Abstract

Objective: Clinical notes, typically written in natural language, often contain substructure that divides them into sections, such as “History of Present Illness” or “Family Medical History.” The authors designed and evaluated an algorithm (“SecTag”) to identify both labeled and unlabeled (implied) note section headers in “history and physical examination” documents (“H&P notes”).

Design: The SecTag algorithm uses a combination of natural language processing techniques, word variant recognition with spelling correction, terminology-based rules, and naive Bayesian scoring methods to identify note section headers. Eleven physicians evaluated SecTag’s performance on 319 randomly chosen H&P notes.

Measurements: The primary outcomes were the algorithm’s recall and precision in identifying all document sections and a predefined list of twenty-nine major sections. A secondary outcome was to evaluate the algorithm’s ability to recognize the correct start and end boundaries of identified sections.

Results: The SecTag algorithm identified 16,036 total sections and 7,858 major sections. Physician evaluators classified 15,329 as true positives and identified 160 sections omitted by SecTag. The recall and precision of the SecTag algorithm were 99.0 and 95.6% for all sections, 98.6 and 96.2% for major sections, and 96.6 and 86.8% for unlabeled sections. The algorithm determined the correct starting and ending text boundaries for 94.8% of labeled sections and 85.9% of unlabeled sections.

Conclusions: The SecTag algorithm accurately identified both labeled and unlabeled sections in history and physical documents. This type of algorithm may assist in natural language processing applications, such as clinical decision support systems or competency assessment for medical trainees.

Introduction

Electronic health records comprise a rich source of clinical information, including observations about patients, laboratory and imaging reports, diagnoses, therapies, and longitudinal descriptions of patients’ illnesses over time. Most clinical records exist as natural language text narratives that providers compose as they interact with patients and interpret test results. As healthcare providers write, dictate, or electronically enter clinical documents, they typically use conceptual, physical, or electronic templates to divide their narratives into commonly-recognized segments, or sections. Clinicians often label the segments with frequently used but nonstandardized terms (“section headers”). For example, history and physical examination (H&P) notes generally contain sections labeled “history of present illness”, “past medical history”, and “physical examination.” Sections can have subsections, such as “cardiovascular exam” within “physical examination” or “substance abuse history” within “social history”. We describe development and evaluation of an algorithm, called SecTag, to parse natural-language clinical documents and label these sections. The SecTag algorithm uses an empirically derived document section header terminology. It identifies both explicitly labeled and unlabeled (i.e., implied) sections of clinical narratives.

Background

Motivation for Automated Section Header Identification

Review of seven decades of physical diagnosis textbooks suggests that common section headers found in clinical
notes have minimally changed over the last century.2–9 In medical, nursing, and dental schools most students learn to write clinical notes using these standard section headers.10–16 Many structured note capture tools use this common organizational structure.17–21 Within given note types, clinical note sections generally follow a logical sequence, such as the head-to-toe anatomical ordering of physical examination components.9 Clinical narrative sections also figure into physicians’ compensation for clinical encounters, based on the United States. Evaluation and Management

figure into physicians’ compensation for clinical encounters, based on the United States. Evaluation and Management

identification of sections within clinical documents provides important context for recognizing and understanding the biomedical concepts they contain. For example, the term “friction rub” (a clinical finding that occurs in multiple anatomical loci) has more specific meaning and clinical implications when one knows if it appeared in the “pulmonary auscultation”, “cardiac auscultation”, “abdominal examination” or “joint examination” section of a note. Similarly, “past medical history” and “family medical history” sections list names of diseases that, based on context, have different implications regarding who has the disease. Clinical note section tagging enhances the functionality of the following applications: (a) clinical decision support systems that require background information to provide appropriate advice (e.g., does the patient have a history of heart failure?),23,24 (b) automated problem list generators,25 (c) systems that support healthcare professionals’ education (e.g., has a trainee ever evaluated a patient with pneumonia or reported hearing a diastolic murmur?),26 (d) clinical documentation tools that convert natural language descriptions into structured representations from a target terminology, such as SNOMED CT®,17,18,20 and (e) efforts to mine electronic medical records for evidence of completion of medical, nursing, and dental schools most students learn to write clinical notes using these standard section headers.10–16 Many structured note capture tools use this common organizational structure.17–21 Within given note types, clinical note sections generally follow a logical sequence, such as the head-to-toe anatomical ordering of physical examination components.9 Clinical narrative sections also figure into physicians’ compensation for clinical encounters, based on the United States. Evaluation and Management

Journal of the American Medical Informatics Association Volume 16 Number 6 November / December 2009 807

scharge summaries,26,29 radiograph reports,30 mammograms,31 and pathology reports.32 It includes a document preprocessor that recognizes some common clinical section tags (e.g., “history of present illness”),33 though no formal evaluation of the segmenter’s accuracy has been published. Several clinically-oriented NLP projects, including, among others, SAPHIRE,34 MetaMap,35 the KnowledgeMap Concept Identifier (“KnowledgeMap”),36,37 a system developed by Nadkarni and colleagues,38 the Mayo Vocabulary Processor,39,40 and systems developed by Chapman and colleagues,41–43 took unique approaches to identifying UMLS (or other standard vocabulary) concepts within clinical texts. While many of these systems index clinical documents at the sentence or noun-phrase level, they often do not explicitly segment clinical notes by section. Thus, such systems may not distinguish the different meanings of “congestive heart failure” when the phrase appears in the family history, past medical history, or the assessment sections of clinical notes. However, some systems correctly classify concepts when a sentence within a note contains clues such as, “there is a family history of colon cancer.”

