The Horizontal and Vertical Nature of Patient Phenotype Retrieval: New Directions for Clinical Text Processing

Christopher G. Chute, M.D., Dr.P.H.,
Division of Medical Informatics Research, Mayo Clinic, Rochester, MN

Abstract

The author reviews the historical problem of identifying appropriate patients for retrieval from a clinical repository of patient records, compares the competing features of document classification and natural language processing, and proposes an alternative approach. The alternative approach 1) codes inquiries in an ontology to lend a vertical axis to retrieval knowledge instead of coding the target body of notes, 2) invokes natural language indexing and lexical normalizations on the corpus of notes that is scalable and tractable, and 3) leverages thesauri of word-level synonyms and near-synonyms to expand term searches "horizontally" around the concept spaces drawn from the ontology in which the queries were "coded."

Introduction

Medical informatics is by its nature a multidisciplinary undertaking. Blois resoundingly illustrated this insight in his inimitable précis[1], "Medicine and the nature of vertical reasoning." Vertical was used in that paper to mean the spectrum of perspectives relevant to medicine, from molecule to society. Influenced by Blois, I mean the same notion of vertical perspective, though as it pertains to concepts – ontologically arranged. I further posit that concepts can express a horizontal nature, manifest in synonymy and near-meaning terms, which I distinguish from vertical ontology. How these directions differ and, more importantly, why that should matter in the practical operation of medical text retrieval form the basis of this paper.

The Phenotype Use-Case

Phenotype is conventionally contrasted with genotype to characterize how genetic variations are manifest in living organisms. It is reasonable to outline levels of evidence for phenotype expression, ranging from the functional behavior of enzymes in vitro to disease manifestation or natural history in the whole organism. For present purposes, I invoke the whole organism model, and further, assume human patient populations.

Epidemiologists have routinely conducted "phenotypic" retrievals on patient records for nearly a century. By this use, I mean the identification of patients within a record repository who express certain characteristics, which typically implies diagnosis but may equally invoke signs, symptoms, interventions, functional status, or clinical outcomes.

I deal in this paper with the general problem of patient phenotype recognition for study cohort retrieval. I suggest a new approach that differs substantially from traditional patient data retrieval strategies.

Medical Record Information

Retrieval systems for identifying patients appropriate for a study from a repository of health records data motivates this paper. Hence, the domain of medical information considered here is restricted to patient medical record data. I do not directly consider the gnarly body of medical knowledge, manifest in literature, reference works, and libraries of decision support resources, though one can readily see how the principles outlined for the patient record use-case could be adapted to broader domains of medical information.

Within the scope of medical records, I further restrict emphasis to textual descriptions. The modern record is replete with continuous variables, iconic and graphical data, diagnostic images, and time-series tracings. While a proper notion of patient cohort retrieval should embrace the full spectrum of patient characterization, the practical inferencing into conceptual categories of non-textual sources is beyond the scope of this paper.

Text, Data, and Medical Things

Textual objects within medical records are a highly inclusive category, ranging from erudite prose on the presentation, treatment, and course of a patient to the banal marker of patient gender by a diminutive "M." Both are textual, and both have bearing on the patient cohort retrieval use-case. However, the latter comprises a degenerate form of classification, in that such byte-size expressions represent an instance drawn from an enumerated value-set. We will deal with coded information in the next section.

The major body of machine-readable reports, notes, and summaries comprises the primary target of interest for identification of patient cohorts. Ranging in length from a few sentences to pages, these text objects contain complex and detailed descriptions of phenotypic characteristics. However, strategies for identifying and harvesting these descriptions vary. I suggest a new approach, invoking ontologies and thesauri.

Classification, Codes, and Concepts

The dominant modality for patient phenotype recognition and retrieval is still based on coded information. The boundaries among categories, codes, and concepts are subtle[2] but substantially affect how such divisions are used practically. Classification is the assignment of a record to a category in an ordered set of concepts, such as an entry in the ICD-9-CM classification. A code is an arbitrary identifier, typically a number, associated with a category. Concepts are the abstract notions intended by the code or category term.

These distinctions become more complex when one attempts to differentiate high-level classifications (such as the ICDs) from detailed nomenclatures (such as SNOMED-CT). The salient difference is one of degree in levels of granularity and specificity along a common axis. However, for our purpose, classifications and nomenclatures both represent underlying ontologies of medical knowledge, albeit with differing specificity.

