Example of Literature Review

The following was extracted from

Chapter 2: Background

2.1 Approaches to EHR summarization

Given the unmet and well-recognized need for comprehensive EHR summarization (Powsner and Tuft 1994; Payne 2000), many research groups have designed and evaluated clinical data summarizers. In this chapter, we sample summarization applications to highlight different features including seminal work, different evaluation strategies, and various input/output data. We also examine the current work and future directions for six challenges of EHR summarization: information redundancy, temporality, missing data, salience detection, rules and heuristics, and deployment of summarization tools.

There are multiple theoretical frameworks for summarization in the clinical domain (Feblowitz et al. 2011) as well as for textual summarization in the general domain (Alterman 1991; Radev, Hovy, and McKeown 2002). In the broader field of summarization, there has been a lot of work in automated text

---

1 A large part of this chapter was originally published in JAMIA. The full citation for this publication is: Pivovarov R, Elhadad N. Automated Methods for the Summarization of Electronic Health Records. JAMIA 2015.
summarization, specifically within the genres of news stories and scientific articles (see (Nenkova and McKeown 2012) for an in-depth review). Clinical summarization, “the act of collecting, distilling, and synthesizing patient information for the purpose of facilitating any of a wide range of clinical tasks” (Feblowitz et al. 2011), presents a different set of challenges from summarization in other domains and genres of texts.

While there exist other discussions on biomedical literature summarization methods (Afantenos 2006; Mishra et al. 2014) and EHR visualizations (Roque, Slaughter, and Tkatsenko 2010; Rind et al. 2010; West, Borland, and Hammond 2014), in this review we focus on characterizing existing clinical summarization systems by outlining the system outputs and evaluations as well as highlighting the remaining challenges that exist in automated summarization.

To categorize the summarizers highlighted in this review, we focus on two common dimensions used in the text summarization literature: extractive/abstractive summarization, and indicative/informative summarization. We define the four categories that describe summary types.

1. **Extractive summaries** are created by borrowing phrases or sentences from the original input text. In the domain of clinical summarization, an extractive approach can identify pieces of the patient’s record and display them without providing additional layers of abstraction.

2. **Abstractive summaries** generate new text that synthesizes the original text. In the domain of clinical summarization, abstractive summaries may provide additional higher-level context to explain the data, such as computed quantities (e.g., trends) or automatically generated text.

Extractive and Abstractive summaries are further categorized as either indicative or informative.

3. **Indicative summaries** point to important pieces of the original text, highlighting significant parts for the reader. In the domain of clinical summarization, indicative summaries may convey, for instance, when key tests were performed or diagnoses were made. Indicative summaries are meant to be used in conjunction with the full patient record.
Informative summaries replace the original text. In the domain of clinical summarization, informative summaries are designed to be used independently of the full patient record, meaning they are used as a replacement for the original full set of raw data.

How to evaluate a summarizer, both its accuracy and its added value in supporting users carry out information-related tasks has also been the subject of investigation in general domain and clinical summarization. Intrinsic evaluations focus on the internal validity of a summarization tool. Typically, experts evaluate the quality of the automatically produced summaries; or themselves create gold-standard summaries, against which automatic ones are compared. In an extrinsic evaluation framework, the usefulness of the summarization tool is assessed through its effectiveness in helping individuals carry out a task. For instance, a clinical summary could be evaluated in an extrinsic fashion by comparing how quickly and accurately trial coordinators can identify patients eligible for a trial with access to patients’ full records or with access to a summary instead.

