Noname manuscript No.
(will be inserted by the editor)

LetterVis: A Letter-Space View of Clinic Letters

Qiru Wang · Robert S. Laramee · Arron Lacey · Owen Pickrell

Received: date / Accepted: date

Abstract The number of Electronic Health Records (EHR) collected by healthcare providers is growing at an unprecedented pace. Clinicians often compose detailed clinic letters to record as much essential information during consultations as they can. This increases the workload of analyzing these letters, performing individual and collective analysis, and clinical decision-making. This paper presents a novel visualization tool, LetterVis, to support the analysis of clinic letters through advanced interactive visual designs and queries. We describe a letter-space that facilities the visual exploration of content and patterns inside a letter. Letters are processed using Natural Language Processing (NLP) techniques and explored in multiple linked interactive views providing different levels of abstraction. The tool includes customized visual designs and views for visualizing antiepileptic drugs (AEDs). We provide a range of filtering and selection options to assist pattern finding and outlier detection. We demonstrate LetterVis with three case studies using anonymized clinic letters, revealing insight that is normally either time-consuming or impossible to observe. Domain expert partners from EHR analysis review the software and are involved in every phase from the initial design to evaluation.

Keywords Inforamtion Visualization · Electronic Health Records · Visual Analytics

Qiru Wang · Robert S. Laramee
Department of Computer Computer Science, University of Nottingham, Nottingham, NG8 1BB, UK
E-mail: {qiru.wang, robert.laramee}@nottingham.ac.uk

Arron Lacey · Owen Pickrell
Neurology and Molecular Neuroscience Group, Institute of Life Science, Swansea University, Swansea, SA2 8PP, UK
E-mail: {a.s.lacey, w.o.pickrell}@swansea.ac.uk

Routinely collected Electronic Health Records (EHR) such as clinic letters contain important information such as demographics, past and present prescriptions and previous check-ups, all of which are valuable in answering clinical research questions. These letters are often stored in a free-text format by clinicians with different writing styles, thus making the extraction of critical information for further analysis a time-consuming and error-prone process. Even with the assistance of machine learning and modern statistical methods, effectiveness remains limited, let alone the challenges associated with transparency, replicability and ethics [27].

Visual analytics (VA) and visualization often involve real-time human observation and intervention. VA has great potential to support clinical decision-making and inform further research under close scrutiny to ensure both quality and transparency. Interactive visual designs can efficiently reduce visual clutter to cope with the exploding data volumes, supporting quantitative as well as qualitative analysis in an interpretable and explainable manner [26]. We propose LetterVis, an interactive letter-space visualization tool specifically designed to enable the efficient exploration and analysis of clinic letters. By letter-space, we mean the standard coordinate system used by clinicians to write clinic letters, e.g. A4 space. See Sec. 4 for more detail.Our collaboration with domain experts informs a novel design that utilizes the information-rich clinic letters to assist in hypothesis generation and verification. We focus on epilepsy in our case studies, but our tool can be easily generalized to support all content in the form of letters. Our contributions include:

- A novel letter-space visualization tool that leverages natural language processing (NLP) techniques to support the exploration of unstructured clinical text in a structured manner,

– Novel and customized visual designs to identify and verify patterns and outliers in a cohort of patients,
– Dynamic analysis and comparison of antiepileptic drug (AED) co-prescriptions through multiple coordinated visual layouts,
– Three replicable case studies to demonstrate LetterVis' ability to support hypothesis verification.

The motivation behind our study is to help EHR researchers explore and analyze AED co-prescriptions in unstructured EHR letter data to identify pattern patterns and outliers, and also to provide an overview, filtering and selection and analysis options. From our interviews with healthcare data analyst experts, they report their recent adoption of advanced visual designs and VA for analysis of AED co-prescriptions yields promising outcomes. However, they point out that usability is an obstacle to further the analysis of AED co-prescriptions, as they often encounter a steep learning curve. This in turn may result in more human errors.