Information Retrieval

The domain of information retrieval is vast. Ruch has recently observed that medical information is materially different from general information, in that medical text tends to have more consistent phrasing and less ambiguous text[3]. Thus generalizations made about the retrieval of newspaper articles or Web documents may not pertain to medical text, and vice versa.

Literature-Finding Legacy

The rise of MedLars in the early 1960s fundamentally changed how health professionals and librarians regarded the medical literature. Manhandling tomes of Index Medicus to identify relevant articles now seems quaint to many of us, and unimaginable to those trained a few short years after my generation. Nevertheless, the present medical literature retrieval infrastructure is profoundly dependent on the human encoding of documents against a well-maintained ontology – MeSH (Medical Subject Headings)[4].

Tremendous efforts have been made to automate the assignment of MeSH codes to the literature, some more successful than others. It is notable that human assignment remains the dominant encoding method, despite its considerable expense and measurable inconsistency. However, the expense is not born by the end user, and the imprecisions are difficult to detect absent high-volume, gold-standard comparison sets. These comparison sets have been published and used[5, 6], though by design they are restricted to a small set of queries and answers because of cost.

Nevertheless, when describing a phenotype retrieval paradigm, comparison to the medical literature metaphor is inescapable. The challenge is to develop methods and interfaces that are perceived to be as useful and friendly as PubMed. This general approach is presented in the "Classify the Data" section within "Standard Retrieval Models" below.

Text Mining in Medicine and Elsewhere

The Department of Defense and other government agencies have sponsored a long-standing series of conferences where the precision and recall of competing algorithms and programs for text retrieval against a standard set of documents are compared[7]. While intended to generalize against any kind of information source, such as newspaper reports of potential interest to the CIA, the exhibited principles should pertain to medical text retrieval. However, these use-cases differ considerably from those encountered in typical medical retrievals.

The general tradition of natural language processing (NLP) of medical text was well reviewed by Friedman[8]. Suffice it to say, the state of the art for well-supported NLP environments can deliver retrieval performance rivaling that of humanly encoded collections[9], though substantial barriers persist[10].

Concept-Based Retrieval

Concept-based retrieval means that documents are recognized as related to a given concept, typically within an ontology. It is to be distinguished from key-word or text-word retrieval, though those tokens often illustrate a general concept. As a principle, it is difficult to differentiate concept-based retrieval from the retrieval of documents based on the assignment of ontology terms, such as the classic MeSH-based literature retrievals. However, the difference between concept-based retrieval and that involving humanly assigned categories is the human element. The former implies a machine assignment of a concept to a term. Preliminary work in our lab on this problem originally invoked statistical methods[11], though we later adopted lexical adjuncts[12] after the fashion of MetaPhrase[13].

The Canon Group in the early 1990s attempted to define a basis for concept-based term recognition using a common information model[14], which many groups have continued[15, 16].

Standard Patient Retrieval Models

The historical choices confronted by researchers who wanted access to a corpus of medical text records included 1) classifying the data, 2) using a nomenclature to encode data fields, or 3) employing natural language processing on the text. A fourth logical option, invoking statistical machine-learning techniques, has been tried, but for the most

part has not proven itself scalable and reliable for large-volume patient retrieval tasks[17].

Classify the Data

Perhaps the oldest among the options, "classifying the data" assigns a high-level category to a text-document instance. The most visible example of this method is the human assignment of MeSH terms to journal articles in the medical literature for retrieval by Medline engines, such as PubMed. In the domain of medical records, ICD-9-CM codes are frequently used, though these assignments tend to occur for the purpose of billing rather than efficient text document retrieval. Furthermore, reimbursement coding tends to treat the entire record as "the document," rather then focusing upon discrete notes or summaries.

For nearly a century Mayo Clinic has used a record classification process for research retrieval. Needless to say, at its introduction in 1907, it had little to do with payment or regulation[18]. Since 1975 we have used an obscure variation on ICD-8, which is a highly modified and extended version of HICDA-2[19].

The difficulties with this approach are many, though chief among them is the coarse granularity of the classifications used. Getting back the heterogeneous group of patients having "Other bowel polyposes" may not well satisfy those seeking instances of Puetz-Jaegers syndrome.