Almost since the inception of EHRs, there has been an interest in creating meaningful succinct summaries for clinicians. The research on automated summary creation has spanned over 30 years and initiated with extracting recent structured events in a patient’s history (Rogers and Haring 1979) evolving into performing natural language processing (NLP) (Liu and Friedman 2004) and automatically linking different data types (Cao et al. 2005; Klann et al. 2013) to create a more holistic view of the patient record. Table 1 lists clinical summarization systems proposed in the research literature in chronological order. We describe each system according to the following axes: the summarization approaches it implements, the type of input data it handles, the type of output summary, the way in which it was evaluated, and whether it was deployed in a clinical environment. Overall, summarization approaches investigated in clinical summarization have primarily been for indicative and extractive summarization. We also note a lack of evaluation, especially in the most recent years. We discuss in further detail the methods used for summarizing clinical data, along with the open research questions present in each of the summarization steps.
<table>
<thead>
<tr>
<th>Summarization approach</th>
<th>Input</th>
<th>Output</th>
<th>Evaluation</th>
<th>Deployed (when is it generated)</th>
<th>General Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>NUCRSS (Rogers and Haring 1979; Rogers, Haring, and Watson 1979)</td>
<td>Extraction of clinical variables, indicative</td>
<td>Real structured EHR data</td>
<td>An 8 page summary of: Problem list, Vital signs, Cardiac-pulmonary-renal diagnoses, Treatments, Routine specialized laboratory examination, Suggestions to physicians regarding patient care</td>
<td>Laboratory study with medical students and physicians showed significant time savings and increased accuracy</td>
<td>Early example of a summarizer. One of the few summary evaluations that demonstrate an impact on quality of care and process outcomes. The study found that the use of the summarizer was able to reduce patient’s length of stay and increase the amount of laboratory tests that were ordered (perhaps because summarizer was able to remind clinicians of all the patients conditions that need monitoring).</td>
</tr>
<tr>
<td>STOR (O’Keefe and Simborg 1980)</td>
<td>Extraction of clinical variables, indicative</td>
<td>Real structured and unstructured EHR data</td>
<td>Loosely customizable, summary which included both time- and problem-oriented views</td>
<td>Clinical study found that clinicians were better able to predict their patient’s future symptoms and laboratory test results when the using medical record in addition to STOR as opposed to just the medical record.</td>
<td>Yes</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>Powsner and Tufte (Powsner and Tufte 1994; Powsner and Tufte 1997)</td>
<td>Extraction of psychiatric variables and recent notes, indicative</td>
<td>Simulated structured, unstructured and genealogy data</td>
<td>A one page summary that visualizes the most salient content (as defined by recency) of the patient record.</td>
<td>None</td>
<td>No</td>
</tr>
<tr>
<td>Lifelines (Plaisant et al.)</td>
<td>Extraction of</td>
<td>Simulated</td>
<td>Holistic</td>
<td>The original Lifelines</td>
<td>No</td>
</tr>
<tr>
<td>Date</td>
<td>Description</td>
<td>Data Type</td>
<td>System Function</td>
<td>User Feedback</td>
<td>Conclusion</td>
</tr>
<tr>
<td>------------</td>
<td>------------------------------------------------------------------------------</td>
<td>------------</td>
<td>---------------------------------------------------------------------------------</td>
<td>------------------------------------------------------------------------------</td>
<td>----------------------------------------------------------------------------</td>
</tr>
<tr>
<td>1996; Plaisant et al. 1998)</td>
<td>Clinical variables, indicative structured data</td>
<td>Interactive patient summaries using a temporal data view on top of the raw EHR data. Displays facts as lines on graphic time axis according to their temporal location and categories/significance are represented by color and thickness.</td>
<td>Application was evaluated for work with juvenile youth records (Plaisant et al. 1996) by a small group of users who reported enthusiasm but mentioned potential biasing by the system’s graphics.</td>
<td>Well-known summarizer tool. The display has served as a model for future timeline-view clinical summarizers. Lifelines2 was created for research and examining many patients together.</td>
<td></td>
</tr>
<tr>
<td>CliniViewer (Liu and Friedman 2004)</td>
<td>Extraction of concepts from text, indicative Real unstructured EHR data</td>
<td>Combined NLP techniques and presented a tree view of a patient’s summaries.</td>
<td>The system was able evaluated on accuracy and speed using real discharge summaries but no</td>
<td>One of the first examples of summaries created using NLP Allows for customizable user</td>
<td></td>
</tr>
<tr>
<td>Problems extracted from the narrative text to the clinician</td>
<td>Displays concepts in context when clicked</td>