Meystre and Haug created a NLP system to generate problem lists by processing clinical documents.25,33 Their first step used a document parser that matched document strings to a list of known headers to identify common sections within clinical documents. Their system assigned all text from the beginning of one section header to the start of the next recognized section header to the former section header. While exemplary, their algorithm misclassified unknown or unlabeled sections.

This paper describes a general algorithm, SecTag, for identifying section headers and delimiting the content associated with those sections. Authors evaluated SecTag’s performance on a corpus of H&P notes. The algorithm can serve as a preprocessor for other NLP applications.

Methods

Methods—SecTag System Overview

Using a locally-developed section header terminology, SecTag identifies section tags within natural language clinical documents.1 The tagger identifies explicitly labeled section headers and deduces presence of implicitly labeled (i.e., missing) section headers. The SecTag algorithm produces output in two formats, native XML with identified section headers annotated within the note’s original text, or as an HL7 CDA-compliant XML document with identified section labels (see online Appendix 1, available at http://www.jamia.org, for hypothetical notes with corresponding SecTag CDA output). The Vanderbilt University Medical Center Institutional Review Board approved the SecTag development process and evaluation.

Methods—H&P Corpus for SecTag Development and Evaluation

To develop the SecTag algorithm and the terminology, we divided a corpus of deidentified electronic H&P documents into a training set and an evaluation set.1 Vanderbilt’s EMR system captures all inpatient and outpatient notes since 2001, and a large portion of notes since the early 1990s. To provide H&P notes for this study, an automated program randomly selected 25,000 notes within the Vanderbilt EMR
system with titles containing "H&P", "admission", or "history." These notes were deidentified by removing the 18 HIPAA safe-harbor categories of information using DE-ID®, a commercially-available software package (University of Pittsburgh Medical Center, Pittsburgh, PA), and other pre- and post-processing refinements.44 Healthcare providers had originally created the H&P notes by manual entry (typing them into the EMR freehand or via a template) or through dictating notes that were then transcribed into the EMR system. We manually reviewed this set of 25,000 notes and retained only those that were actual H&Ps (n/H11005 10,767).

This set was divided into a training set (n/H11005 9,567) used for development of the terminology and SecTag, and an evaluation set (n/H11005 1,200, of which 540 were actually used in the study).

Methods—Section Header Terminology
We previously described development of the section header terminology, which included all relevant section tags for H&P documents.1 Terminology development derived candidate header terms from LOINC®,45 the QMR findings hierarchy,46–48 history and physical examination textbooks,2–8 the section header list identified by Meystre and Haug,33 extensive automated and manual review of the H&P from our training set, and from review of general and subspecialty H&P templates (n/H11005 83) that exist within Vanderbilt’s web-based clinical note writing application.17 The SecTag terminology data model followed principles established by Cimino49 and others.50 Similar to the UMLS,51 the SecTag header terminology included a concept-oriented structure with polyheirarchical parent–child relationships (e.g., “jugular venous pulse exam” was a child of both the “cardiovascular exam” and the “neck exam”). The SecTag section header terminology contained 1,109 concepts with 4,332 terms. It contains mapping to existing LOINC® and the E/M coding system terms (where possible). The SecTag terminology is available free-of-charge by contacting author JD.

Methods—SecTag Algorithm
The SecTag algorithm sequentially processes documents in five steps (per Figure 1): (a) identify sentence boundaries and elements of lists (e.g., “1. Congestive heart failure”); (b) identify all candidate section headers using lexical tools, spelling correction, and NLP techniques; (c) calculate the Bayesian probabilities that each sentence belongs to any given section; (d) disambiguate unclear section headers, using the Bayesian probabilities; and (e) identify the end (terminal boundary) of each section. We used the training set of H&P notes to develop and iteratively improve the algorithm. The SecTag algorithm is described briefly here; a more detailed description is provided in Appendix 2, available online at http://www.jamia.org.

