92	0.97	0.80	0.88	0.99	0.91	0.90 (0.87–0.94)	0.002
Matching correct asserted concepts	0.91	0.92	0.95	0.94	0.94	0.98	0.88	0.95	0.99	0.96	**0.94 (0.92–0.96)**	0.001
Matching meaningful correct concepts	0.87	0.93	0.90	0.87	0.93	0.97	0.83	0.90	0.99	0.91	0.91 (0.88–0.94)	0.002
Expert weighted scheme	0.91	0.95	0.95	0.94	0.92	0.98	0.88	0.94	0.98	0.95	**0.94 (0.92–0.96)**	0.001
Logistic Regression	0.93	0.89	0.98	0.91	0.91	0.92	0.89	0.93	0.92	0.84	0.91 (0.89–0.94)	0.002
TF–IDF weighting	0.88	0.93	0.92	0.88	0.91	0.97	0.81	0.89	0.98	0.92	0.91 (0.88–0.94)	0.002
TF–IDF, asserted concepts	0.91	0.95	0.96	0.93	0.91	0.98	0.87	0.93	0.98	0.92	**0.93 (0.91–0.95)**	0.001
All relevant documents												
Topic concept only	0.69	0.69	0.83	0.82	0.80	0.77	0.51	0.50	0.80	0.59	0.70 (0.62–0.78)	
Topic concept only	0.85	0.90	0.88	0.85	0.89	0.94	0.80	0.87	0.95	0.87	0.88 (0.86–0.91)	0.001
Total matching concepts	0.87	0.85	0.91	0.94	0.92	0.96	0.84	0.92	0.97	0.93	**0.91 (0.89–0.94)**	0.000
Total matching asserted concepts	0.85	0.90	0.89	0.85	0.91	0.94	0.81	0.87	0.97	0.89	0.89 (0.86–0.92)	0.001
Matching correct concepts	0.87	0.83	0.92	0.94	0.94	0.96	0.85	0.92	0.98	0.94	**0.92 (0.89–0.95)**	0.000
Matching correct asserted concepts	0.85	0.95	0.89	0.85	0.92	0.94	0.85	0.91	0.97	0.88	0.90 (0.87–0.93)	0.001
Matching meaningful correct concepts	0.85	0.90	0.94	0.91	0.84	0.99	0.89	0.93	0.96	0.90	**0.91 (0.88–0.94)**	0.001
Expert weighted scheme	0.88	0.86	0.91	0.94	0.91	0.96	0.84	0.92	0.97	0.93	**0.91 (0.89–0.94)**	0.000
TF–IDF weighting	0.86	0.92	0.90	0.86	0.89	0.96	0.82	0.88	0.96	0.91	0.90 (0.87–0.93)	0.000
TF–IDF, asserted concepts	0.86	0.95	0.92	0.93	0.90	0.99	0.86	0.91	0.96	0.91	**0.92 (0.89–0.94)**	0.000

TF–IDF, term frequency, inverse document frequency weighting algorithm.

Table 3

Impact of individual note sections on overall relevance of the document. OR, odds ratio, calculated by multivariable ordinate logistic regression. CCP, core clinical problems. The document count represents the total documents matched for all 10 core clinical problems, such that a single document could be counted up to 10 times.

Section	Document count for all CCPs			Adjusted OR	*p*
	Primary	Relevant	Irrelevant		
Chief complaint	89	33	60	3.42 (2.51–4.67)	0.000
History of present illness	186	181	470	2.00 (1.77–2.25)	0.000
Past medical history	66	52	181	0.86 (0.70–1.07)	0.176
Personal and social history	15	22	113	1.28 (0.73–2.22)	0.387
Family medical history	49	41	140	0.91 (0.61–1.36)	0.649
Medications	15	8	46	0.83 (0.48–1.43)	0.497
Review of systems	92	127	594	0.72 (0.65–0.80)	0.000
Physical examination	170	208	897	1.61 (1.41–1.85)	0.000
Laboratory data	67	39	83	0.95 (0.75–1.20)	0.668
Problem lists	21	14	21	1.52 (1.04–2.24)	0.033
Assessment and plan	216	200	469	1.93 (1.75–2.13)	0.000
Total unique documents	986	925	3074		

Table 4

Failure analysis of recall errors in concept identification.