<td>Evaluation with clinicians was conducted.</td>
<td>An interactive display of both navigational capabilities for the unstructured and structured text and extraction of clinical variables, indicative of:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>---------------------------------------------------------</td>
<td>------------------------------------------</td>
<td>-------------------------------------------</td>
<td>-----------------------------------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HNC Patient Worksheet (Wilcox et al. 2005)</td>
<td>One of the few example of a clinical outcome tested in the evaluation</td>
<td>A retrospective cohort study found that compliance with HbA1c testing was higher for patients who had a worksheet printed than for those who didn't.</td>
<td>CLEF (Hallet and Scott 2005; Rogers, Puleston, and Rector 2006; Hallett 2008)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Extracts demographic, problem, medication, laboratory test, actionable advisory data.</td>
<td>Yes each patient visit</td>
<td>None</td>
<td>Abstraction from clinical text and extraction of relevant clinical data</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Works on top of the MedLEE (Friedman et al. 1994) NLP engine which handles modifiers</td>
<td></td>
<td></td>
<td>Represents histories as a</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Views</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
| KNAVE-II (Shahar et al. 2006) | Abstraction and extraction of clinical variables, informative | Real structured data on bone marrow transplant patients | Interactive data display of abstracted and raw protocol-based care data containing a tree-semantic network of events organized temporally and semantically. | Lists requirements that are very relevant to general designers of clinical summaries – the list was generated via initial requirements elicitation process. Users produced quicker answers, had somewhat No Performs semantic, temporal, and context abstraction. Requires Domain-specific ontologies. Consists of a knowledge base, abstraction generator,
<table>
<thead>
<tr>
<th>Study</th>
<th>Description</th>
<th>Input</th>
<th>Output</th>
<th>Result</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>BabyTalk (BT-45) (Hunter et al. 2008; van der Meulen et al. 2010)</td>
<td>Abstraction of ICU data streams, informative</td>
<td>Real raw neonatal ICU data streams</td>
<td>Automatically generated natural language to describe ICU data streams for easier comprehension by the nursing staff.</td>
<td>A laboratory study found that human-generated text summaries of ICU streams helped nurses predict their patient’s trajectories’ better. The team is working to create automatically generated text summaries that perform as well as human-generated summaries.</td>
<td>No</td>
</tr>
<tr>
<td>Were et al. (Were et al. 2010)</td>
<td>Extraction of clinical</td>
<td>Real structured</td>
<td>Patient summary for use in an HIV</td>
<td>A pre-post study design using time-motion study</td>
<td>Yes (each)</td>
</tr>
<tr>
<td>Method</td>
<td>Abstraction from text and extraction of clinical variables, indicative</td>
<td>EHR data from OpenMRS</td>
<td>An interactive data display that summarizes and integrates various pieces of the EHR including images and free text.</td>
<td>A pilot study on Timeline found that although the initial learning curve was high, with time, the clinicians were able to perform image review quicker and were more confident in their clinical conclusions than when they used the EHR display</td>
<td>Timeline had manually coded rules while AdaptEHR aims to automatically infer rules and relationships from ontologies and graphical models, the publication states that the conditional probability tables are not yet defined. Has four dimensions of representing data: time space</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>Timeline/AdaptEHR (Bui, Aberle, and Kangarloo 2007; Bashyam et al. 2009)</td>
<td>Abstraction from text and extraction of clinical variables, indicative</td>
<td>EHR data from OpenMRS</td>
<td>A pilot study on Timeline found that although the initial learning curve was high, with time, the clinicians were able to perform image review quicker and were more confident in their clinical conclusions than when they used the EHR display</td>
<td>No</td>
<td>Timeline had manually coded rules while AdaptEHR aims to automatically infer rules and relationships from ontologies and graphical models, the publication states that the conditional probability tables are not yet defined. Has four dimensions of representing data: time space</td>
</tr>
<tr>
<td>HARVEST (Hirsch et al. 2014)</td>
<td>Extraction of concepts from text and clinical variables, indicative</td>
<td>Real structured and unstructured EHR data</td>
<td>A problem-based, interactive, temporal visualization of a longitudinal patient record.</td>
<td>A task-based, time evaluation found no difference in ability to extract, compare, synthesize and recall clinical information when using HARVEST in addition to the EHR, when carried out with subjects who had no prior experience with the summarization tool</td>
<td>Yes (real time)</td>
</tr>
</tbody>
</table>