To disambiguate unclear section headers and to predict where unlabeled sections occur, SecTag employs both statistical and hierarchical models of section headers. The statistical model, derived from automated processing of the training set, contains the prior probability of each possible section header occurring in a document, the probability of each section header occurring in each location in the document, and the probabilistic sequential order for section headers appearing in documents (e.g., “chief complaint” occurs before “history of present illness”). The SecTag algorithm uses the header terminology’s parent–child hierarchical relationships (e.g., “substance abuse history” is part of “social history”) for rule-based inferences (e.g., the label “mother” likely represents “mother’s medical history” if found within the “family medical history” section) and to calculate the path length between two section concepts (i.e., the number of section concepts [nodes] traversed between two section headers in the terminology).
Identifying Sentences and Lists
Clinical notes are often not well-formatted or well-structured, especially when created by busy practitioners through dictation without a template. To “normalize” clinical documents, SecTag first identifies individual sentences and lists within the document. Sample H&P notes used in this study were often word-wrapped through the insertion of newline characters, so SecTag uses a variant of the KnowledgeMap sentence identifier, to combine sentence fragments separated by newline characters.36 To aid in section boundary prediction, SecTag identifies paragraph boundaries and enumerated lists of words, phrases, or sentences (e.g., “1. Congestive heart failure . . . . . . 2. Diabetes . . . .”). A group of sentences within a list likely belongs to the same section or subsection (e.g., if list X was immediately preceded by “Assessment and Plan”, all its components are likely “assessment_and_plan” elements).

Identifying Candidate Section Headers
The SecTag algorithm next processes the document, sentence by sentence, to find all possible section headers. First, SecTag attempts to identify all explicitly labeled section tags by searching for strings that begin sentences, consist of only capital letters, or end in a dash, colon, or period (because manual review found these delimiters common in section headers in both dictated and electronically generated notes). The system also identifies within-sentence strings that match section header terms, such as “RR: 12, Temp: 37”, which match the “respiratory rate” and “temperature” subheaders of the “vital signs” within “physical examination”. To further improve section tag recognition, the system employs three techniques to identify document header strings that do not exactly correspond to entries in the SecTag header terminology: (1) recognize derivational word variants and synonyms that map to known terms;36 (2) correct spelling errors using the open-source spell checker Aspell;52,53 and (3) remove common “stop words” (such as prepositions and determiners) and certain modifiers such as possessive words, numbers (written as either a word or a number), anatomical references (e.g., “right”, “superior”, or “bilateral”), and other common adjectives (e.g., “recent”, “other”). Each candidate section header within the SecTag lexicon receives a score based on its similarity to each phrase in the clinical document; strings with direct matches receive higher scores than those matched via transformations, such as spelling correction or word variant generation.

The SecTag algorithm also uses syntactic and semantic methods (based on the SecTag terminology) to increase recognition of implied section headers. It applies NLP algorithms to detect section tags that occur within a sentence by examining noun phrases, such as “chief complaint” from “Mr. X is here for a chief complaint of dyspnea.” Using the section relationships present in the section terminology, SecTag can find common ancestral section headers when multiple valid section header candidates exist (Figure 2).

When processing a document, SecTag keeps track of all instantiated (i.e., already matched) sections and whether a given section is “open” (or “active”—meaning that new sentences are still being assigned to it). Sections can be nested such that multiple sections are open simultaneously, e.g., both “physical exam” and “cardiovascular exam” can be open.

Bayesian Section Tag Prediction
The SecTag algorithm uses naive Bayes scoring to predict unlabeled (implied) section headers, to choose among multiple possible section header candidates for a given text segment, and to discard candidate section headers with low scores. The algorithm calculates Bayesian probability, sentence-by-sentence, for all of a document’s section header candidates Section, based on probabilities of words occurring in each section in the training set \( P(\text{Section} | \text{words}) \). The training set also determined the probability of any section following a previously encountered exact-matched section in the document \( P(\text{Section} | \text{priorSections}) \), where priorSections represents the preceding exact-matched sections in the document. The complete SecTag calculation is

\[
P(\text{Section}_i | \text{words}, \text{priorSection}_j) = P(\text{Section}_i) \times \prod_{k=1}^{m} P(\text{Section}_i | \text{priorSection}_k) \times \prod_{j=1}^{\text{words}} \frac{P(\text{word}_j | \text{Section}_i)}{P(\text{word}_j)}
\]

Training SecTag did not require manual curation—see Appendix 2 for details.

Disambiguating Unclear Section Headers
Many H&P strings potentially map to multiple section header concepts; for example, “cardiovascular” can refer to “cardiovascular_exam”, “cardiovascular_plan”, “cardiovascular_system_review”, etc. Other section headers may be implied, such as when “2/6 holosystolic murmur” connotes “cardiac examination” within “physical examination.” The SecTag algorithm calculates a score for all ambiguous sections labels and for all unlabeled sections predicted by NLP or semantic rules. The algorithm uses this score to either select the best scoring section candidate or discard the candidate as a “poor match” (Figure 1). The score for each section involves three major components: (1) the similarity of the clinical note phrase to the candidate SecTag term string, (2) the Bayesian score (calculated as above), and (3) the distance of the candidate section header to other nearby note section headers, calculated as the path length between the concepts in the SecTag terminology. The SecTag algorithm then instantiates the best-ranking section header as “active” (unless it is considered a poor match due to a low Bayesian score or is eliminated through a series of rules—Appendix 2, Section D). For
example, one rule excludes any predicted section candidate that occurs elsewhere in the document as an exact-matched header. Appendix 2 describes the disambiguation methods in detail.

