Doccurate: A Curation-Based Approach for Clinical Text Visualization

Nicole Sultanum, Devin Singh, Michael Brudno, and Fanny Chevalier

Abstract—Before seeing a patient, physicians seek to obtain an overview of the patient's medical history. Text plays a major role in this activity since it represents the bulk of the clinical documentation, but reviewing it quickly becomes onerous when patient charts grow too large. Text visualization methods have been widely explored to manage this large scale through visual summaries that rely on information retrieval algorithms to structure text and make it amenable to visualization. However, the integration with such automated approaches comes with a number of limitations, including significant error rates and the need for healthcare providers to fine-tune algorithms without expert knowledge of their inner mechanics. In addition, several of these approaches obscure or substitute the original clinical text and therefore fail to leverage qualitative and rhetorical flavours of the clinical notes. These drawbacks have limited the adoption of text visualization and other summarization technologies in clinical practice. In this work we present *Doccurate*, a novel system embodying a curation-based approach for the visualization of large clinical text datasets. Our approach offers automation auditing and customizability to physicians while also preserving and extensively linking to the original text. We discuss findings of a formal qualitative evaluation conducted with 6 domain experts, shedding light onto physicians' information needs, perceived strengths and limitations of automated tools, and the importance of customization while balancing efficiency. We also present use case scenarios to showcase *Doccurate*'s envisioned usage in practice.

Index Terms—Visual Curation, Clinical Text, Text Visualization, Medical Narrative

1 Introduction

Clinical practice is a complex activity that requires a grasp of both the medical issues afflicting a patient as well as contextual factors influencing their health such as family status, economic situation, and mental health. To encompass this complexity, text is the preferred mode of documentation (in contrast to more structured data formats), given its ability to preserve contextual richness while concisely communicating the "health narrative" – *i.e.*, the progression and interplay of medical events alongside contextual factors – as well as the flexibility to accommodate a physician's individual documentation needs [26].

On the other hand, this communication power begins to collapse as soon as the scale of text jumps from a few sentences to hundreds of protracted documents, making it increasingly difficult to obtain a sufficient overview of a patient. Given the pervasive time pressure surrounding medical practices, physicians are unable to conduct comprehensive reviews of the patient record [38] which may lead to patient safety issues [2]. To overcome this challenge, a significant body of research has looked into the task of obtaining overviews from clinical text using data visualization and text summarization (as reviewed by Rind *et al.* [35] and Pivovarov *et al.* [29]), which requires (often automated) pre-structuring of the clinical text for information retrieval.

While fully automated text processing has emerged as a powerful tool in the text visualization tool set, it also introduces a myriad of challenges. First, there is a need for error management to help identify, diagnose, and act on automation mistakes. Despite the advances in natural language processing (NLP), automated text processing is still an active field of research and yields substantial error rates (*e.g.*, about 80-85% precision and recall for state-of-the-art named entity recognition (NER) on biomedical text [13]). Second, physicians have unique information needs that depend on factors such as medical specialty, the patient's case, and the physician's own mental models.

- *Nicole Sultanum, Devin Singh and Michael Brudno are with the Hospital for Sick Children. Email: devin.singh@sickkids.ca*
- *Nicole Sultanum, Michael Brudno and Fanny Chevalier are with the University of Toronto. Email: {nicolebs,brudno}@cs.toronto.edu, fanny@dgp.toronto.edu.*

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

Medical information can be organized in a number of ways, including (but not limited to) time-oriented, source-oriented and problem-oriented structures [39], all supporting different and important clinical tasks. Given the variety of different structuring strategies, working with pre-boxed collections of categories and parameterizations offers only partial solutions to the aforementioned problems. Even when automation controls are exposed for user customization, physicians often lack the time or technical expertise to diagnose errors and fine tune automation effectively [38]. Ultimately, the adoption of such technologies in real world clinical practice is hindered by a lack of flexibility and transparency of automated components. We argue that these considerations should be seen as an integral part of the user experience and should be fully incorporated into the design.

In this paper, we propose a more verifiable and customizable approach to leverage automated text processing and unstructured medical knowledge in clinical text visualization. Our approach is inspired by past work in visual curation [11, 28] — *i.e.,* user-in-the-loop iterative refinement of automated processes, aided by visualization — and proposes the creation of *reusable and physician-defined thematic filters* that leverage medical taxonomies to cut through the clutter and aggregate related information in pre-processed tagged text.

We also support automation transparency and verification by communicating key details of automated output. We present *Doccurate*, a clinical text visualization prototype that applies these principles to support the visualization of large text patient records, and discuss findings of a formal qualitative evaluation with 6 medical practitioners. Results cover physicians' experiences and opinions on customization, trust in automation, and the use of data visualization for text, indicating that our approach has potential to help ease the adoption of automation into clinical workflows. Based on study findings, we also present use case scenarios we developed with a domain expert to illustrate clinical tasks that we envision *Doccurate* could be particularly helpful for.

2 Clinical Text and Curation

Ideally, physicians should be able to retrieve a sufficiently complete and accurate picture from a patient *chart*, *i.e.,* the collection of notes documenting a patient's history, in just a few minutes before a consultation or emergency intervention. In this section, we discuss the shortcomings that make this task particularly challenging. We then review prior efforts in leveraging automation to provide physicians with meaningful information extracted from large text-based clinical records while highlighting limitations of these methods and opportunities for improvement. We discuss orthogonal research in visualization aimed

at providing visual summaries of large text corpora (with varied levels of automation) in other domains. Finally, we introduce our proposed approach aiming to reconcile the power of automation and physician-defined customization to support individual information needs.

2.1 Challenges of Clinical Text

Prior to seeing a patient, physicians seek to obtain sufficient overview of the patient's history to guide the consultation by reviewing the patient chart [4,25]. They focus on trying to capture the key medical issues into a *problems list* and put together a *proto-narrative* (mental or drafted) of the patient's history encompassing known and probable causes to be then reassessed with the patient [25]. However, because clinicians have limited time to achieve sufficient overview [4], they adopt strategies to optimize their study, such as skimming notes for medical issues, focusing on recent information, and directly asking the patient for details [38]. These strategies usually cover most crucial information, but provide limited coverage of details buried in the patient record or the patient's memory that may be relevant. This represents a patient safety concern [2], and risks escalate further when physicians review longer patient charts with hundreds of documents [29]. Skimming notes, in particular, is further complicated by non-standardized formatting, content redundancy and information scatter.

2.2 Physicians & Automation

In view of this information overload, automated strategies have been proposed to facilitate physician access to text content in patient charts, including text summarization and visual summaries [29]. However, these strategies require significant oversight of text structuring processes before their output can be conveyed to the physician.

Early seminal work such as Lifelines [30–32] and Powsner & Tufte's Graphical Summary [33] required content to be manually extracted from text before it could be visualized. In later research, automation was gradually introduced to reorganize documents in a patient chart according to medical problems [5], generate chronicles from events extracted from clinical text and regenerate standardized text summaries [14, 15], extract medical issues to generate problem-based timelines [14, 15, 18] and problem-based word clouds [17], and leverage disease graphical models to infer causality [18]. These works either operate on the assumption that the structured data is correct or acknowledge the need of a human annotator to oversee the process. MedStory [38], our previous investigation to leverage medical text alongside visualization, relies on named entity recognition (NER) to identify concepts in the text and provides extensive linking to the originating documents, but does not offer mechanisms to correct or improve the outcomes of the automation.

Arguably, the lack of comprehensive validation of automated outcomes is a significant obstacle to NLP adoption [29], which in turn deters *in the wild* evaluations. Most validation efforts so far have focused on assessing accuracy of automated tools and few works have looked into how NLP-based technologies can be better integrated into clinical practice. A few notable exceptions include expert validation of automated output, such as a recent study comparing physicians' perceived usefulness of manually created problem lists, automatic lists generated by IBM Watson, and a physician's own list [7]. They found physicians rated the automatic list higher than the generic manual problem list, but still consistently rated their own lists higher than the automated ones; this may suggest a need to leverage some physician agency and sense of ownership over fully automated approaches.

Another trend is investigating human-in-the-loop strategies to understand and customize automation tools. An evaluation of NLPReVis [40], a system designed for physicians to refine their own NLP models, found that participants were able to improve the models from a small initial training set, but still involved a lengthy hour long process of dedicated work that did not integrate into a physician's standard chart review workflow. Overall, a number of challenges that deter NLP adoption in clinical practice still persist, including lack of generalizability (requiring NLP models to be rebuilt and retrained for different tasks), lack of customization features for end users, and lack of understanding of how these NLP tools can effectively fit into user workflows [6].

2.3 Text Visualization in Other Domains

Text visualization is a complex and multifaceted problem with a significant research history. Of relevance to our target problem, we highlight works that aim to (a) summarize large temporally-oriented text collections, (b) visually support text navigation, and (c) mediate use of automated text processing components.