Error type	Total errors	Unique errors	Examples and comment
Typographical errors	36 (18%)	31 (29%)	Misspellings such as: “Diarrhea,” “2Allergies,” “pleracy”, “emisis”
Grammatical parsing errors	17 (9%)	10 (9%)	“He fever,” “coughing fits,” “progressed to cough” (cough identified as a verb and ignored)
Line break[a]	11 (6%)	11 (10%)	“discussed blood \n pressure,” “shortness of \n breath”
Form data	80 (40%)	2 (2%)	“Insomnia Mood changes Emotional lability Anxiety Depression” → “mood insomnia” instead of “insomnia” and “mood changes”
Complex phrase parsing	24 (12%)	23 (21%)	“back and lower leg pain” → KMCI found “Back” instead of “back pain”; “pain upon deep breathing” → KMCI found “pain” and “depth of inspiration” instead of “chest pain on breathing”
Disambiguation	12 (6%)	12 (11%)	“UA” → KMCI sometimes identified as “upper arm” instead of “urinalysis”; “IBS” → KMCI sometimes identified as “Ib serotype” instead of “Irritable Bowel Syndrome”
Under specified terms	8 (4%)	8 (8%)	“… not had **reflux** with omeprazole…” → KMCI identified “Reflux, NOS” instead of “Gastroesophageal reflux disease”
Inadequate synonymy	10 (5%)	10 (9%)	“heart cath” (UMLS did not contained the abbreviation “cath”), “viral resp infection” (UMLS did not contained the abbreviation “resp”)
Total	198	109	

KMCI, KnowledgeMap concept identifier.

[a] Most documents in this corpus are word-wrapped with newline characters (identified by “\n” in the examples). The algorithm attempts to recombine sentences with line breaks inserted within them but occasionally failed to do so, and the algorithm does not allow concepts to span multiple sentences.

Table 5

Selected diseases relevant to core clinical problems. Only correct concept matches are included in this list. Diseases located only in the “review of systems” section are excluded from this list.

Disease name	Student 1 (asserted)	Student 1 (negated)	Student 2 (asserted)	Student 2 (negated)	Student 3 (asserted)	Student 3 (negated)	Total asserted	Total negated
Hypertensive disease	38	0	89	3	129	2	256	5
Pneumonia	46	1	45	3	42	2	133	6
Gastroesophageal reflux	33	3	32	0	67	3	132	6
Myocardial infarction	15	1	53	1	56	0	124	2
Chronic obstructive lung disease	21	0	10	0	80	0	111	0
Diabetes mellitus	5	0	42	2	36	0	83	2
Urinary tract infections	9	0	51	4	17	4	77	8
Congestive heart failure	10	0	26	0	16	0	52	0
Mental depression	13	0	14	0	16	0	43	0
Bronchial asthma	2	1	29	0	11	0	42	1
Pulmonary embolism	4	1	5	2	28	12	37	15
Cerebrovascular accident	5	1	18	0	9	0	32	1
Influenza	6	0	4	0	21	1	31	1
Hepatitis	2	0	12	0	16	0	30	0
Hypothyroidism	5	0	4	0	20	0	29	0
Herpes zoster disease	4	0	12	0	12	0	28	0
Peptic ulcer disease	9	0	9	0	8	4	26	4
Anemia	7	1	11	0	8	0	26	1
Diverticulitis	12	0	5	1	6	0	23	1
Upper respiratory infections	11	2	7	3	5	1	23	6
Pneumocystis carinii pneumonia	0	0	9	0	13	0	22	0
Pyelonephritis	0	0	14	1	3	0	17	1

Table 6

Selected findings relevant to core clinical problems. Findings located only in the “review of systems” section are excluded from this list.

Disease name	Student 1 (asserted)	Student 1 (negated)	Student 2 (asserted)	Student 2 (negated)	Student 3 (asserted)	Student 3 (negated)	Total asserted	Total negated
Fever	47	41	79	29	51	23	177	93
Skin rash	34	141	52	134	52	74	138	349
Vomiting	11	6	72	19	20	17	103	42
Chest pain	9	14	45	2	45	6	99	22
Cough	18	11	27	5	32	5	77	21
Edema	11	43	35	74	21	36	67	153
Constipation	10	9	21	16	26	7	57	32
Leukocytosis	6	0	23	0	23	1	52	1
Diarrhea	6	4	22	15	21	5	49	24
Shortness of breath	7	3	24	2	15	5	46	10
Back pain	21	5	16	4	6	0	43	9
Dysuria	13	7	22	28	5	5	40	40
Abdominal pain	6	2	17	3	12	4	35	9
Wheezing	8	39	10	12	15	2	33	53
Pruritis	6	0	16	5	11	2	33	7
Dyspnea	11	15	9	1	12	2	32	18
Erythema	6	2	16	20	7	2	29	24
Syncope	4	3	14	8	9	1	27	12
Cardiac murmurs	0	0	15	67	9	22	24	89
Jaundice	2	10	6	80	12	45	20	135
Crackles	0	40	15	2	5	3	20	45
Productive cough	3	2	6	0	10	0	19	2
Tachycardia	1	0	8	0	8	0	17	0
Hematuria	2	2	3	10	7	1	12	13
Hematochezia	3	5	4	1	5	1	12	7
Nuchal rigidity	1	7	2	26	4	14	7	47
Hepatomegaly	0	6	0	30	2	32	2	68