1 Related Work

Our work primarily focuses on EHR Visualization (EHR Vis). Surveys by Rind et al. [19], West et al. [28], and Rostamzadeh et al. [20] provide an overview of EHR Vis. Our work also overlaps with other topics including Visual Analytics (VA) and Text Visualization. McNabb and Laramee [14] provide a comprehensive overview of information visualization and visual analytics, which includes three surveys focusing on healthcare visualization.

Electronic Health Record Visualization: To the best of our knowledge, there is no standard definition of an EHR since its inception in the 1960s [16]. A popular definition defines an EHR as, *"A longitudinal collection of electronic health information about individual patients and populations for supporting the analysis of healthcare, education and research"* [8, 10].

Bernard et al. [1] build a visual-interactive system that enables physicians to train models for prostate cancer identification. Glueck et al. present a trilogy of visual analysis tools for phenotype comparison: PhenoBlocks [5] with a novel differential hierarchy comparison algorithm accompanied with a customized sunburst radial hierarchy layout, PhenoStacks [4] with a novel topology simplification algorithm to eliminate duplicates, and incorporates natural language queries for searching, and PhenoLines [6] adds the support for the visualization of temporal evolution of phenotypes.

Natural Language Processing: Natural Language Processing (NLP) is defined by Liddy as *"a theoretically motivated range of computational techniques for analyzing and representing naturally occurring texts at one or more levels of linguistic analysis for the purpose of achieving human-like language processing for a range of tasks or applications"* [13]. NLP plays a significant role in the visualization and visual analytics of EHR data archived as free text. EHR-NLP approaches usually include the use of existing tools, customized classification algorithms and curation-based extraction [11].

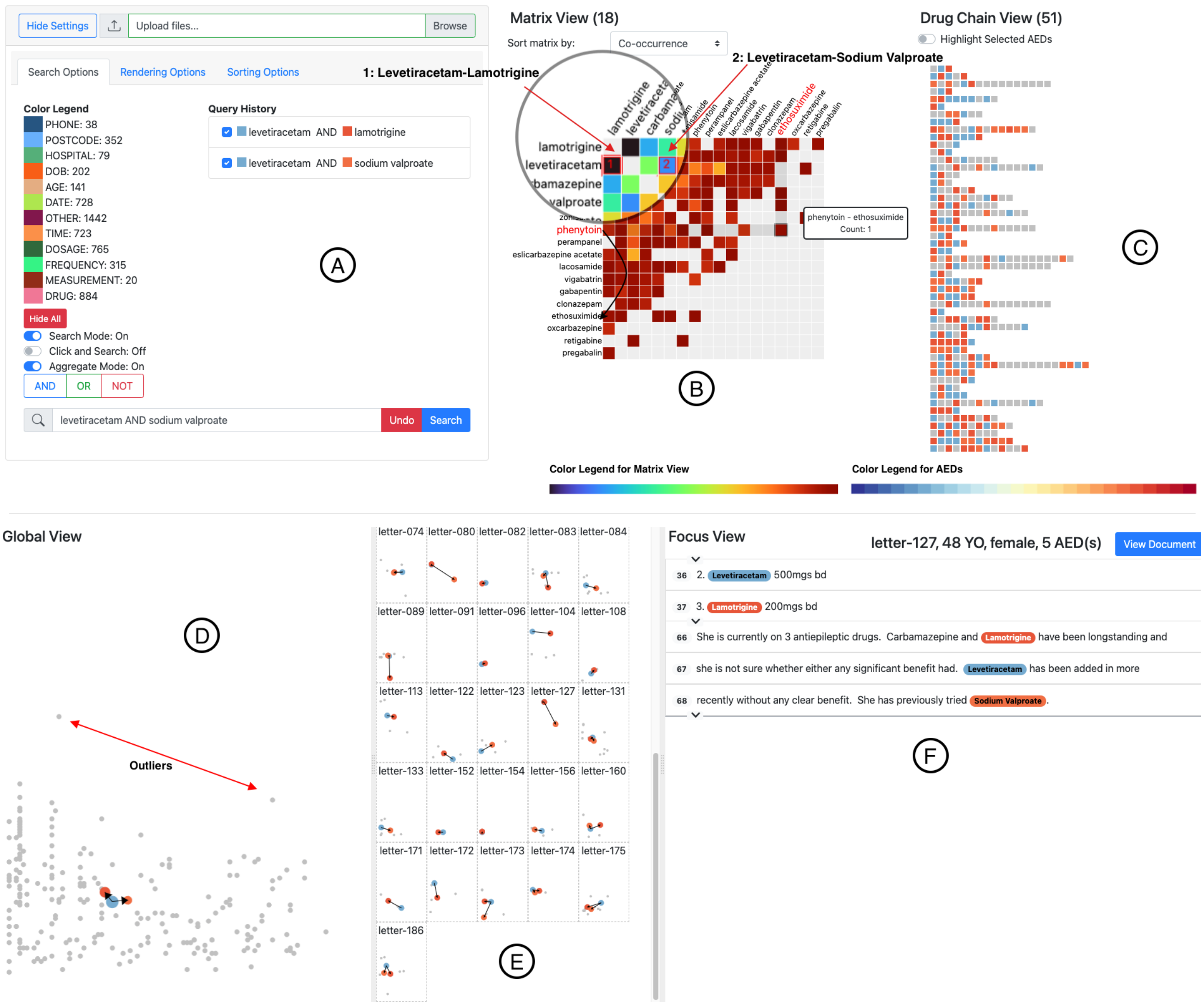
Through our domain expert partners, we learn that GATE, an open source text analysis tool that supports text mining of biomedical documents via natural language processing (NLP), developed by Cunningham et al. [3], is one of the most popular choices when it comes to EHR data pre-processing. GATE is capable of extracting structured information from unstructured free text, eg. clinician's notes and discharge letters. However, it lacks interactive visualization features to assist advanced analysis.