Relating to text summarization, word clouds are a popular visualization strategy that have been frequently used with streamgraphs to visualize overarching themes in the text along their evolving prevalence over time. After ThemeRiver [16] pioneered the representation, several others leveraged it to summarize microblog streams [3, 8], email communication [24], news media [9] and academic documents [23]. To better emphasize individual theme prevalence, some works evolved from the original stacked layout and instead separate individual streams into dedicated tracks [9,23] . Although useful at providing high-level *at-a-glance* overview, word cloud-based approaches are arguably limited for tasks that requires deeper narrative understanding of the underlying text, *e.g.*, relationships between entities, motives, and rationale.

Fewer works undertake the problem of navigating text collections to support low-level reading tasks such as highlighting locations of interest and supporting direct navigation to any segment of text. Notable instances include VarifocalReader [22] that provides a multi-level visual index of a large text document to conciliate overview and details in a single view, and Belmonte's Twitter visualization of Obama's 2014 State of the Union (SOTU) address [3] that complements a speech transcript with graphical views showing concurrent Twitter activity at each segment of the speech. While these works inform valuable strategies to support text through visualization, none have been designed to consider the unique characteristics of the medical domain.

Most of the works above employ some form of automated text processing (*e.g.*, topic models [24]) but do not provide mechanisms to verify automation performance. To tackle this issue, recent works reconsider traditional visualization workflows that assume the correctness of automated text processing to instead acknowledge the limitations of automation and enable user interventions over automated input as an integral part of the process. TimelineCurator [11] proposes a semi-automatic approach to creating timelines that leverages an average performing temporal extraction algorithm to identify events, and offers tools to both verify the output and to curate events to be shown. ConceptVector [28], on the other hand, supports the creation of lexicons via word embedding techniques that are driven by user-defined keyword concepts. Both strategies propose to seamlessly integrate into existing workflows—*i.e.*, creating timelines [11] and building lexicons while analyzing document sets [28]—and include an iterative refinement loop to progressively improve the underlying semantic structures representing knowledge from text. Naturally, none of these applications are suited or relevant for medical applications as-is, and although they are much faster than manual labour they still require significant time investment before user efforts are rewarded.

2.4 Our approach: Visual Curation meets Medical NLP

We take inspiration from TimelineCurator [11], ConceptVector [28], and their semi-automatic visual curation models to inform the desirable characteristics of our proposed approach. We argue that the use of NLP in clinical practice should be framed in a more physician-centered manner that (a) allows for a continued adaptation of automation to personal and evolving information needs, (b) fosters a more verifiable and adequate reliance on automated tools, (c) integrates into existing clinical workflows, and (d) supports efficiency. To fulfill this vision, we propose a new interaction model supported by data visualization to allow physicians more agency over automated processes, by:

Empowering user oversight. Automated NLP tools are efficient at extracting structure from large collections of plain text, but also fail often. Both the visualization and the original underlying text support users in detecting and correcting errors themselves, providing *autonomy* and encouraging *proper, balanced reliance* on automation.

User-defined structures. The equivalent of an *authoring tool* for physicians would be to provide the ability to define customizable facets

of the medical narrative to represent perspectives the physician is interested in conveying at any given time. Flexibility is crucial to support their varied and complex information needs.

Considering the unique requirements of the medical domain, we also account for the following complementary aspects:

Integration into clinical workflows. Instead of seeing curation as a standalone activity separate from the patient care process (*e.g.*, retraining NLP models on one's spare time), curation should seamlessly integrate into daily clinical workflows and activities such as patient chart reviews.

Efficiency and Reusability. If curation activities are to be seen as integral and continued effort within clinical workflow, its burden should be minimized and it should ideally help save time in the long run. Individual curation actions should be efficient and leverage past curation efforts as much as possible.

3 *Doccurate*

We designed *Doccurate* as a data visualization tool to support overview of large patient charts. In addition to the overarching goals defined in the previous section, we also leverage design considerations from prior work to support clinical text overview (the result of formative and summative assessments with 22 physicians) [38] and physician interaction with NLP automation [40]. Our design goals are as follows:

(G1) Preserve the original text. Apart from text being a familiar medium, clinical notes are regarded as "medical evidence" [38] and should be available in their original form (instead of being replaced with other forms of summarization or only in smaller parts).

(G2) Foster suitable trust over automated output. Physicians should be offered means to assess the extent to which they can rely on automation. We support this by selectively exposing internal aspects of the automation and providing extensive linkage to the original text.

(G3) Provide information in different levels of granularity. Support both information in-a-glance for high-level overview as well as tools for content exploration and information seeking.

(G4) Support user-driven customization. Physicians' information requirements and personal preferences are diverse, so it is important to support some level of interface tailoring to suit these needs.

(G5) Convey time and progression. Understanding how medical problems evolved over time is an important aspect of clinical overview and, therefore, temporal references should be thoroughly supported.

(G6) Support content faceting. The patient chart is structured like a chronicle, with medical notes representing localized snapshots of the patient's health status and thus may encompass information on several issues. When the focus is on one particular medical problem, physicians should have the ability to cross-cut the record to get a more coherent picture of that issue.

3.1 Filter Collections (FCs) as Curated Content Facets

Following our established emphasis on customization and reusability, we propose **Filter Collections** (or **FCs** for short) as our curation building blocks, consisting of *physician-defined semantic filters to create faceted views of clinical content*. FCs leverage text tagged by automated entity recognition algorithms and their mappings to medical taxonomies such as SNOMED-CT [19]. These taxonomies provide thorough encodings of domain specific knowledge that can be used to improve text processing outcomes, *e.g.*, topic models of medical record data [12]. In this work we make use of the hierarchical structure of taxonomies to define *semantic scopes of interest*. As follows, an FC basically defines *a set of related umbrella concepts pertaining to a shared meaning*; by leveraging parent-child relationships in taxonomies, we can extend this meaning to a large collection of children concepts covered under the umbrella and rapidly group related concepts together.

The detailed nature of these taxonomies and fine-grained hierarchical breakdown confers considerable expressive power to these umbrella scopes and is flexible enough to represent a myriad of clinical scenarios. For example, consider the SNOMED-CT hierarchy excerpt in Fig. 1. A general practitioner may be interested in grouping all cardiovascular issues together, and would include the top level concept `Disorder of cardiovascular system` into a *"Cardiovascular"* related FC. A cardiologist, on the other hand, would benefit from a more detailed breakdown of the different types of cardiovascular disorders, *e.g.*, by creating a dedicated FC for heart diseases separate from thrombotic or blood-vessel related disorders. Encompassing several such concepts into one FC allows for combinations that transcend the original structure of the taxonomy, for instance by combining mentions of thrombotic diseases (nested under `Disorders`) and anti-platelet agents (nested under `Substances`). In summary, provided there is an efficient mechanism to access and navigate the taxonomy, we expect physicians to be able to quickly create sophisticated FCs using a handful of carefully selected umbrella concepts. Once an FC is created, it may also be reused for other patients and could help save time in the long run.

A downside of this approach is that it requires text to be tagged before it can be captured by an FC; the entity recognition algorithm may fail to tag a valid term, or a valid concept may not yet exist in the leveraged taxonomy. We address both issues by (a) allowing users to tag parts of the text with valid concepts, and (b) allowing FCs to also include keywords to complement umbrella concepts.

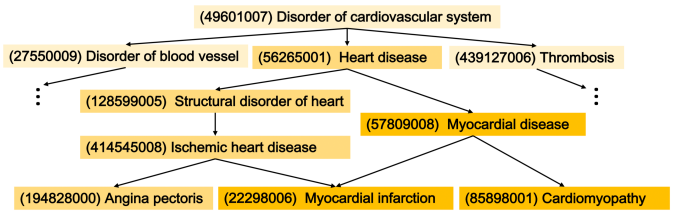


Fig. 1. An excerpt of a SNOMED-CT concept hierarchy.

3.2 Text Pre-processing

We preprocessed patient chart documents for entity recognition using Apache cTAKES [10], an open-source named entity recognition (NER) tool for clinical text; past benchmarking efforts [37] indicate good but not excellent performance (F-scores between 0.715 and 0.824), making it an ideal testbed to assess how our proposed approach supports physicians interfacing with error-prone NLP pipelines. Tagged terms representing medically relevant problems are associated to a UMLS [27] concept (a metathesaurus that unifies several medical taxonomies under common identifiers) and the associated SNOMED-CT concepts. While UMLS provides a convenient unique identifier that unifies various equivalent SNOMED-CT concepts, it is not cleanly arranged as a semantic hierarchy like SNOMED-CT, and we use the latter to specify FCs; SNOMED-CT also features a fine-grained hierarchical structure with over 300,000 terms distributed across over 15 levels, allowing for greater semantic flexibility.

3.3 Interface Design

Doccurate's interface is divided into 4 panels (Fig. 2): (A) the **Control Panel**, listing patient demographics, document filters, and filter collections (FCs); (B) the **Timeline**, providing an overview of FC content; (C) the **Text Panel**, listing the documents in the patient chart; and (D) the **Curation Panel**, to create and edit FCs. We make reference to our corresponding design goals ***(G1*** through ***G6)*** as features are presented.