**Table 2.1 A sampling of clinical summarization applications.** The table is organized by publication date. The inputs, outputs, methods, and evaluation strategies are listed along with notable additional information for each summarizer.
2.2 Methodological challenges to EHR summarization

The following sections present some unsolved challenges in clinical summarization. A conceptual framework proposed by Feblowitz et al. (Feblowitz et al. 2011) defines a set of actions that successful summarizers should accomplish with raw information: Aggregate, Organize, Reduce/Transform, Interpret, Synthesize. We discuss methodological challenges with automated summarization within the context of this framework.

Specifically,

- To successfully aggregate disparate clinical data sources, the ability to recognize and account for similarity is imperative. Such similarity occurs at different levels within narratives: from word-level similarity to concept to statement-level; as well as in other data types and across. We focus our discussion on textual similarity.

- The organization and interpretation of the aggregated data requires extraction and reasoning over clinical events and their temporality. We examine extraction of temporal information from text along with representation and reasoning over clinical events.

- The organization and interpretation of the aggregated data also requires that missing data points be accounted for. Patients are sometimes seen with predictable regularity but are most often seen at erratic intervals. Missing data points are often filled in by imputation, adding missing data indicators, deleting information with missing data, or other strategies.

- In the reduction and transformation of data and its synthesis, it is critical to decide which pieces of information are important and must be contained in the summary. Some methods for automatically detecting importance have relied on linguistic structure while others use probabilistic modeling techniques.

- To provide context for interpretation and synthesis of clinical data, it is useful to employ existing knowledge and create rules for the summarization. Knowledge-based heuristics often provide a way to specify time constraints, concept relationships, and abstractions.
Finally, to successfully implement summarizers into clinical care, challenges of deployment need to be addressed. Because in vendor EHR systems there are limited opportunities to deploy innovative and experimental technology, there have been few attempts to translate patient record summarization systems into the clinic; however, to demonstrate utility, it is imperative to implement and study clinical summarization tools in the real world care setting.

2.2.1 Identifying and aggregating similar information

We review approaches to identifying and aggregating similar information on three different levels of language abstraction: words, concepts, and statements, as investigated within and outside the field of clinical summarization.

Word-level Similarity

In clinical NLP, much work has been devoted to identifying lexical variants that are similar in meaning (Friedman and Elhadad 2014). The Unified Medical Language System (UMLS), an agglomeration of different biomedical terminologies (Lindberg, Humphreys, and McCray 1993), for example, provides essential knowledge towards that goal by grouping words into concepts. The UMLS aggregates terms from different vocabularies and maps them to semantic concepts, each labeled with a Concept Unique Identifier (CUI). For instance, the terms MI, myocardial infarction, and heart attack all share lexical similarity, and map to the same underlying CUI. Within clinical summarization, normalization of words to concepts has only recently been investigated (Hirsch et al. 2014; Zhang et al. 2011).

An alternative, and most common approach in clinical summarization, is to identify word-level similarity by finding redundant strings of words. Patient records often contain redundant spans of text – this can be explained by the fact that documentation is often formulaic but also by the common habit of clinicians to copy and paste text from one note to another (Hirschtick 2006). Multiple different automated methods have been employed to identify copy and pasted words within clinical notes. A plagiarism detection tool called CopyFind has been used to identify overlapping phrases in input texts (Thornton et
al. 2013). More recently, global (Wrenn et al. 2010) and local (Zhang et al. 2011; Cohen, Elhadad, and Elhadad 2013) bioinformatics-inspired alignments have been proposed for identifying redundant sections along with language modeling techniques for assigning probabilistic similarity scores for phrase pairs (Zhang et al. 2011).

Concept-level Similarity

Concept-level similarity represents a more abstract level of similarity than similarity between words and strings. For instance, the concepts “epilepsy” and “seizure” – despite being two different UMLS concepts – share much semantic similarity when conveyed in a patient record.

In certain well-defined domains, clinical summarization approaches have relied on aggregating concepts, helping further the goal of synthesis (Shahar et al. 2006; Hsu et al. 2012) primarily through well-defined ontologies. For broader domains, how to identify that two semantic concepts are similar enough to be aggregated remains an open question. Furthermore, in text processing, mapping from words to concepts remains difficult because of the strong ambiguity of language (Friedman and Elhadad 2014).