Identifying Section Termination

As SecTag sequentially processes the document, it assigns each encountered sentence to all currently active sections. To determine when to terminate (close) an active section, SecTag employs Bayesian scores and a series of rules. Termination rules are specific to the type of section currently active; for example, short sections such as "chief complaint" are terminated when SecTag encounters a blank line between paragraphs, while longer sections like "history of present illness" can cross paragraph boundaries. The algorithm uses "knowledge" of document formatting to keep an active section open or to close it (e.g., it will not close in the middle of a numbered list, but will do so at its end if subsequent markings suggest a new section header). If a newly encountered word or phrase has a sufficiently high Bayesian score to indicate that it is a new section header that is not a child of any active section header, SecTag may close some or all the open headers. Conversely, to prevent assignment of a sentence to a poorly matching section, an active section header may be dropped from consideration (terminated) if its Bayesian score falls to a sufficiently low-level. Thus, some sentences may not be assigned to any section if the best ranking section receives a poor score.

**Table 1**  Recall and Precision for Each Major Section

<table>
<thead>
<tr>
<th>Section Name</th>
<th>n</th>
<th>Number Labeled (%)</th>
<th>Recall (95% CI)</th>
<th>Precision (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chief complaint</td>
<td>283</td>
<td>280 (99%)</td>
<td>100% (98–100)</td>
<td>100% (98–100)</td>
</tr>
<tr>
<td>History present illness</td>
<td>353</td>
<td>281 (80%)</td>
<td>99% (98–100)</td>
<td>93% (90–96)</td>
</tr>
<tr>
<td>Past medical history</td>
<td>296</td>
<td>262 (95%)</td>
<td>99% (98–100)</td>
<td>99% (97–100)</td>
</tr>
<tr>
<td>Family medical history</td>
<td>255</td>
<td>250 (98%)</td>
<td>100% (99–100)</td>
<td>98% (95–99)</td>
</tr>
<tr>
<td>Parent medical history</td>
<td>192</td>
<td>7 (4%)</td>
<td>95% (90–0.98)</td>
<td>90% (84–94)</td>
</tr>
<tr>
<td>Sibling medical history</td>
<td>96</td>
<td>3 (8%)</td>
<td>87% (71–95)</td>
<td>97% (84–100)</td>
</tr>
<tr>
<td>Child medical history</td>
<td>3</td>
<td>1 (33%)</td>
<td>100% (29–100)</td>
<td>100% (29–100)</td>
</tr>
<tr>
<td>Health maintenance</td>
<td>92</td>
<td>91 (99%)</td>
<td>100% (96–100)</td>
<td>100% (96–100)</td>
</tr>
<tr>
<td>Personal and social history</td>
<td>267</td>
<td>252 (94%)</td>
<td>100% (98–100)</td>
<td>99% (96–100)</td>
</tr>
<tr>
<td>Substance use history</td>
<td>254</td>
<td>138 (54%)</td>
<td>94% (91–97)</td>
<td>98% (96–100)</td>
</tr>
<tr>
<td>Medications</td>
<td>282</td>
<td>254 (90%)</td>
<td>100% (98–100)</td>
<td>99% (96–100)</td>
</tr>
<tr>
<td>Allergies and adverse reactions</td>
<td>254</td>
<td>249 (98%)</td>
<td>100% (99–100)</td>
<td>100% (99–100)</td>
</tr>
<tr>
<td>Review of systems</td>
<td>462</td>
<td>437 (95%)</td>
<td>100% (98–100)</td>
<td>95% (92–97)</td>
</tr>
<tr>
<td>Physical examination</td>
<td>336</td>
<td>307 (91%)</td>
<td>100% (99–100)</td>
<td>99% (97–100)</td>
</tr>
<tr>
<td>Vital signs</td>
<td>333</td>
<td>201 (60%)</td>
<td>99% (97–100)</td>
<td>92% (89–95)</td>
</tr>
<tr>
<td>General examination</td>
<td>268</td>
<td>211 (79%)</td>
<td>99% (97–100)</td>
<td>100% (98–100)</td>
</tr>
<tr>
<td>Dermatologic examination</td>
<td>216</td>
<td>172 (80%)</td>
<td>99% (97–100)</td>
<td>95% (92–98)</td>
</tr>
<tr>
<td>Lymph nodes/heme examination</td>
<td>142</td>
<td>131 (92%)</td>
<td>99% (96–100)</td>
<td>96% (99–96)</td>
</tr>
<tr>
<td>HEENT examination</td>
<td>767</td>
<td>595 (78%)</td>
<td>98% (97–100)</td>
<td>98% (96–99)</td>
</tr>
<tr>
<td>Cardiovascular examination</td>
<td>293</td>
<td>235 (80%)</td>
<td>100% (98–100)</td>
<td>98% (96–99)</td>
</tr>
<tr>
<td>Gastrointestinal examination</td>
<td>295</td>
<td>225 (76%)</td>
<td>99% (98–100)</td>
<td>97% (94–98)</td>
</tr>
<tr>
<td>Chest examination</td>
<td>374</td>
<td>291 (78%)</td>
<td>99% (97–100)</td>
<td>98% (96–99)</td>
</tr>
<tr>
<td>Genitourinary examination</td>
<td>138</td>
<td>116 (84%)</td>
<td>99% (96–100)</td>
<td>94% (89–97)</td>
</tr>
<tr>
<td>Neurological examination</td>
<td>320</td>
<td>244 (76%)</td>
<td>97% (94–100)</td>
<td>95% (91–97)</td>
</tr>
<tr>
<td>Psychological examination</td>
<td>67</td>
<td>62 (93%)</td>
<td>100% (95–100)</td>
<td>99% (92–100)</td>
</tr>
<tr>
<td>Musculoskeletal examination</td>
<td>82</td>
<td>51 (62%)</td>
<td>95% (87–99)</td>
<td>92% (84–97)</td>
</tr>
<tr>
<td>Extremity examination</td>
<td>314</td>
<td>210 (67%)</td>
<td>97% (95–99)</td>
<td>94% (90–96)</td>
</tr>
<tr>
<td>Laboratory, imaging, and pathology results</td>
<td>393</td>
<td>246 (63%)</td>
<td>98% (96–99)</td>
<td>88% (84–91)</td>
</tr>
<tr>
<td>Analysis, assessment, and plan</td>
<td>600</td>
<td>503 (84%)</td>
<td>98% (96–99)</td>
<td>96% (94–97)</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>7969</td>
<td>6325 (79%)</td>
<td>98.6% (98.3–98.8)</td>
<td>96.2% (95.8–96.6)</td>
</tr>
</tbody>
</table>