The **Timeline** provides a time-oriented overview of lists of tagged content encompassed by FCs ***(G5 G6)***, organized in three levels of detail ***(G3)*** (Fig. 4). At the 1st level, all FCs are visible, along with a subset of their most frequent terms. The 2nd level appears when an FC is selected and displays all tags encompassed by that FC. The 3rd level appears when a tag is selected, displaying short snippets around the selected item and enabling access to a complete list of all snippets for that tag (by clicking ⋯). Clicking a timeline snippet redirects the physician to the corresponding passage in the **Text Panel** ***(G1)***. At all levels, content is organized into tracks encompassing (a) a frequency stream, (b) representative samples of tagged text, and (c) a track title (either the FC title on the 1st level or the corresponding tag description on the 2nd and 3rd levels) ***(G6)***. By perusing both the tag description and the encompassed text samples (Fig. 2(B.3)), the physician may

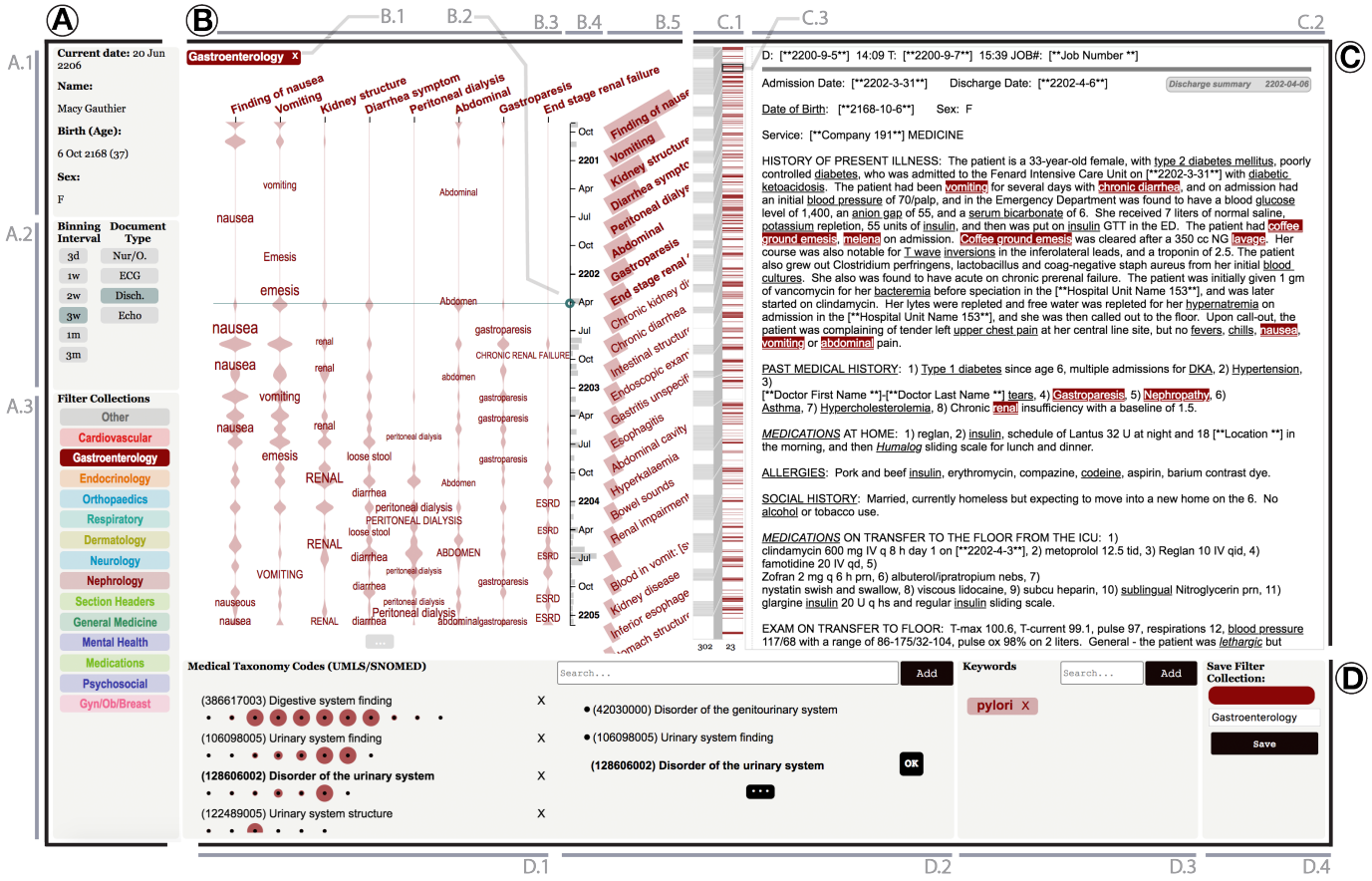


Fig. 2. *Doccurate*'s interface: **(A) Control Panel** with *Demographics* (A.1), options to adjust the Timeline's *binning interval* and the *visible documents types* in the Text Panel (A.2), and the *complete list of FCs* sorted by frequency (A.3); **(B) Timeline** with *breadcrumbs* indicating level of detail (B.1), scrollable list of *items encompassed by the current level* (B.3); items *tracks* with frequency streams and representative labels (B.5), a dark *line marker*indicating time of current visible document (B.2) and *time axis*, featuring a *document histogram* (B.4); **(C) Text Panel**, with the double *text overview bar* and respective*document counts* at the bottom (C.1), and all visible chart documents (C.2); **(D) Curation Panel** for a selected FC, with subpanels for the *list of codes* (D.1), *hierarchy adjustment* for a selected code (D.2), *list of keywords* (D.3) and *colour/title* editing (D.4).

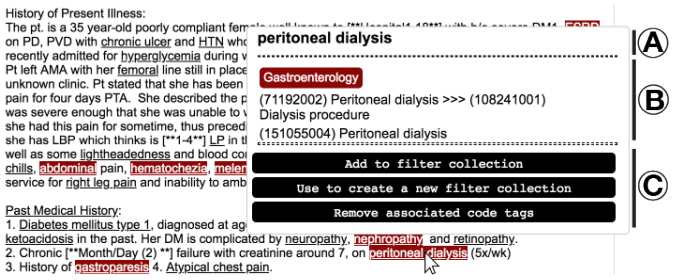
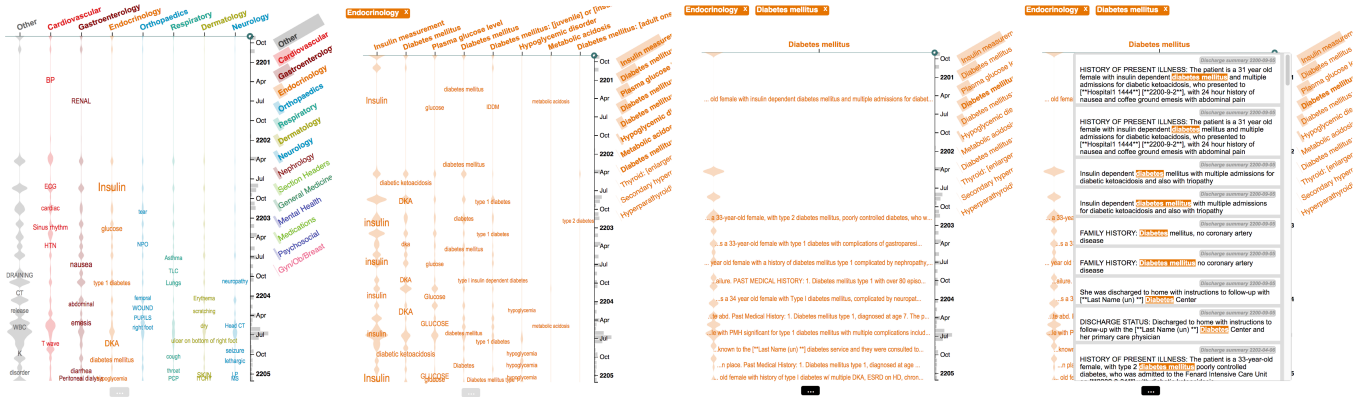


Fig. 3. The curation information panel, displayed when text is selected in the **Text Panel**. It lists (A) the selected term, (B) its associated FCs and concept codes, and (C) valid curation actions for the term.

be able to identify tagging mistakes on the go ***(G2)***. All items in a level are listed on the right (Fig. 2(B.5)), sorted by frequency (most frequent first), and can be scrolled through to access other items in that level (visible items are **bolded**). Behind the time axis (Fig. 2(B.4)), a document histogram shows the number of documents at specified time intervals. The interval can be adjusted in the **Control Panel** (Fig. 2(A)) and updates both the document histogram and the track streams.