Detection of semantic redundancy has been investigated through two approaches: knowledge-free and knowledge-based. Knowledge-free similarity metrics have been developed for textual input. They rely on Harris’ 1968 hypothesis which stipulates that concepts that appear in similar contexts are similar (Harris 1968). In practice, concepts are compared in a vector space, where each concept is a vector representing the context in which the concept typically occurs. This method has been implemented multiple times in the clinical domain to identify similar UMLS concepts (Pedersen et al. 2007; Patwardhan and Pedersen 2006; Pivovarov and Elhadad 2012). Knowledge-free approaches are attractive when there is little ontological knowledge available. Alternatively, knowledge-based methods leverage existing resources to determine the similarity of two concepts. For instance, if the two concepts are present in an ontology, similarity can be assessed through the structure of the ontology. Other knowledge-based methods include examining similarity of the two concepts’ definitions. We refer the reader to detailed reviews of concept-based similarity (Pedersen et al. 2007; Pesquita et al. 2009). We discuss our hybrid methodology for finding similar concepts using knowledge-free, knowledge-based, and
ontological definitions in Chapter 3. However, despite the active research on this topic, these concept-level similarity methods have not been yet translated to most clinical summarization systems.

**Statement-Level Similarity**

A pervasive aspect of a patient record is the high level of statement redundancy across notes. For instance, two pathology reports for a given patient share many similar statements. Beyond the formulaic nature of documentation, statement-level redundancy also occurs because of copying and pasting from previous notes with some minimal editing of the copied statements.

In clinical summarization, there has been little work on this important aspect of similarity identification. Recently, a topic modeling approach was proposed to identify and control for such redundancy across patient notes (Cohen et al. 2014). In the general NLP community, identifying statement level similarity has been studied through the tasks of paraphrasing identification and textual entailment (Androutsopoulos 2010). Many of the methods in text summarization for identifying both unidirectional (textual entailment) and bidirectional (paraphrasing) similarity employ a hybrid of methods for word-level and concept-level redundancy such as string similarity, logic-based methods, and context-vector (Dagan et al. 2010).

Along with the need for higher order language similarity work in the clinical domain, there is an ongoing push to personalize similarity detection. It is well established that semantic similarity is context-dependent (Janowicz 2008) and a recent study suggests that redundancy be examined as a function of the patient’s previous history (Farri 2012). While identification of similar contexts based on the patient’s health is an ongoing direction of research (see Chapter 3), there is further work to be done in identifying context-specific similarity on higher-order semantic levels. Identifying similar words, concepts, and removing redundancy by patient-tailored information aggregation is an important direction for future EHR summarization methodology.
2.2.2 Organizing and reasoning over temporal events

Patients’ health evolves on many different time scales. Some health events such as pneumonia present themselves sporadically while chronic conditions like diabetes develop and worsen over a period of years. The importance of presenting clinical data in a time-dependent fashion has been recognized for a long time (Fries 1974; Cousins and Kahn 1991; Samal et al. 2011) however accurate temporal representation remains an open problem (Zhou and Hripcsak 2007; Sun, Rumshisky, and Uzuner 2013; Wu et al. 2014). Automatic creation of a clinical data timeline from textual and structured clinical records requires temporal event extraction, ordering, and reasoning.

Temporality is an active research area in the genre of news summarization given the quick news cycle and fast-paced evolution of news stories (Allan, Gupta, and Khandelwal 2001). However, news summarization research cannot always be readily translated into the health domain, as the challenges in health data are unique (Combi and Shahar 1997; Cios and Moore 2002). For example, different note types and specialties have different temporal relationships: pathology reports are often about one moment in time without reference to historical ailments whereas discharge summaries describe an entire inpatient hospital stay and instructions for future care. Styler et al. identified four complexities with extracting temporal information in clinical data: (i) diversity of time expressions, (ii) complexity of determining temporal relations among events, (iii) the difficulty of handling the temporal granularity of an event and (iv) general NLP issues (Styler, Bethard, and Finan 2014).