"Number labeled" refers to the number of sections that had an author-placed label in the document. HEENT = Head, Eyes, Ears, Nose, Throat.
determined and annotated the accuracy of SecTag-identified section boundaries (whether the algorithm had correctly marked the start and end of the text belonging to each section). Using the web-based tool, the evaluators indicated where section headers (unlabeled major sections or author-labeled sections) should have occurred, but were not detected by the SecTag algorithm. To assist the evaluators with their ratings, the physicians were given a reference list of all possible terminology section headers. This list was organized hierarchically (according to the canonical relationships in the SecTag section header terminology) and visually weighted such that more common sections were in larger fonts (based on their frequency in the training documents).

To improve concordance among evaluators, project members provided a training example, followed by a session in which all physician-evaluators scored the same five initial H&Ps. Their ratings were compared, and we established that adequate agreement existed among evaluators. Next, we configured the evaluation sequence such that every eighth document that each evaluator subsequently reviewed came from a common shared document pool. This allowed calculation of inter-rater agreement.

We instructed the evaluators to examine the clinical note’s original labels (when present) as a guide to the document author’s intended section header when evaluating SecTag’s choice of a section header. For instance, if the document included the section label “Head, Eyes, Ears, Nose, Throat” but that section contained only an ear exam, the evaluators were instructed to mark SecTag’s annotation as correct if it contained either the broader (“HEENT exam”) or more specific (“ear exam”) concept. Likewise, “He has a 40 pack-year history of smoking” could be accurately tagged as the “tobacco use” segment of the “social history” even if the author had placed it in the history of present illness section.

SecTag Evaluation Measurements

The primary evaluation outcomes, based on physician-evaluators’ ratings as the gold standard, were SecTag’s recall and precision for all major sections listed in Table 1 (whether explicitly labeled in the document or implied), and the recall and precision for all sections (whether major or not) that were identified by the system. We classified SecTag’s correctly identified section headers (i.e., where physicians agreed with SecTag) as true positives (TP), and instances where physicians did not agree with SecTag labeling as false positives (FP). Sections labeled by physician-evaluators but not labeled by SecTag comprised “omitted sections (OS).” The study defined the following: (1) OVERALL RECALL: the number of TP divided by the total number of sections (TP + OS); (2) OVERALL PRECISION: the ratio of SecTag’s correctly labeled section count to its total number of proposed section tags [TP/(TP + FP)]. MAJOR SECTION RECALL: [MTP/(MTP + OMS)], where MTP is the number of major section true positives, OMS is the number of omitted major sections, and (MTP + OMS) represents the total number of gold standard major sections in the document; (4) MAJOR SECTION PRECISION: [MTP/(MTP + MFP)], where MFP is the number of major section false positives.