The **Text Panel** contains all chart documents, concatenated in a long scrolling list and chronologically sorted ***(G1, G5)***. A text overview bar (Fig. 2(C.1)) indicates the length and position of chart documents (and current viewport, Fig. 2(C.3)) relative to the entire chart. The **Timeline** also features horizontal markers indicating the creation date of current visible documents (Fig. 2(B.2)) ***(G5)***. From the **Control Panel**, specific document types can be selected for viewing ***(G6)***: on Fig. 2(A.2), only discharge summaries (`Disch.`) are selected for viewing, which also updates the right half of the text overview bar to list only the visible documents (Fig. 2(C.1)) and document counts for total and filtered documents (at the bottom). When an FC is active, the text overview bar displays the location of all tagged text in that FC as coloured lines to convey density and highlight tag clusters; users can navigate to any part of the visible chart (*e.g.*, to inspect tag clusters) by clicking on the corresponding location in the text overview bar. Tagged text is also highlighted in the document view (Fig. 2(C.2)).

Users can perform a number of curation-related activities directly from the text. If a tagged term is clicked, an information panel appears (Fig. 3) indicating FCs that encompass that term, all SNOMED-CT concepts associated to it, and parent concepts that triggered FC inclusion (following ">>>"). This information allows the troubleshooting of wrong or dubious FC inclusion, either due to incorrect code assignment or parent-concept over/under scoping ***(G2)***. Relevant curation actions are also available, including *adding the selected term to an existing FC*, *creating a new FC including that term*, and *removing assigned codes* (Fig. 3(C)). For non-tagged text, drag-selecting a text snippet will display the same information panel, with for the first two FC curation actions being performed with the term as a *keyword* instead, and the third action changed to *add* a concept code instead of removing. This way, curation actions and error corrections (that propagate to all mentions of the same concept/term) can be performed in context while the physician is reading the note ***(G1 G2 G4)***.



(a) (b) (c) (d)

Fig. 4. The Timeline view for *Macy Gauthier*, as Dr. J reviews the record for the first time (Section 6, *Use Case Scenarios*). It illustrates the three levels of detail (a-c) and the snippets panel which can be activated from the 3rd level (d). Label placement on timeline tracks corresponds roughly (but not exactly) to the date of the document containing that label, as a force directed layout attempts to repositions labels to minimize overlap.

When an FC is active (either selected from the 1st **Timeline** level or from the FC list in the **Control Panel**), the **Curation Panel** is updated to display its list of encompassed parent concept codes (Fig. 2(D.1)) and keywords (Fig. 2(D.3)) ***(G4)***. Concept codes also feature prevalence by hierarchy level: each inner circle (black) is a child level, and the size of the outer circle (coloured) conveys frequency of child terms encompassed at that level). Existing codes and keywords can be removed, and new ones can be directly added from the **Curation Panel** via text input. In addition, concept codes can be edited for scope by replacement with more general (parent) or more specific (children) concepts: the panel on Fig. 2(D.2) allows for hierarchy adjustment of selected concepts. Users can also edit the title and assign a color to the FC from a palette of 16 hues based on *Google Charts* colors (Fig. 2(D.4)); we chose strong saturated colors that still allowed for reasonable hue discrimination under reduced opacity.

Finally, *Doccurate* features a special FC called **"Other"** which lists tagged terms that have not been covered by any existing FC. It provides a thoroughness "sanity check" to improve existing FCs (particularly at early curation stages) and guides the creation of new ones ***(G4)***.

	Tasks			Role	Experience
	1 (15min)	**2** (10min)	**3** (10 min)		
P1	**Patient 1**	**Patient 2**	**Patient 2**	Physician	5 years
P2	**Patient 1**	**Patient 3**	**Patient 3**	Resident	1st year
P3	**Patient 3**	**Patient 2**	**Patient 2**	Resident	1st year
P4	**Patient 3**	**Patient 1**	**Patient 2**	Resident	1st year
P5	**Patient 1**	**Patient 3**	**Patient 2**	Resident	1st year
P6	**Patient 3**	**Patient 1**	**Patient 2**	Resident	2nd year

Fig. 5. Variations across participant sessions (Patient 1: yellow, 2: blue, 3: green), including reviewed records and role/experience.

4 Evaluation Study

We designed *Doccurate* with the ultimate goal of enabling physicians to adequately leverage automation to support more complete overviews of large patient charts. To assess the potential and limitations of our visual curation approach we conducted an evaluation with domain experts carrying out time-constrained overview tasks over a number of patient records using our research prototype. The goal of the study was to expose physicians to a number of chart review scenarios similar to what they would encounter in real practice while using *Doccurate*.

Reflecting upon the aforementioned challenges, our study was constructed to address the following research questions:

1. **Curation workflow:** How do physicians adopt (and adapt to) the new features into their chart review workflow?
2. **Curation structures:** How do different physicians choose to customize their FCs? What levels of detail are useful? Would leveraging a predefined group of FCs be a feasible solution?
3. **Automation transparency**: Is the FC curation model understandable? Are physicians able to appropriately gauge trust in automation and adjust their expectations accordingly?
4. **Efficiency**: What is the impact of curation on perceived workflow efficiency? Do physicians feel compelled to engage in low-level curation? Do physicians feel this FC model could help them save time in the long run?
5. **Comprehensiveness**: How does the FC model contribute to obtaining a more complete overview of the patient?

4.1 Participants

We recruited 5 residents and 1 physician (6 total) in General Practice (GP) to participate in the study, from 7 different Canadian healthcare institutions (2 participants had double affiliation, and 2 shared the same institution); 4 participants were in their 1st residency year, 1 in the 2nd, and the physician had 5 years of practice. We chose to recruit GPs due to their broader range of medical interests. All participants used electronic medical records in their practice on a daily basis. Reported preparation time before seeing a patient in clinic spanned 2-15min (avg. 7min) for returning patients and 5min-1h (avg. 20min) for new patients. Participants were given a $40 gift card for their participation. From here on, we refer to individual participant sessions as P1-P6.

4.2 Method

We adopted an iterative qualitative approach with predefined tasks. Study sessions were approximately 2h long and spanned three scenarios involving obtaining a patient overview from a medical record. After 3 sessions we made usability improvements (mainly bug fixing and the addition of the snippets list, Fig. 4(d)) plus a few adjustments to the session structure to encompass a wider variety of scenarios. Three distinct patient records were used in this study (with approximately 300 documents each), retrieved from MIMIC-III [20], a large database of medical data for critical care. Each session encompassed:

1. An extensive **walkthrough** of Doccurate (30-40min);
2. **Task 1 (FC Creation)**: creating FCs from scratch, while reviewing a patient record A (15min);
3. **Task 2 (FC Reuse)**: reusing FCs created in Task 1 on a new patient record B (10min);
4. **Task 3 (FC Presets)**: using a pre-curated and more complete set of FCs on either record B (P1-P3) or a new record C (P4-P6) (10min);
5. a **closing interview**.

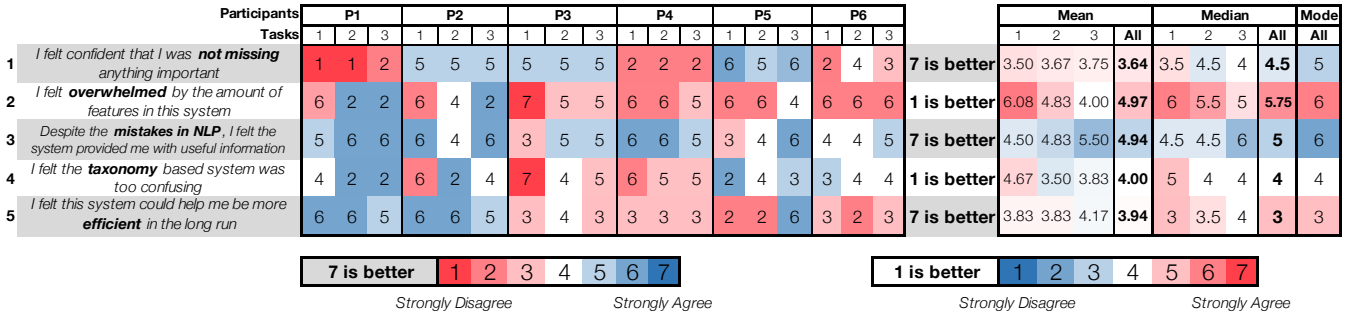


Fig. 6. Questionnaire results, per participant x task (the same questionnaire was filled after every task, to capture any emerging trends). For all questions, the scale ranged from 1 (*Strongly Disagree*) to 7 (*Strongly Agree*), but the meanings of the questions alternate in polarity. To facilitate analysis we adjust for polarity using color, which we map to blue for positive outcomes and red for negative ones.

Overall, tasks were tailored to explore both filter collection *creation* and *reuse*. During each task, participants were asked to fill a summary sheet with bullet point notes encompassing relevant *medical events*, diagnosed *medical problems* and relevant information on *social history*, to be filled as the patient history was unveiled. A short task-specific interview and standard questionnaire followed each task.