After the extraction of event time, there is a need for performing relative temporal ordering (Savova, Bethard, and Styler 2009). Event ordering is difficult in part due to inexact wording, but also because clinical knowledge is often needed to infer how long conditions may last (e.g., a diabetes diagnosis is often not discussed at every visit but a clinician is aware that diabetes is a chronic condition, not an intermittently reoccurring condition each time the “diabetes” term is mentioned or the diabetes ICD-9 code is recorded) (Hripcsak et al. 2009). Some recent work in event ordering includes the representation of temporal disease progression separately for each problem by Sonnenberg et al., an
approach they call “clinical threading” (Sonnenberg et al. 2012) and frame-like semantic representations with rule-based temporal extraction to arrange problems on a timeline (Jung et al. 2011). Raghavan et al. identify and temporally order cross-narrative medical events across documents in clinical text using weighted finite state transducers (Raghavan et al. 2014).

Reasoning and abstraction of extracted clinical events to highlight disease progressions and trends is critical for creating succinct clinical summaries. Abstractions of temporal data can include combining events within a certain time frame and performing interval-based abstractions such as combining multiple chemotherapy drug mentions into a chemotherapy regimen time span (Klimov, Shahar, and Taieb-Maimon 2010) or reasoning about the length of time that symptoms lasted and their relation to diagnosis (Zhou, Parsons, and Hripcsak 2008). The questions of which events should be combined and what an appropriate time frame is remain difficult and currently resolved by leveraging clinical knowledge and ontologies. Time-dependent clinical summarization is a continually evolving research area and there is opportunity for automatically identifying, accurately ordering, and performing reasoning over temporal clinical events.

2.2.3 Accounting for and interpreting missing data

Clinical records are sparse: documentation only occurs when a patient is seen by a clinician, thus clinical records miss the overwhelmingly large amount of observations about a patient across their lifetime. When summarizing sparse data, a critical complication is how to interpret and reason over the missing data (Wells et al. 2013). In some cases, missing data is not important and can safely be ignored by a summarization system (e.g., a patient has no change in health status in between visits). In other cases, the presence of missing data hints at a salient aspect about the patient that needs to be highlighted within the summary (e.g., patient is too sick to come to their visit). How to interpret and determine the salience of missing data is a challenge, and one not investigated thus far in clinical summarization.

In the field of general statistics, there are three types of missing data: Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing Not at Random (MNAR) (Little and Rubin
Most techniques for dealing with missing data assume that data are MCAR or MAR distributed, and include (i) variations of complete-case analysis, where only data with no missing values are used, (ii) single imputation, where missing data are imputed based on the values observed (using the mean, median, linear interpolation, etc.), and (iii) likelihood-based methods which compute maximum likelihood estimates for missing data (Enders 2006).

In the clinical domain, there is mounting evidence that most of the data are MNAR (Lin and Haug 2008; Pivovarov et al. 2014). For these data, the missingness is informative, meaning that there is an underlying reason that the data are missing but that this reason is simply unobserved. Some techniques that use informative missing data properties to infer properties about clinical data have been proposed. Chapter 4 of this dissertation presents our work on identifying patterns of missing data and using them to help infer healthcare status of patients. A common way of using missing data in the clinical domain has been to look at how long values should last based on recorded measurements or documentation frequency. For example, laboratory test measurements have been studied to gather appropriate imputation time (Hug 2006) and to infer health status features (Weber and Kohane 2013). Van Vleck studied duration and persistence of problems in notes (Van Vleck and Elhadad 2010) as a function of missing data, while Klann (Klann and Schadow 2010) and Perotte (Perotte and Hripcsak 2013) both studied the duration of ICD-9 codes. Klann estimated the durations for which each ICD-9 code remains valid and Perotte automatically classified ICD-9 codes into chronic and acute conditions. The modeling work that most explicitly demonstrates informativeness in missing data examined the accuracy of prediction models when: (i) ignoring missing data, (ii) interpolating missing data, or (iii) incorporating a missing data indicator, and reported that the missing data indicator method performed best (Lin and Haug 2008). To properly provide context and infer trend lines, as demonstrated by Poh and de Lusignan for kidney disease data (Poh and de Lusignan 2011b; Poh and de Lusignan 2011a), or to make predictions in clinical summaries it is critical to incorporate missing data literature and techniques into summarizer applications. The utility of modeling missing data explicitly is clear, however this conclusion is not being translated into clinical summarization research yet.
2.2.4 Reducing information to only the most salient