The cost and effort associated with human classification of complex medical records is not a scalable or sustainable undertaking in modern research infrastructure. Furthermore, even the most experienced nosologists exhibit inconsistency when confronted with very large coding spaces such as MeSH or SNOMED.

Use a Nomenclature to Encode Data Fields

The differences between encoding data fields and classifying a document involve scope (short phrases and elements vs. an entire document) and granularity (detailed nomenclature vs. high-level classification). The archetype of this activity is the selection of what I have called sentinel fields, such as Reason for Visit, Indication for Order, or Dismissal Diagnosis, and then encoding the phrases or terms in those fields using a nomenclature such as SNOMED-CT. This pluraity implies that a given patient may have many codes associated with a document, and certainly more codes associated with his or her record.

The major advantages to sentinel field encoding are the specificity enabled by focal retrievals and the appropriateness of compositional expressions to represent specific field descriptions. The disadvantage is a paucity of algorithmic methods to achieve such algorithmic coding.

Employ Natural Language Processing

While the general problem of NLP systems was well summarized by Friedman[8], Cimino and colleagues have applied NLP to the general patient cohort retrieval problem[20]. Baud has adopted the traditional NLP approaches to use a shallow semantic model, closely aligned with what I propose in this paper[21]. He further suggests that such a "light model" may be sufficient to achieve the leveraging of traditional NLP with semantic enhancements[22].

Baud directly addresses the most interesting aspect of NLP work, which attempts to meld "dumb" NLP with semantic knowledge. Johnson has previously attempted to characterize semantic type in context to disambiguate multiple word sense terms[23]. The next step beyond this is to invoke a thesaurus of synonyms and near-synonyms (plesionyms), which preserve semantic distance distinctions.

A Proposed New Approach

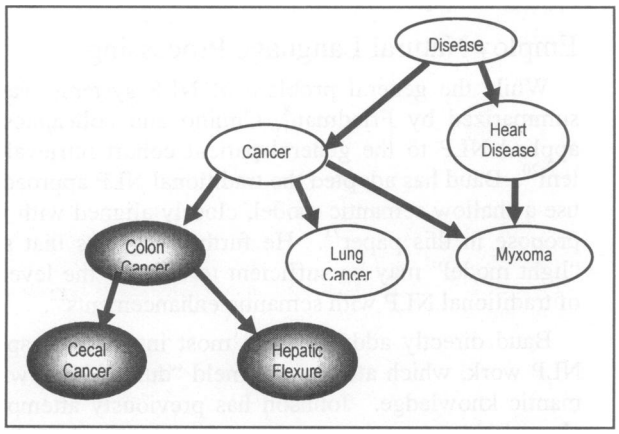
Confronted with the non-scalable task and imprecision of coding large corpus of medical text using ontology classifications, or relying on NLP techniques that may not fully exploit the knowledge and relationships within an ontology, we chose to turn the problem around. Briefly, our approach is based on 1) coding the question rather than the corpora to invoke ontology knowledge (the vertical axis), 2) using NLP indexing and lexical normalization techniques on the corpora, and finally 3) invoking thesauri of synonyms and plesionyms to expand horizontally the scope of words and terms considered in the retrieval.

Code the Question

The advantages of using well-formed ontologies to leverage information retrievals have been well established by the medical literature (MeSH) and healthcare reimbursement (ICD-9-CM) retrieval communities. Traditionally, one codes the body to be searched and the question into the same classification, and then matches codes. Computationally, these retrievals are efficient, since one is executing Boolean logic on vectors of pointers.

However, the requirement to code both the corpus and the question does not scale for large repositories of patient descriptions. The cost of human coding is prohibitive. Nevertheless, one can preserve the advantages of concept explosion, semantic neighborhood navigation, sibling concept inclusion, and the appropriate scope and scale of concepts requested for retrieval by coding the question alone. The figure illustrates a trivial example of concept explosion for colon cancer, using a shallow ontology akin to the ICD. This step corresponds to what we are calling a vertical concept expansion, borrowed from the graphic convention of having parent concepts above their children.

In this example the problem of coding the question



is certainly scalable, which would permit the recovery of all the semantic children associated with the question as "known" by the ontology. A related question is then which ontology? This question bears careful evaluation, but logical candidates include MeSH, SNOMED-CT, or even ICD-9-CM for certain classes of query.