The adjustments made throughout the study include:

Assessing alternative scenarios. The initial study setup (P1-P3) was designed so that a reference could be established between Tasks 2 and 3 that would give participants the chance to compare their own FCs versus a more complete set created by someone else. For P4-P6, we felt the need to assess impressions around the pre-curated set of filters to support overview of a completely new patient record, and therefore introduced a third unique record on Task 3 for P4-P6; we used the same record in Task 3 for the latter sessions (P4-P6) to further facilitate intra-participant comparison, while still alternating records for Tasks 1 and 2. A breakdown of the different records used per participant is provided in Fig. 5, along with participant's background information.

Removing potential biases. During the walkthrough for P1-P2, we briefly presented examples (from the presets) to illustrate what is possible with FCs. However, we found that some FCs created in these sessions were similar to our examples and wondered if the patterns would hold without prompting, thus omitting examples for P3-P6.

Study session duration. Given the time-consuming nature of this study and participants' tight schedules, we made adjustments for brevity: (a) we removed a 5-10min pre-task that allowed P1-P3 some time to review Task 1's record A prior to curation, and instead, used record A for the walkthrough (instead of a separate test record for P1-P3) to compensate for the extended exposure while reducing overall session time. We also found that 10min was a short interval to create a comprehensive set of FCs, and chose to alternate between two of the more similar records for Tasks 1 and 2 to increase chances of overlap and reuse taking place within such a limited time frame.

These diversifications were effected so we could increase our chances of uncovering additional facets of the problem, which we believe has ultimately enriched our analysis.

4.3 Findings

We present findings and discuss implications pertaining to our previously established research questions.

4.3.1 Creating and Reusing FCs: Considerations on Structure

For Tasks 1 and 2, participants were asked to create filter collections from scratch. For these Tasks, we found that physician-created filters tended to be more specific than expected (especially those not prompted with examples), focusing more on symptoms and syndromes (*e.g.*, 'Diabetes', 'Hypertension', 'Seizure') alongside other more general ones on body systems or medical specialties (*e.g.*, 'Cardiac', 'Neuro', 'Psychosocial') (Fig. 7). This is a reflection of physicians' typical information retrieval workflow as they go over the record and seek to retrieve a list of medical problems for that patient [38]. Despite all but P3 reporting a reasonable grasp of the taxonomy-driven tagging to enable FC creation, P1 and P2 stated they needed more time to master it, P3 found it inefficient and P6 did not use the taxonomy hierarchy. Questionnaire ratings on the topic are also mixed (Fig. 6, item 4).

P1	P2	P3	P4	P5	P6
Diabetes	Psych	Renal Disease	GI	Back	Diabetes
Cardiac	Diabetes	Diabetes	Neuro	Mental Health	Cardiac
Pain	Internal Medicine	Asthma	Cardiac	Diabetes	Asthma
Psychosocial	Cardiology	Alcohol use disorder	Diabetes	HTN	Neuro
Respiratory	Collection 1	Abuse	Renal	MSK	Mental Health
			Social	Seizure	
			Family History		
			Surgery		

Fig. 7. Final list of curated FCs (end of Task 2) with user assigned colors; grey (default color) is auto assigned to new FCs, hence its prevalence.

In summary, not all participants appreciated the process of creating filters from scratch, and there was some confusion around how to best structure it. This may have partly contributed to the feeling of being overwhelmed (Fig. 6, item 2), but would likely improve with time given the positive trend seen from Tasks 1 to 3 for this criterion.

For Task 3, we presented a preset list of FCs, designed with a more general, specialty-based structure (FC list in Fig. 8). Feedback was more positive this time around, and most participants appreciated the information retrieved by the preset FCs (Fig. 6, item 3, Task 3). Participants also appreciated having a base to work on instead of creating filters from scratch, and several found the more general structures to be useful: P5 commented that this initial set was particularly valuable, to showcase the capabilities of the system; P1 found the presets could be a helpful starting point to explore records of undiagnosed patients. On the other hand, most participants acknowledged that applicability of the presets is limited and appreciated the ability for further customization (P1 P2 P4 P5). Reflecting this thought, several performed some form of curation to the presets during Task 3 (P2 P5 P6).

4.3.2 Workflow: visualization and text

Participants were asked to create and use FCs while reviewing the chart and taking notes, but were otherwise free to use the available tools however they wished. For Tasks 1 and 2, we found that most curation actions sprung from the text while participants were reviewing the chart (*e.g.*, pausing reading to add a term to a new or an existing FC) instead of the Curation panel. This may be partly due to the lack of familiarity with the taxonomy itself—all participants but P6 had not heard of SNOMED-CT prior to the study—but is most likely due to the convenience factor of performing curation while perusing the record. The Curation panel was actually deemed useful by a few participants (P5 P6) for its keyword search, as it indicates the frequency of keywords in the drop down menu.

On Task 3, and with a more comprehensive collection of FCs at hand, the Timeline was frequently used by all participants. They followed a similar pattern that encompassed selecting FCs from the top of the list (*i.e.*, starting with the most frequent FCs), inspecting the items found, drilling down further and either redirecting to the original mention

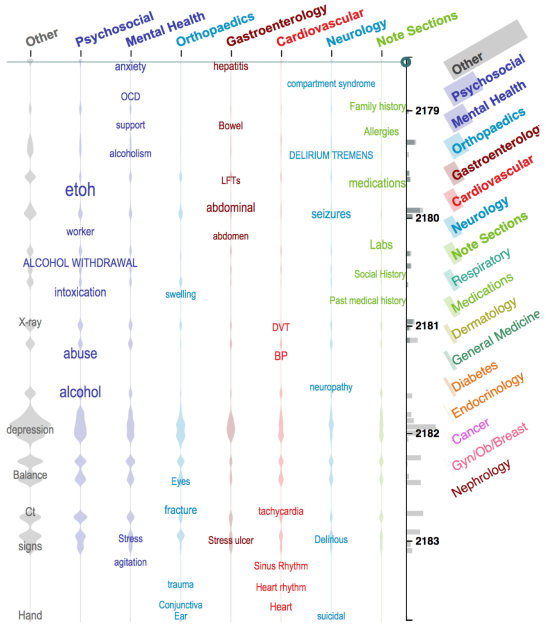


Fig. 8. An overview of a complex patient with preset FCs. Highlights on *"alcohol"*, *"intoxication"*, *"alcohol withdrawal"* and *"delirium tremens"* provide consistent hints of a significant and adverse history of alcohol abuse, with multiple episodes of withdrawal and related symptoms.

in the text when finding something worth investigating or moving on to the next FC. At times, the snippets and 2nd level FC view were sufficient to assess information, which saved time. In between FC transitions and after redirections, participants also spent time reviewing the information around the redirection point, which often prompted serendipitous discoveries. Compared to the traditional chart reading workflow, participants obtained a wider coverage of the entire chart, as redirections were not necessarily pointed at the most recent note. While there are a few confounding factors at play (the novelty effect and our initial request to use the FCs), given the positive findings discussed earlier on the use of preset FCs, it is possible that this workflow shift would also happen in a real world scenario.

4.3.3 The patient narrative, and the big picture

Physicians strive to grasp the patient's *medical history*, that is, the progression of medical events and factors that led to the current patient status. All participants found that the Timeline's keyword approach was limited in terms of conveying the narrative, but could still convey a broad picture of the general issues and the severity of the case: *"this [timeline overview] actually (provided) me with a snapshot of things that are going on, which was a lot more useful than [drilling down details]; (..) this was a little bit more useful, to get just a high level overview of what the actual history looks like"* (P6) . The extent of the Timeline's contribution varied across participants, however. For instance, P4 found that the picture was too vague and still needed to *"read a discharge summary in its entirety"* for a good grasp of the story , while P5 was able to put a proto-story together *"right off the bat"* simply from the co-occurrence of related terms and and how big (*i.e.*, how frequent) they appeared in the Timeline: *e.g.*, noting *"polysubstance abuse"* and particularly *"alcohol abuse"*, followed by related events such as *"alcohol withdrawal"*, *"delirium tremens"* and *"seizures"* indicated that this is a complex patient with a long and severe history of alcohol abuse that triggered extreme and recurrent symptoms (as illustrated on Fig. 8).

Supporting at-a-glance content is perhaps the most appreciated aspect of *Doccurate*. Among the preferred features of the system, participants cited the Timeline Overview page (P2 P5 P6) and the sorting of FCs and issues by frequency (P3 P4). Some participants who favoured reading also appreciated the ability to complement or confirm knowledge, via the preset FCs (P4) and the "Other" FC (P1). This speaks to the value of providing awareness and coverage of the entire record that can complement the reading, in order to mitigate the recency bias of focusing only on the last few notes.