Salience identification has been heavily researched in the general domain text summarization literature. Early methods for identifying important topics relied on counts: frequency (Luhn 1958) and term frequency-inverse document frequency, which corrects for word specificity (Jones 1972). Other methods have focused on structure, such as document structure (Edmundson 1969) or syntax structure (Marcu 1997) to identify important phrases. Syntactic information gleaned from the input document can identify which parts of a sentence are salient and which may be safely removed from a summary (e.g., a relative clause). It is unclear, however, how these approaches translate to the clinical domain, where syntactic structure is unconventional. Using prior knowledge of the input document structure (e.g., biomedical papers have an introduction, followed by a methods section) to weigh the salience of information pieces based on where they are conveyed in the document is, however, promising in the clinical domain (yet not investigated thus far). Clinical notes follow a pre-specified structure; a diagnosis mention might be more relevant when conveyed in the past medical history than in the family history for instance. A different method for salience identification, still within the general domain summarization field, leverages discourse by considering sentences in input documents through a network, where lexical similarity between sentences is represented by the network edges. In this representation, salient sentences are the ones with the highest centralities (Radev, Jing, and Budzikowska 2000; Erkan and Radev 2004).

An alternative method for identifying relevant information relies on probabilistic modeling techniques such as Hidden Markov Models for identifying topics and topic changes in a set of documents (Barzilay and Lee 2004) or hierarchical Latent Dirichlet Allocation (LDA)-type models for identifying novel information with respect to older documents (Delort and Alfonseca 2012). Our work on employing Bayesian learning techniques to the construction of effective automated summaries is described in Chapter 5.

The one type of salience detection that has been explicitly studied in the clinical domain is based on cue phrases. Cue phrases are pieces of text that signify that what follows is likely to be important. For
example, “In conclusion” often precedes an important summarizing statement (Edmundson 1969). In clinical documentation, de Estrada et al. developed a system called Puya that found cue phrases indicating normality or abnormality in the physical exam sections of notes (de Estrada, Murphy, and Barnett 1997). Another way of detecting salience relies on n-gram language modeling to identify the most recent information in the record, under the assumption that the newest information is the most salient for the provider to see (Zhang, Pakhomov, and Melton 2012; Zhang 2014). A visualization prototype used this n-gram model to automatically highlight text that was found to be novel, drawing the provider’s attention to the new findings (Farri et al. 2012).

Defining salience in an operative fashion for automated summarization is an open question. In the general domain, there is evidence that humans sometimes disagree about what pieces of information are indeed salient, and that salience is often task-specific (Nenkova and Passonneau 2004). Similarly, in the clinical domain, determining what is important for a clinician is also probably quite task-specific. Nevertheless, it is safe to say that salience of elements in the patient record is related to capturing the health status of the patient and how it changes through time (Farri 2012; Suermondt et al. 1993). How to do so automatically, that is how to link textual and individual raw low-granularity observations to high-level clinical abstractions is one of the paramount challenge of informatics research. For instance, there has been little formal investigation of clinically specific markers of importance such as absolute change of a laboratory test value, the rate of change, the rate of mention of a particular concept, and other importance cues.

### 2.2.5 Using existing clinical knowledge

The informatics community has invested enormous effort into codifying clinical knowledge in a variety of terminologies and ontologies. This knowledge representation effort has been successful in helping efforts like phenotyping combine terminological knowledge, expert reasoning, and machine learning to create actionable disease definitions (Pathak, Kho, and Denny 2013). Similarly in summarization work, it is important to make use of these available clinical knowledge representations and
use them to generate rules and heuristics. Chapter 5 examines our efforts to incorporate coded clinical knowledge into an unsupervised, data-driven automated summarization engine.