SecTag Evaluation Statistical Analyses and Sample Size Calculation

Prior to the study, we determined that analysis with 90% power at a precision of 0.9 within a 95% confidence interval of 0.05 would require evaluators to label 471 section headers (or about 26 typical documents from our training set). We chose a much larger set to determine accurate confidence intervals for major sections, which might not be present in every document.

We calculated inter-rater agreement via Cohen’s Kappa. We used the Wilcoxon’s rank sum test to compare nonparametric data (expressed as median and interquartile range), the Student’s t test for parametric data (expressed as mean ± standard deviation), and the $\chi^2$ statistic for categorical data. Confidence intervals were calculated using the binomial exact method. All statistical analyses were performed with Stata, version 9.2 (StataCorp LP, College Station, TX).

Results

SecTag Recall and Precision

The physician-evaluators scored 319 unique H&P documents. Of these, 66 were authored by attending physicians and 252 by house staff; only one note was written by a medical student. Reviewers classified 88% of the notes as full H&Ps, 6% as attending attestations, and 6% as brief admission notes. Forty-four percent of the notes were general medicine service H&Ps, 43% were from nonsurgical subspecialties, 10% were surgical service H&Ps, and 4% were intensive care unit admission notes. Sixty-three (20%) of the H&Ps had been entered using electronic templates; the remaining 80% were dictated and transcribed into the EMR. The SecTag algorithm identified 16,036 sections from the evaluation set (median 52 sections per document); 7,858 (49%) were major sections, such as chief complaint, physical exam, and assessment and plan (Table 1). The 16,036 identified sections contained 355 different header concepts from the SecTag terminology. The training and evaluation sets were similar in numbers of words, numbers of sections, and numbers of labeled sections per H&P document.

Of the 16,196 total sections identified, physician-reviewers classified 15,329 sections as true positives, 707 as false positives, and 160 as omitted sections (i.e., false negatives). The SecTag algorithm recall and precision were 99.0 and 98.6% for all section concepts and 98.6 and 96.2% for major section headers (Table 2). The algorithm more effectively identified labeled sections than unlabeled ones (recall 99.8 vs. 96.6%, $p < 0.001$). Table 1 shows the recall and precision for each major section type. The algorithm effectively identified labeled and implicit major section headers with recall 98.6% (range 87–100%) and precision 96.2% (range 90–100%). Document authors often failed to provide section labels for substance abuse history, vital signs, laboratory and radiology results, and first-degree relative family medical history (only 5% were labeled). The SecTag algorithm identified these unlabeled sections primarily with noun phrase processing and Bayesian prediction. Although differences were small, recall was slightly better for nonmajor sections (recall 99.3 vs. 98.6%, $p < 0.001$) and precision slightly better for major sections (95.0 vs. 96.2%, $p < 0.001$). When processing a document, the SecTag algorithm often identified major sections by finding more specific subsections and deducting...
Table 2 - Recall and Precision for SecTag Section Header Tagging

<table>
<thead>
<tr>
<th></th>
<th>Label in Document</th>
<th>No Label in Document</th>
<th>Total (Document)</th>
</tr>
</thead>
<tbody>
<tr>
<td>All Sections</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gold standard</td>
<td>11476</td>
<td>4720</td>
<td>16196</td>
</tr>
<tr>
<td>SecTag identified (TP+FP)</td>
<td>11456</td>
<td>4580</td>
<td>16036</td>
</tr>
<tr>
<td>Number tagged correctly</td>
<td>11353</td>
<td>3976</td>
<td>15329</td>
</tr>
<tr>
<td>Number tagged incorrectly</td>
<td>103</td>
<td>604</td>
<td>707</td>
</tr>
<tr>
<td>Number where SecTag</td>
<td>20</td>
<td>140</td>
<td>160</td>
</tr>
<tr>
<td>omitted correct tag (FN)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Recall</td>
<td>99.8% (99.7–99.9)</td>
<td>96.6% (96.0–97.2)</td>
<td>99.0% (98.8–99.1)</td>
</tr>
<tr>
<td>Precision</td>
<td>99.1% (98.9–99.3)</td>
<td>86.8% (85.8–87.8)</td>
<td>95.6% (95.3–95.9)</td>
</tr>
<tr>
<td>Major sections only</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gold standard</td>
<td>6325</td>
<td>1644</td>
<td>7969</td>
</tr>
<tr>
<td>SecTag identified (TP+FP)</td>
<td>6321</td>
<td>1537</td>
<td>7858</td>
</tr>
<tr>
<td>Number tagged correctly</td>
<td>6250</td>
<td>1310</td>
<td>7560</td>
</tr>
<tr>
<td>Number tagged incorrectly</td>
<td>71</td>
<td>227</td>
<td>298</td>
</tr>
<tr>
<td>Number where SecTag</td>
<td>4</td>
<td>107</td>
<td>111</td>
</tr>
<tr>
<td>omitted correct tag (FN)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Recall</td>
<td>99.9% (99.9–1.00)</td>
<td>92.4% (91.1–93.8)</td>
<td>98.6% (98.3–98.8)</td>
</tr>
<tr>
<td>Precision</td>
<td>98.9% (98.6–99.1)</td>
<td>85.2% (83.5–87.0)</td>
<td>96.2% (95.8–96.6)</td>
</tr>
</tbody>
</table>

TP = True positive; FP = False positive; FN = false negative. The gold standard was composed of physician review of SecTag-identified sections and manual identification of sections not labeled by SecTag.