4.3.4 Efficiency, Trust, and the limits of automation

By the end of Task 3, all participants had identified at least a few classification mistakes, encompassing:

- Terms incorrectly tagged by the NER engine, *i.e.*, assigned wrong codes. Common occurrence for acronyms, *e.g.*,'PCP' tagged as *"Pulmonary pneumocystosis"* but meaning *"primary care provider"*
- Terms that are correctly tagged but uninformative overall for that patient or task, *e.g., "Eye"*, *"Back"*, *"CT scan"*.
- Terms that appear be noteworthy but are later found to be negative findings, *e.g., "chest pain"* versus *"no chest pain"*.
- Terms that are tagged with two similar codes that are both captured under the same FC and thus generate redundant appearances that clutter the timeline view, *e.g.*, *"Diabetes Mellitus"* and *"Diabetes Mellitus:[juvenile] or [insulin dependent]"* (Fig. 4)
- Compound terms that are tagged separately, and therefore less precise; *e.g., "back"* and *"pain"*, vs. *"back pain"*
- Terms that are inherently ambiguous; e.g., *"abuse"* may stand for both *"victim of abuse"* and *"substance abuse"*.

Some of the above issues pertain to previously documented limitations of cTAKES, namely, disambiguation of similar concepts and dealing with compound terms [37]. These mistakes led to participants being overall more guarded about the content in FCs and caused added frustration, since participants felt the need to spend time double checking that the occurrences in the timeline were correct; some participants also wished for a shorter path to verification (P4 P5). Most participants stated that the system overall did not make them faster in their assessment (P3 P4 P5 P6), that they *"never trust the system 100%"* (P2) and that they *"still have to read the note"* (P4). This is also reflected in the somewhat poor scores for efficiency in the long run (Fig. 6, item 5). On the other hand, this frustration reflects a healthy adjustment of expectations, and fulfills one of our original goals to let users adequately gauge trust in automation: : *"It's hard for me to take somebody else's word or a computer work (...), I need my own kind of control to make sure... because it's important, right? To know that I'm capturing everything that's significant about the patient"* (P5). All participants were able to identify and assess automation mistakes, and ultimately found that the information provided by the system was valuable despite the errors (Fig. 6, item 3).

4.3.5 Comprehensiveness and Trust

We found mixed results on whether *Doccurate* increased confidence in a physician's own assessment. On one hand, the questionnaire indicates little to no effect in improving confidence across tasks (as per mean/median scores for each task, on Fig. 6, item 1), and the aforementioned concerns on automation mistrust affected how Timeline suggestions were taken into account. On the other, some of the features to support comprehensiveness were underused (*e.g.*, the "Other" FC) and most participants stated that FCs were useful to indicate points that might have been missed otherwise (P1 P2 P5 P6). Interestingly, P2 commented that *Doccurate*'s comprehensiveness actually made him less confident in his assessment, given the many additional elements not mentioned on the last note (which he reported typically relying on).

Looking further at individual self confidence scores (Fig. 6, item 1), we find a polarized distribution with either confident (P2 P3 P5) or unconfident (P1 P4 P6) assessments. Upon further inquiry, we found there were extraneous trust factors at play. For example, half of the participants mentioned exercising general caution regarding the inherent uncertainty in the patient charts and the limitations of focusing only on recent notes (P1 P2 P5), *e.g.*, *"if I had gone to the very last note, would it have encompassed all of this information already? I don't know"* (P1); conversely, the other half (P3 P4 P6) felt comfortable relying on the last note, and claimed that they don't always need to know absolutely everything about the patient, *e.g.*, *"(..) the last discharge summary is usually pretty decent in terms of what are the (past) medical*

history issues so far (..)". Other participants commented on the quality of the notes themselves, that they were poorly structured and therefore hard to read (P2), and not amenable to automation (P5). This speaks to the complexity of the space, which warrants further investigation overall, but also touches on the variety of information needs and the importance of providing enough flexibility to support such needs, which we sought to achieve via customization.

4.3.6 Suitable Application Scenarios

While we found no clear consensus on what medical contexts participants judged *Doccurate* would be more useful for in their practice, there was a general trend towards scenarios that benefit from comprehensiveness. This included roles such as general practice (P1 P4), nursing (P6), social work and other psychosocial-driven specialties (*e.g.*, psychiatry) (P3 P6), and activities such as seeing a patient for the first time (P2 P5) and reviewing undiagnosed patients (P1). On the other hand, two participants also commented that it was useful for directed searches (P2 P4), therefore potentially supporting both ends of the spectrum.

5 Discussion

Despite the limited nature of our study with its qualitative focus and small participant pool, we found compelling evidence that our curation-based approach can bring value to clinical practice. In particular, we were happy to observe that physicians were able to easily identify automation errors, that curation actions (including FC creation and error management) were wholly integrated into the chart review process, and that physicians found value in the information collected by the FCs.

The study also shed light on how *Doccurate* may be best leveraged in practice and further improved. We found that while some participants struggled with creating FCs from scratch, they appreciated the preset collection of filters. We posit that an initial setup with general, ***systems-based presets along with other general purpose FCs*** (*e.g.*, providing pointers to section headers such as *"Social History"* and *"Allergies"*) could be offered as a base, a set which physicians can progressively tailor to suit their individual needs. In this case, it would be useful to ***visually differentiate preset filters from physician-defined filters***.

Customization is still essential, as we envision physicians would create a potentially large collection of problem-based FCs, *e.g.*, to cover chronic conditions (such as *Diabetes*) or correlated effects (such as *Cardio Risk Factors*). That said, the FC creation and customization process should be more efficient and scalable. One strategy would be to ***leverage physician-defined FCs towards the creation of a shared FC library*** and to ***suggest "community" FCs that are relevant to a patient***. This could facilitate FC uptake and potentially speed up chart review workflows, but may also introduce additional overhead given the uncertainty involved in dealing with someone else's FCs. Another idea is to ***suggest FCs by clustering related non-tagged items*** (*i.e.*, terms under the "Other" FC), possibly using topic models. While prior work found that physicians can judge pertinence and relevance of topic items [1], additional information to assess suggestions such as ***quantifying and conveying uncertainty would increase confidence***. On the topic of automation support, providing more active ***user support for error correction*** is also important; despite being able to pinpoint errors, participants were often not sure how to best address them.

Regarding cognitive burden, participants did find the experience overwhelming. While extended usage would likely mitigate this effect, a few improvements could be beneficial, such as ***hiding the curation panel when not in use*** to make up more space for content and ***providing a visual representation for the concept codes hierarchy***.

Finally, we also identified areas for further investigation, particularly around the trade-off between efficiency and trust, as well as the practical limits of automation. While several of the classification mistakes encompass solvable problems (*e.g.,* negation detection), many others spring from task-, preference- and intent-specific contextual factors that are challenging to predict and are unlikely to be solved simply with "better automation". Ultimately, one should consider that no system will be 100% accurate and that providing adequate mechanisms for inspection is essential to NLP-powered visualizations. Another challenge lies in defining ground truth and accuracy measures prior to deployment in real practice, since it is difficult to predict all relevant in-the-wild contextual factors [29]. We posit that our curation strategy could serve both as a transition technology and as a means to collect training data for active learning [21]. This trade-off could have the potential to break the dependency in the NLP non-adoption cycle of which (a) automation cannot be integrated into practice due to lack of comprehensive validations, but (b) comprehensive validations cannot be fully performed without deployment "in the wild" [29]. These are interesting questions to pursue further, possibly in a longitudinal study.

6 Use Case Scenarios

Following our reflections in the Discussion, we present use case scenarios to demonstrate how we envision *Doccurate* could be used in practice. These scenarios were created in collaboration with a clinical practitioner (the 2nd author of this work) so as to be representative of real clinical decision-making workflows. We created a *fictional* patient identity and context around a *real* patient chart (anonymized for dates, names and places), chosen among the three charts we used in the evaluation study. We highlight references to *text snippets* in italicized form, `FC keywords or taxonomy codes` in typewriter font, and **FCs** in bold face.

6.1 FC Curation for Chart Review

Macy Gauthier is a 37 year old lady with a history of severe and poorly controlled diabetes, and a number of consequent hospital admissions. She has recently moved into town, and found herself a new local family doctor (*i.e.*, general practitioner), Dr. J, to follow-up on her case. Dr. J just received a copy of Macy's medical chart, and takes a few minutes to review it with *Doccurate* prior to their consultation; it is June 2206.

Macy's chart is initially loaded with a set of general, systems-based FCs (as listed on Fig. 2). On the Timeline, mentions of *type 1 diabetes* and *DKA* (Diabetic Ketoacidosis, an acute life-threatening complication of type 1 diabetes) are immediately noted. Drilling into the **Endocrinology** FC and quickly reviewing associated snippets, Dr. J notes the prevalence of a diabetes-related history and decides to create a dedicated FC, **Diabetes**, to facilitate follow-up of this chronic condition. Relevant **Endocrinology**-highlighted mentions on the current visible note were added to this new FC, including *insulin*, *hypoglycemia* and all mentions of *diabetes mellitus* (by adjusting the hierarchy of the `Diabetes Mellitus type 1` tag to its parent node, `Diabetes Mellitus`). Given the prevalence of these issues, Dr. J suspects that there might be other diabetes-related complications in the record nested under other FCs that could be added to **Diabetes**, and makes a mental note to keep an eye on them while reviewing other issues.