Index and Normalize the Text

While we explicitly eschew coding the text corpus against an ontology, one cannot leave raw text without any preprocessing. Fortunately, the real-time Web-based search engines that can scan the world's collection of Web pages in seconds testify to the tractatibility of making word-level indices to those texts. Such indices can be quite naïve about normal lexical form variants, such as plurals, tense or case changes, and gender. The Specialist Lexicon[24] of the NLM has proven to be the definitive resource for achieving a rapid and reliable normalization of medical text terms.

We have explored whether one should normalize first, and index the normal form, or index first, and then normalize a copy of the index. While the computationally most efficient path might be the first, we have opted for the second approach, which preserves a text retrieval index of the "raw" form of all words contained in the corpora. The lexical normalization is robust, but not perfect, and may occasionally distort a word or term. The post-hoc normalization permits graceful recognition and recovery of these error states when they occur.

Invoke Thesauri

Because we opted not to code the corpora, we confront the question as to what else must be done to match coded questions against a minimally preprocessed body of text. At a minimum, we can be confident that exact word matches can be accommodated, as can lexical variants, thanks to the Specialist Lexicon (pre- or post-hoc). What then to do with words that are neither exact matches nor lexical variants?

Many ontologies, such as MeSH, include term-level synonyms in their content. For example, MeSH uses Colon Cancer as an "entry term" or synonym to the canonical form Large Bowel Malignant Neoplasm. Thus a ready supply of term-level synonyms can be harvested from the ontology; this harvest will greatly improve retrieval. However, a list of entry terms is clearly not sufficient, as term-level synonyms rarely include all the word-level permutations that occur naturally. For example, Colon Malignant Neoplasm is unlikely to exist as a synonym, even though it could be inferred from the word-level similarity between Colon and Large Bowel.

This raises the problem of word-level synonymy. Consider Renal Cancer and Kidney Cancer. Most users would be happy to retrieve cases referenced by one, even though they had asked the other. The permutation of such word-level synonyms becomes astronomical in phrases of even moderate length. Furthermore, what of terms that are close but not exact synonyms? Would a user just as soon have Renal Neoplasms, in addition to Cancers? As indicated previously, such closely similar terms are known as plesionyms.

We propose that thesauri of such synonyms and plesionyms, used in combination with term-level synonyms derived from an ontology, would comprise a horizontal expansion of appropriate terms around a concept space to simulate the behavior we would have seen if the corpora had been coded in the first place.

Thesauri Construction

The construction of thesauri that can usefully function in this horizontal and vertical paradigm is a formidable challenge[25]. We have begun the creation of such a thesaurus at Mayo, driven by the identifiable examples in our corpora of 13M patient notes.

An immediate question is when does a plesionym cross a "vertical distance boundary" to become a different concept? For example, is cancer an ontologic child of neoplasm, or are cancer and neoplasms reasonably considered near-synonyms? The answer, naturally, is that it depends. Ideally, the words might function as plesionyms when one would like them to, and not at other times. Relative synonymy introduces the difficult concept of semantic distance as a criterion for plesionym definitions.

Regrettably, none of the well-established mechanisms for determining semantic distance will operate across the micro-distances involved for plesionymy. Therefore, we are experimenting with two candidate methods:

1. when retrievals that invoke a plesionym pair exhibit better precision and recall than when not invoked, the pair are true plesionyms.

This method has the advantage of unassailable validity (almost to the point of tautology) but is monumentally non-scalable. It requires the labor- intensive verification of true retrievals, and then may suffer from being question-context dependent.

2. when terms appear in the same lexical company within sentences and expressions, they have a probability of near-synonymous behavior.

This method is wonderfully scalable and relatively inexpensive, but suffers from poor validity. It is not likely to be question-dependent, though the low yield from its usage may preclude its practicality for plesionym discovery.

We see thesaurus building, with scalable and valid methods for plesionym authoring, as the next great challenge for high-volume NLP against "naïve" or un-coded textual datasets.

Privacy and Confidentiality

We would be remiss were we not to mention the ever-present concerns about the misuse of such tools, or the failure to respect the privacy and confidentiality of patient data content. Our work at Mayo is undertaken with de-identified notes where practical, and otherwise done under IRB-approved protocols.