The SecTag algorithm correctly identified the starting and ending boundaries for 92.7% (14,203/15,328) of the correctly labeled sections (Table 3). The system better predicted the boundaries for labeled sections than unlabeled ones (p < 0.001). The most common error was an ending error (5.5% [846/15,328] of all correctly tagged sections, 75% of all boundary errors), meaning SecTag either predicted the ending of the section (failed to include relevant content) or too late (included content that did not belong to that section). An analysis of 112 randomly selected incorrectly placed boundaries revealed that 56% of the boundary errors excluded relevant content (i.e., the section was terminated too early) from the section while 31% included too much information. About 15% of the boundary errors were due to nonclinical content in the section, such as outline headers, medical record numbers, or page numbers.

The inter-rater reliability on accuracy between all reviewers was good (Kappa = 0.70, p < 0.0001). Each evaluator reviewed an average of 36 H&Ps of the 60 assigned, and 11 documents were scored by multiple evaluators. Inter-rater reliability for placement of section boundaries was lower than for section header identification (Kappa = 0.49, p < 0.0001).

Precision of Section Identification Techniques

Table 4 shows the precision of each of SecTag’s component algorithms. Spelling correction was the worst performing algorithm, with a correct section prediction occurring with a precision of 62%. Correct spell-correction mediated matches included both multi-word and single-word matches (e.g., “cheif complant”, “labarotory”); all incorrect spell-correc-

Table 3 - Accuracy of Section Boundary Detection for Correctly Labeled Sections

<table>
<thead>
<tr>
<th>Section Boundaries</th>
<th>Label in Document (%)</th>
<th>No Label in Document (%)</th>
<th>Total (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct</td>
<td>10983 (96.7%)</td>
<td>3221 (81.0%)</td>
<td>14204 (92.7%)</td>
</tr>
<tr>
<td>Incorrect start</td>
<td>20 (0.2%)</td>
<td>197 (5.0%)</td>
<td>217 (1.4%)</td>
</tr>
<tr>
<td>Incorrect end</td>
<td>344 (3.0%)</td>
<td>502 (12.6%)</td>
<td>846 (5.5%)</td>
</tr>
<tr>
<td>Incorrect start and end</td>
<td>6 (0.05%)</td>
<td>56 (1.4%)</td>
<td>62 (0.4%)</td>
</tr>
<tr>
<td>Total</td>
<td>15329</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

CI = confidence interval; NLP = natural language processing.
tion mediated matches were single-word matches from document text. Eight spell-correction mediated errors were due to incorrectly disambiguating between the possible sections for an accurate spelling correction (e.g., “ucolor”, meaning urine color, became “skin color” since urine color was not a defined section), four were abbreviation/acronyms not present in the terminology (including a deidentified person’s initials), and six were the result of SecTag choosing a wrong spelling correction.

SecTag generated 1,664 possible sections headers for which it considered the “best” candidate section header a poor match and thus discarded it. A poor Bayesian score was the most frequent reason to discard a possible section label (58% of discarded sections). The authors evaluated 20 random notes to determine if the poor matches were appropriately discarded or not. Manual review suggested that 93% of the poorly matched section headers were appropriately discarded; 7% could have been instantiated as a section rather than discarded, but none were major section headers.

Discussion

The current study is one of the first large-scale efforts to formally evaluate a clinical note section header identification algorithm. To identify section labels in documents and predict where unlabeled sections belong, SecTag effectively combines NLP methods, concept matching approaches involving variant recognition, and scoring algorithms that include a naive Bayesian classifier. Using a large sample of general H&P documents from the EMR system of a single institution, we found that SecTag accurately identified the great majority of common section headers and boundaries. The algorithm employed a standardized section header terminology, which represented H&P section header concepts well, at least within the test institution.

Accurate section identification is a key first step toward greater automated or semiautomated clinical note processing. In future applications, section identification might be coupled with other NLP tools to improve decision support programs or to enhance clinical research. For example, a decision support system, operating on contextual understanding of concepts within a note, could suggest that a patient with a family history of colon cancer in a first degree relative or with a past personal history of ulcerative colitis should undergo early and more frequent colorectal cancer surveillance.