The following stop is **Cardiovascular**, since (a) diabetes is a significant risk factor for heart disease, and (b) it figures as a prevalent FC. However, a quick glance of the items encompassed in the Timeline doesn't appear to include any serious manifestations, and an inspection over `myocardial infarction` (*i.e.*, heart attack) reveals no positive mentions. Dr. J decides that the few positive cardiac events are not significant enough to establish a significant link to diabetes, and decides not to add any cardiac issue to **Diabetes** at this time. Before moving on, one quick inspection on *cath* mentions, which was expected to refer to cardiac catheterization procedures, was found to actually refer to *foley cath*, which is a urinaty catheter intervention. Noticing the automation failed to distinguish this nuance, Dr. J decides to removes the code tags on *cath* mentions to avoid further false positives and adds the compound keyword `foley cath` to **Nephrology** instead.

This last intervention led Dr. J to check the **Nephrology** FC. There, Dr. J finds mentions of *nephropathy*, and upon snippet inspection finds out that this issue is caused by her diabetes, along with neuropathy and retinopathy (which were mentioned in the same paragraph). All three terms are added to **Diabetes**. Still under **Nephrology**, several recent mentions of *ESRD* (End Stage Renal Disease, a very significant finding which could be linked to diabetic nephropathy) and *hemodialysis* were found in the latter history (from July 2203 onwards). Dr. J decides this history is significant enough to warrant the creation of a **Renal Dialysis** FC to keep track of the dialysis interventions, and proceeds to add the corresponding terms as well as related *creatinine* measurements to it.

Following the earlier mentions of neuropathy, Dr. J decides to quickly check the **Neurology** FC for related findings, and *seizures* emerges as the most frequent item. It appears with a high but relatively narrow frequency burst mid 2204 (that is, two years prior), indicating the issue was significant but appears to have been stabilized. The binning interval was adjusted from 3 weeks to 1 week to provide more resolution, and shows the term was consistently mentioned for a few weeks around June 2204. Dr. J quickly reviews the related seizure history directly from the notes to learn more, and confirms that the more serious episodes indeed happened during that time, but there were also other suspected "seizure-like" events later on that should be monitored. Knowing that severe diabetes can trigger such episodes, Dr. J wonders if *seizures* should be added to **Diabetes**. Upon further inspection it was found that the etiology for seizures was uncertain given Macy's numerous health issues, and so another FC, **Seizure**, was created to keep track of this issue separately. At this time, the front desk clerk calls to inform the patient has arrived. By now, Dr. J has identified a number of chronic conditions to guide this first patient visit, and hopes to establish good rapport with the patient by showing preparedness.

6.2 Follow up and for Ongoing Care

Dr. J has a productive first visit with Macy and is able to rapidly form a list of key medical issues. A plan is made to follow-up in 4-weeks time with Macy scheduled to continue with her ongoing dialysis with her Nephrologist and complete bloodwork prior to her next appointment. At the end of the visit, Dr. J adds a **Laboratory** FC to facilitate easier tracking of her bloodwork results. Dr. J, now has a custom curated *Doccurate* chart for Macy, highlighting her key medical issues which will streamline review of her chart prior to the next follow-up appointment. In addition, if Dr. J, is away from clinic and a physician colleague is covering his patients, they will have access to an already curated *Doccurate* chart highlighting Dr. J's medical priorities for Macy thus facilitating a rapid chart review as needed.

6.3 Overview for Emergency Care

Two weeks after Macy's appointment with her new family doctor, she is found by her partner to be unresponsive on the floor of her bedroom. He calls for emergency medical services and is instructed to begin CPR. An ambulance arrives and she is brought to the nearest community Emergency Department. Upon arrival she is taken to a resuscitation room with multiple physicians and nurses. Macy's partner has not yet arrived to the hospital and there are no other relatives present to provide the team with a medical history. Dr. D is one of the emergency physicians and is tasked with searching through Macy's extensive medical chart. She opens Macy's *Doccurate* chart and sees the Timeline (1st level) featuring the FCs curated by Dr. J.

At first glance Dr. D immediately notices *Insulin* in large font representing a relative increased frequency of mention throughout Macy's chart. She also sees the curated **Diabetes** FC that was previously created by Macy's family doctor. Knowing that Macy has *Type 1 Diabetes Mellitus* prompts Dr. D to call this piece of information out to the team as Macy's loss of consciousness (LOC) may be related to hypoglycemia. Opening the **Diabetes** FC reveals that Macy has an array of complications related to her diabetes that may be contributing to her LOC. Dr. D notes that Macy has had *DKA* in the past, which is a condition of severe hyperglycemia that can cause cerebral edema and acidosis leading to LOC. Dr. D also notes unexpectedly that Macy at a young age of 37 has end stage renal disease and is on dialysis as a complication related to her diabetes. This can lead to severe electrolyte disturbances and LOC. Lastly, Dr. D. notes that Macy has a chronic foot ulcer, which may be a potential source of infection and sepsis leading to LOC. Given Macy's clinical status, Dr. D worries that Macy may be having a cardiovascular emergency. She quickly explores the **Cardiovascular** FC and reviews the associated snippets to discover that Macy does not have any known life-threatening cardiovascular disease on history. Dr. D gathers this information rapidly from the *Doccurate* visualizations and informs her fellow Emergency Physicians at Macy's bedside resuscitating her.

This information related to Macy's complex medical history is vital for the emergency medicine team as it provides a framework for generating differential diagnoses as to why Macy has LOC and helps guide rapid decision-making. Using this information the team realizes that Macy may be in DKA given the history seen on her *Doccurate* visualization. Administering routine volumes of IV fluids will be dangerous for her as it can worsen her cerebral edema and the team changes their resuscitation approach. Bloodwork is rapidly completed demonstrating a dangerously high blood sugar with an acidosis and the diagnosis of DKA confirmed. The appropriate life saving treatment of an insulin infusion is started and Macy slowly recovers over the upcoming days.

7 Limitations

One limitation of this work is the small scope and duration of the study; with our focus on user experience, we did not conduct a fully fledged quantitative analysis, nor assessed participant's performance with the tool. While findings are promising, questions such as whether this model could be efficient in the long run are difficult to answer without a larger cohort and a longitudinal study setup.

Another criticism of *Doccurate* was that some participants felt it did not easily convey the patient "narrative". While supporting patient narratives was not the primary focus of this work, it is an important aspect to deliver as part of the overall solution for physicians. We believe curation could help support that goal, for instance by allowing physicians to specify highlights in the text that could be presented in a storytelling manner.

Finally, since *Doccurate*'s curation operations dynamically operate on a large collection of documents and tags, several of them appeared laggy to users. This negatively affected satisfaction and how participants perceived efficiency. Adjustments on this front should lead to significant improvement in perceived value.

8 Conclusion and Future Work

In this paper we propose and assess the value of curation-based approaches for physicians to visualize and peruse clinical text. We present *Doccurate*, a semi-automatic approach that uses NER-structured text and allows for the creation of semantic filters based on structured knowledge encoded in medical taxonomies. An evaluation of *Doccurate* with 6 domain experts revealed that the approach has potential to leverage text via a flexible and conscientious use of automation that seamlessly integrates into clinical tasks.

To our knowledge, this is the first exploration of its kind for the medical space and as such represents but an initial step towards more efficient and flexible access to clinical text. We propose a few avenues for future work. First, we debate on the value of extending FC expressive power by leveraging *multiple taxonomies* at a time or even *ontologies* (*i.e.*, encompassing arbitrary relationships, not just parent/child). While this may seem like a straightforward improvement, we argue that one of the strengths of current FCs is their simplicity and posit that the added power of leveraging multiple relationships may not outweigh the burden of added complexity.

Second, we argue there is room to better support trust over automated processes, especially if additional components are added to the pipeline, *e.g.*, FC suggestions and negation detection. Possibilities could include quantifying and conveying *uncertainty* of automated processes [36] and providing *explanations* of automated decisions [34] that weigh curation input into their reasoning models. User-centered research leading to a better understanding of how physicians perceive and correct errors should also be leveraged to improve NLP automation. Finally, we believe our proposed curation-based approach for text exploration could be useful to other domains that require perusing large collections of narrative text, such as journalistic inquiry and intelligence analysis.

Acknowledgments

We would like to thank Juliana de la Vega for her assistance with the user studies, V. Bilbily and M. Tao for their contributions to the manuscript, and the DGP lab as a whole for the helpful discussions and ideas. We also thank the anonymous reviewers for their thoughtful comments. This work was partly funded by a *Natural Sciences and Engineering Research Council (NSERC) Discovery Grant*.

REFERENCES

[1] C. W. Arnold, A. Oh, S. Chen, and W. Speier. Evaluating topic model interpretability from a primary care physician perspective. *Computer methods and programs in biomedicine*, 124:67–75, 2016.