1. Blois MS, *Medicine and the nature of vertical reasoning.* New England Journal of Medicine, 1988. 318(13): p. 847-851.
2. Rector AL, *Clinical Terminology: Why is it so Hard?* Methods of Information in Medicine, 1999. 38(4-5): p. 239-252.
3. Ruch P, *et al., Comparing general and medical texts for information retrieval based on natural language processing: an inquiry into lexical disambiguation.* Medinfo, 2001. 10(Pt): p. 261-5.
4. Nelson S, *Medical Subject Headings (MeSH®).* 2002, National Library of Medicine: Bethesda, MN. p. 2.
5. Hersh WR, *et al., A performance and failure analysis of SAPHIRE with a MEDLINE test collection.* J Am Med Inform Assoc, 1994. 1(1): p. 51-60.
6. Hersh WR, *et al., Factors associated with successful answering of clinical questions using an information retrieval system.* Bull Med Libr Assoc, 2000. 88(4): p. 323-31.
7. Harman DK, *The First Text Retrieval Conference (TREC-1).* . 1993, Washington, D.C.: U.S. Department of Commerce.
8. Friedman C, Hripcsak G, and Shablinsky I. *An Evaluation of Natural language Processing Methodologies.* in *Proceedings of the 1998 AMIA Annual Symposium.* 1998. Orlando, FL: Hanley & Belfus, Inc.
9. Baud RH, *et al., A toolset for medical text processing.* Stud Health Technol Inform, 2000. 77: p. 456-61.
10. Hripcsak G, Kuperman G, and Friedman C, *Extracting Findings from Narrative Reports: Software Transferability and Sources of Physician Disagreement.* Methods of Information in Medicine, 1998. 37: p. 1-7.
11. Chute CG and Yang Y, *An Evaluation of Concept-based Latent Semantic Indexing for Clinical Information Retrieval.* Symposium on Computer Applications in Medical Care, 1992. 16: p. 639-643.
12. Elkin PL, Bailey KR, and Chute CG, *A randomized controlled trial of automated term composition.* Journal of the American Medical Informatics Association, 1998. SympSuppl: p. 765-769.
13. Tuttle MS, *et al., Metaphrase: an aid to the clinical conceptualization and formalization of patient problems in healthcare enterprises.* Methods Inf Med, 1998. 37(4-5): p. 373-83.
14. Evans DA, *et al., Toward a medical-concept representation language. The Canon Group.* J Am Med Inform Assoc, 1994. 1(3): p. 207-17.
15. Rassinoux A-M, *et al., Modeling concepts in medicine for medical language processing.* Methods of Information in Medicine, 1998. 37(4/5): p. 361-372.
16. Rector AL, *et al., Medical concept models and medical records: An approach based on GALEN and PEN&PAD.* Journal of the American Medical Informatics Association, 1995. 2(1): p. 19-35.
17. Chute CG and Yang Y, *An Overview of Statistical Methods for the Classification and Retrieval of Patients Events.* Methods of Information in Medicine, 1995. 34(1/2): p. 104-110.
18. Kurland L and Molgaard C, *The patient record in epidemiology.* Scientific American, 1981. 245(4): p. 54-63.
19. Activities. CoPaH, *H-ICDA, Hospital Adaptation of ICDA, Second Edition.* 1968, Ann Arbor, MI: Commission on Professional and Hospital Activities.
20. Zeng Q and Cimino JJ, *Evaluation of a system to identify relevant patient information and its impact on clinical information retrieval.* Proc AMIA Symp, 1999: p. 642-6.
21. Baud RH, *et al., Conceptual search in electronic patient record.* Medinfo, 2001. 10(Pt): p. 156-60.
22. Baud RH, *et al., A light knowledge model for linguistic applications.* Proc AMIA Symp, 2001: p. 37-41.
23. Johnson SB, *A semantic lexicon for medical language processing.* J Am Med Inform Assoc, 1999. 6(3): p. 205-18.
24. McCray A, Srinivasan S, and Browne A, *Lexical methods for managing variation in Biomedical Terminologies.* Journal of the American Medical Informatics Association, 1997. SympSuppl: p. 193-201.
25. Bodenreider O, *et al., Beyond synonymy: exploiting the UMLS semantics in mapping vocabularies.* Proc AMIA Symp, 1998: p. 815-9.