Several holistic summarization efforts leveraged terminologies to identify concepts that are semantically related (e.g., medications that treat particular conditions) (Klann et al. 2013) or rules to determine salience (e.g., identify and highlight the salient results that are abnormal) (Plaisant et al. 1998). However, summarization engines built for particular diseases benefit most often from manually crafted rules and disease-specific knowledge bases as they enable tailored, task-dependent systems. The KNAVE-II application (Shahar et al. 2006), created for synthesis of bone marrow transplant patients, relies on an expert-maintained knowledge base for creating a semantic navigation system and concept abstraction. The Timeline system (Bui, Aberle, and Kangarloo 2007) is also built on a manually coded set of rules which identify salient concepts for different diseases, and perform temporal event reasoning. In addition, summaries that are setting and user specific often use expert-driven rules to ascertain which pieces of data should be shown at which time and to whom. Although the incorporation of clinical expertise into summarization is often a laborious process and sometimes only covers specific domains of expertise, it provides critical help in addressing some of the similarity, temporality and salience challenges. Of relevance to this review, we note that while existing summarizers rely on established knowledge resources, there is an active field of research to create these resources either by translating clinical expertise or acquiring the resources from data (Noy et al. 2009; Mortensen et al. 2012; Tao et al. 2013).

2.2.6 Deploying summarization tools into the clinic

The ultimate goal of any clinical summarization tool is implementation and usage by clinicians at the point of care. To date, however, there has been no widespread adoption of automated summarizers, especially for the large holistic temporal summarizers (Samal et al. 2011). Pervasive deployment is often hindered by the commercial EHRs systems that have been adopted across the country. Building real-time computational tools to work atop commercially built EHR systems is still a daunting task as these vendor
EHR systems are often not built to support interaction with outside applications. In addition, as the systems are closed off, dissemination of summaries across different hospitals and EHRs is a challenge as well. However, there is promising work with the i2b2-SMART platform that enables easier translation across institutions; researchers have developed a system to automatically link different data types across the EHR (mainly diseases and medications) and display a newly organized view of the patient record (Klann et al. 2013).

To create meaningful and practical summaries that assist clinicians during their point of care needs, summarizers need to provide real-time information with patient record updates immediately available in the summary. This is an especially difficult task when the summary tool works with natural language, as the processing must be completed quickly and accurately. Current work with distributed infrastructures, like Apache Hadoop, provides promising results for immediate summarization (Hirsch et al. 2014).

Another large barrier to translation of summarizer research into the clinical domain is rigorous evaluation. Hospitals often call for evidence of a useful summarizer before investing expensive resources into the implementation of the summarizer, but without adoption a summarizer is extremely difficult to evaluate. As is clear from Table 1, clinical summarization literature lacks standard evaluation metrics and there are very few extrinsic evaluations, a similar finding to a review of biomedical literature summarization by Mishra et al. (Mishra et al. 2014). Given the restriction of limited adoption, it is not clear on which dimensions clinical summarizers should be evaluated. Initially, in order to avoid costly development and implementations with marginal benefit, it is imperative to study the need for a summarizer tool, context of usage, and clinician workflow. However, without eventual implementation into clinical care, showing any process- or health-level outcomes is not possible and therefore how to perform useful evaluations remains unclear: should, for instance, summarization systems focus on accurate information extraction, facilitating information exploration (e.g., which concepts are most relevant to the clinician), or user-friendly designs? Although the rigorous user-interface and cognitive process evaluations that are necessary for creating new summarization systems often require deployment and study of actual use in practice, there exists guidance in the literature on cognitive aspects of clinical
reasoning that can inform summarization system creation. Prior work on general medical cognition (V. L. Patel, Arocha, and Kaufman 2001), clinical decision-making (Arocha, Wang, and Patel 2005; Kushniruk 2001), human-computer interaction for interface design (Patel and Kushniruk 1998; Jaspers et al. 2004; Thyvalikakath et al. 2014), handoff communication (Abraham et al. 2011; Abraham et al. 2014), clinical workflow analysis (Militello et al. 2014; Unertl et al. 2009) and some recent qualitative work specifically on clinical document synthesis which has identified common cognitive pathways for EHR document synthesis (Farri 2012) and patterns of EHR data access (Reichert et al. 2010) can guide the development of summarization systems. However, we emphasize that without actually studying the clinical context and manner in which clinicians use summarizers (either in the laboratory with prototype systems or in the clinic with deployed systems), it will be challenging to develop better evaluation strategies and better summarizers.