Future tools might use the SecTag algorithm to improve competency assessment regarding trainee’s clinical education. Such a tool might scan medical students’ or resident physicians’ notes to evaluate the completeness of their workups for patients with a specific condition. The tool might examine all trainee-generated notes on patients with back pain, to assess whether trainees elicited certain key faculty-designated history and physical examination elements (e.g., the presence of saddle anesthesia, incontinence, weakness, or weight loss). The tool could detect whether trainees discussed appropriate “red flag” diagnoses in the assessment and plan section. Toward this end, students and faculty at our institution are using a “learning portfolio” that couples SecTag with the KnowledgeMap concept identifier to track students’ clinical experience.

The ability to assign a block of text within a clinical document to a predefined concept within a section terminology hierarchy may improve concept identification, much in the same way as the Linguistic String Project found improved understanding by programming specific sublanguage grammars. For example, the acronym “BS” in the respiratory/chest examination section of the physical examination likely means “breath sounds” but means “bowel sounds” in the abdominal examination section, and possibly “blood sugar” (glucose) in the laboratory results section; likewise, such descriptors as “normal”, “non-tender”, or “not enlarged” may occur within many sections but indicate distinctly different clinical meanings and evoke different differential diagnoses based on context. For example, when parsing a “vital signs” section, a program might treat any floating point number between 35.0 and 40.0 as a Celsius temperature measurement, and a percentage between 60 and 100% as an oxygen saturation determination, especially if preceded by a segment labeled as a respiratory rate.

The current study’s failure analysis revealed several venues for potentially improving SecTag’s performance. Some SecTag errors were secondary to suppression of note text by DE-ID software “bugs”. The latter program treated certain disease eponyms as being a person’s name and “de-identified” them into uninterpretable terms; it sometimes also removed acronyms that were section labels (e.g., “GI”), ostensibly because the acronym matched the initials of a person (e.g., the patient, a physician, or a nurse). A second cause of failure was the spelling correction algorithm, which performed suboptimally. Because the Aspell algorithm used in the study did not support words containing numbers (e.g., the “S4” heart sound), SecTag performance degraded when these words were inappropriately “spell-corrected” to words without numbers. We have since adjusted the algorithm to omit such words from spell-correction. A few common medical words were missing from the spell-check vocabulary; these were added after completion of the study.

Finally, dictated documents often contained various forms of deidentified patient names, medical record numbers, and page numbers as new page headers—these often caused errors in section tagging. Imperfect sentence parsing also caused some errors in section boundary detection.

The Bayesian algorithm predicted correct placement of unlabeled sections with an accuracy of 81% for 631 different possible sections. This was not as accurate as other methods SecTag uses for identifying unlabeled section headers. However, the Bayesian score was critical in discarding erroneous candidate sections. A possible cause of error in the Bayesian prediction was an imperfect gold standard, derived automatically from iterative tagging of the training corpora. While the naive Bayes approach performed acceptably, more sophisticated algorithms, such as support vector machines, may perform better.

Limitations of the Current Study

There were several limitations of this study. The study used clinical notes from a single medical center; formatting, styles, typical content, and section header names may be different in other settings. We attempted to mitigate this bias by deriving the SecTag header terminology from external, nationally available sources such as common textbooks and existing standard vocabularies such as LOINC® and QMR.
Second, the prior probabilities for Bayesian scoring were derived from automatically tagged documents (instead of manually tagged documents). While this allowed quick derivation of a large tagged corpus accurate for most sections headers, a manually tagged corpus, would likely be more accurate, were it feasible to create one. Since the nonprobabilistic tagging performed better on major sections than subsections, SecTag is biased towards predicting parent concepts. We only evaluated the SecTag application on H&P documents; other document types (e.g., discharge summaries or operative reports) are likely to require additional training for optimum performance.

The performance of SecTag in identifying subsection tags may have been overestimated in some areas due to physicians’ original use of electronic templates in generating notes (20% of this corpus). Furthermore, evaluators were specifically told that matching specific laboratory and radiology subheaders was not a goal of this study. Labeling laboratory and radiology subheaders is less important because, first, these categories are already well represented in existing terminologies such as LOINC®, SNOMED CT®, and QMR, and, second, most EMRs provide laboratory and radiology results in structured and labeled formats. Physician-evaluators scored the SecTag output instead of creating a de-novo gold standard, which may bias them in favor of algorithm. Finally, evaluators may have been subject to an information bias in rating nonmajor section headers since they were instructed to concentrate on identifying major section headers.

Conclusions
The current study provides one of the first formal evaluations of a clinical note section header tagging algorithm. The SecTag algorithm accurately identified both labeled and unlabeled sections in randomly chosen H&P documents. The SecTag terminology contained appropriate matches, with very few exceptions, for the section header labels actually used by clinicians during patient care in their H&P notes. Additional research should evaluate and extend the current SecTag terminology and algorithms for other document types. In the future, the SecTag algorithm could be coupled with a robust concept identification system or a NLP system to provide a better contextual basis for disambiguating clinical terms embedded within source clinical documents.

References