[2] J. W. Beasley, T. B. Wetterneck, J. Temte, J. A. Lapin, P. Smith, A. J. Rivera-Rodriguez, and B.-T. Karsh. Information chaos in primary care: implications for physician performance and patient safety. *The Journal of the American Board of Family Medicine*, 24(6):745–751, 2011.

[3] N. G. Belmonte. Extracting and visualizing insights from real-time conversations around public presentations. In *Visual Analytics Science and Technology (VAST), 2014 IEEE Conference on*, pp. 225–226. IEEE, 2014.

[4] C. Bossen and L. G. Jensen. How physicians 'achieve overview': a case-based study in a hospital ward. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pp. 257–268. ACM, 2014. doi: 10.1145/2531602.2531620

[5] A. A. Bui, D. R. Aberle, and H. Kangarloo. Timeline: visualizing integrated patient records. *IEEE Transactions on Information Technology in Biomedicine*, 11(4):462–473, 2007. doi: 10.1109/titb.2006.884365

[6] W. W. Chapman, P. M. Nadkarni, L. Hirschman, L. W. D'avolio, G. K. Savova, and O. Uzuner. Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions, 2011.

[7] M. V. Devarakonda, N. Mehta, C.-H. Tsou, J. L. Liang, A. S. Nowacki, and J. E. Jelovsek. Physicians assessment of IBM Watson generated problem list. Technical report, IBM Research, 2016.

[8] M. Dörk, D. Gruen, C. Williamson, and S. Carpendale. A visual backchannel for large-scale events. *Visualization and Computer Graphics, IEEE Transactions on*, 16(6):1129–1138, 2010.

[9] W. Dou, X. Wang, D. Skau, W. Ribarsky, and M. X. Zhou. LeadLine: Interactive visual analysis of text data through event identification and exploration. In *Visual Analytics Science and Technology (VAST), 2012 IEEE Conference on*, pp. 93–102. IEEE, 2012.

[10] T. A. S. Foundation. Apache cTAKES, 2013.

[11] J. Fulda, M. Brehmel, and T. Munzner. TimelineCurator: Interactive authoring of visual timelines from unstructured text. *IEEE transactions on visualization and computer graphics*, 22(1):300–309, 2016. doi: 10.1109/TVCG.2015.2467531

[12] M. Glueck, M. P. Naeini, F. Doshi-Velez, F. Chevalier, A. Khan, D. Wigdor, and M. Brudno. Phenolines: Phenotype comparison visualizations for disease subtyping via topic models. *IEEE Transactions on Visualization & Computer Graphics*, (1):1–1.

[13] M. Habibi, L. Weber, M. Neves, D. L. Wiegandt, and U. Leser. Deep learning with word embeddings improves biomedical named entity recognition. *Bioinformatics*, 33(14):i37–i48, 2017.

[14] C. Hallett. Multi-modal presentation of medical histories. In *Proceedings of the 13th international conference on Intelligent user interfaces*, pp. 80–89. ACM, 2008. doi: 10.1145/1378773.1378785

[15] C. Hallett, R. Power, and D. Scott. Summarisation and visualisation of e-health data repositories. In *UK E-Science All-Hands Meeting, 18-21 Sept 2006, Nottingham, UK*, 2006.

[16] S. Havre, B. Hetzler, and L. Nowell. Themeriver: Visualizing theme changes over time. In *Information Visualization, 2000. InfoVis 2000. IEEE Symposium on*, pp. 115–123. IEEE, 2000.

[17] J. S. Hirsch, J. S. Tanenbaum, S. L. Gorman, C. Liu, E. Schmitz, D. Hashorva, A. Ervits, D. Vawdrey, M. Sturm, and N. Elhadad. Harvest, a longitudinal patient record summarizer. *Journal of the American Medical Informatics Association*, pp. amiajnl–2014, 2014. doi: 10.1136/amiajnl-2014-002945

[18] W. Hsu, R. K. Taira, S. El-Saden, H. Kangarloo, and A. A. Bui. Context-based electronic health record: toward patient specific healthcare. *IEEE Transactions on information technology in biomedicine*, 16(2):228–234, 2012. doi: 10.1109/titb.2012.2186149

[19] S. International. SNOMED-CT, 2002.

[20] A. E. Johnson, T. J. Pollard, L. Shen, H. L. Li-wei, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark. MIMIC-III, a freely accessible critical care database. *Scientific data*, 3:160035, 2016.

[21] M. Kholghi, L. Sitbon, G. Zuccon, and A. Nguyen. Active learning: a step towards automating medical concept extraction. *Journal of the American Medical Informatics Association*, 23(2):289–296, 2015.

[22] S. Koch, M. John, M. Worner, A. Muller, and T. Ertl. VarifocalReader—in-depth visual analysis of large text documents. *Visualization and Computer Graphics, IEEE Transactions on*, 20(12):1723–1732, 2014.

[23] S. Liu, Y. Chen, H. Wei, J. Yang, K. Zhou, and S. M. Drucker. Exploring topical lead-lag across corpora. *Knowledge and Data Engineering, IEEE Transactions on*, 27(1):115–129, 2015.

[24] S. Liu, M. X. Zhou, S. Pan, Y. Song, W. Qian, W. Cai, and X. Lian. Tiara: Interactive, topic-based visual text summarization and analysis. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 3(2):25, 2012.

[25] T. Mønsted, M. C. Reddy, and J. P. Bansler. The use of narratives in medical work: a field study of physician-patient consultations. In *ECSCW 2011: Proceedings of the 12th European Conference on Computer Supported Cooperative Work, 24-28 September 2011, Aarhus Denmark*, pp. 81–100. Springer, 2011. doi: 10.1007/978-0-85729-913-0_5

[26] Z. Morrison, B. Fernando, D. Kalra, K. Cresswell, A. Robertson, A. Hemmi, and A. Sheikh. An evaluation of different levels of structuring within the clinical record:final report for the NHS Connecting for Health evaluation programme. Technical report, 2012.

[27] U. N. L. of Medicine. Unified Medical Language System (UMLS), 2009.

[28] D. Park, S. Kim, J. Lee, J. Choo, N. Diakopoulos, and N. Elmqvist. Conceptvector: text visual analytics via interactive lexicon building using word embedding. *IEEE transactions on visualization and computer graphics*, 24(1):361–370, 2018.

[29] R. Pivovarov and N. Elhadad. Automated methods for the summarization of electronic health records. *Journal of the American Medical Informatics Association*, 22(5):938–947, 2015. doi: 10.1093/jamia/ocv032

[30] C. Plaisant, D. Heller, J. Li, B. Shneiderman, R. Mushlin, and J. Karat. Visualizing medical records with lifelines. In *CHI 98 Cconference Summary on Human Factors in Computing Systems*, pp. 28–29. ACM, 1998. doi: 10.1145/286498.286513

[31] C. Plaisant, B. Milash, A. Rose, S. Widoff, and B. Shneiderman. Lifelines: visualizing personal histories. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 221–227. ACM, 1996. doi: 10.1145/238386.238493

[32] C. Plaisant, R. Mushlin, A. Snyder, J. Li, D. Heller, and B. Shneiderman. Lifelines: using visualization to enhance navigation and analysis of patient records. In *Proceedings of the AMIA Symposium*, p. 76. American Medical Informatics Association, 1998. doi: 10.1016/b978-155860915-0/50038-x

[33] S. M. Powsner and E. R. Tufte. Graphical summary of patient status. *The Lancet*, 344(8919):386–389, 1994.

[34] M. T. Ribeiro, S. Singh, and C. Guestrin. Why should I trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144. ACM, 2016.

[35] A. Rind, T. D. Wang, W. Aigner, S. Miksch, K. Wongsuphasawat, C. Plaisant, B. Shneiderman, et al. Interactive information visualization to explore and query electronic health records. *Foundations and Trends® in Human–Computer Interaction*, 5(3):207–298, 2013.

[36] D. Sacha, H. Senaratne, B. C. Kwon, G. Ellis, and D. A. Keim. The role of uncertainty, awareness, and trust in visual analytics. *IEEE transactions on visualization and computer graphics*, 22(1):240–249, 2016.

[37] G. K. Savova, J. J. Masanz, P. V. Ogren, J. Zheng, S. Sohn, K. C. Kipper-Schuler, and C. G. Chute. Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513, 2010.

[38] N. Sultanum, M. Brudno, D. Wigdor, and F. Chevalier. More text please! understanding and supporting the use of visualization for clinical text overview. In *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 2018. doi: 10.1145/3173574.3173996

[39] H. Tange. How to approach the structuring of the medical record? towards a model for flexible access to free text medical data. *International journal of bio-medical computing*, 42(1):27–34, 1996.

[40] G. Trivedi, P. Pham, W. W. Chapman, R. Hwa, J. Wiebe, and H. Hochheiser. NLPReViz: an interactive tool for natural language processing on clinical text. *Journal of the American Medical Informatics Association*, 25(1):81–87, 2017.