Zhang et al. incorporate NLP algorithms in their AnamneVis [29] to extract structured medical information from doctor-patient dialogs and medical reports to assist the visualization of patient medical history. Trivedi et al. [25] introduce NLPReViz that interactively trains NLP models for classifying information extracted from clinical records. Sultanum et al. present Doccurate [22] that provides an accurate and sufficient overview for individual patients based on the user-supplied extraction rules. The focus of our work is not NLP itself, but rather extracting a structured visual representation of the data hidden inside letters with the help of NLP.

The key difference between our work and previous work is the introduction, development and evaluation of letter-space, and its application to clinic letters. We work closely with health data analysts to curate a specific list of extraction rules for processing epilepsy related EHRs. To facilitate the analysis of the result, we incorporate interactive visual designs along with an advanced query interface that is compatible with Apache Lucene [23], a text search engine library that is known for its flexible and efficient searching algorithms. The library came to our attention after one of our expert health data analyst partners' recommendation.

2 EHR Data Description

Our data includes 200 clinic letters written by clinicians specialized in neurology and provided by our healthcare data analyst expert partners.Each letter represents a single patient visit to a neurologist. A typical letter contains identifiable information including patient name, age, gender, address, NHS number, AED prescriptions (past and present), symptoms, diagnosis and other health-related information. Due to the sensitive



Hide Settings | Upload files... | Browse

Search Options | Rendering Options | Sorting Options

Color Legend
- PHONE: 38
- POSTCODE: 352
- HOSPITAL: 79
- DOB: 202
- AGE: 141
- DATE: 728
- OTHER: 1442
- TIME: 723
- DOSAGE: 765
- FREQUENCY: 315
- MEASUREMENT: 20
- DRUG: 884

Hide All

Search Mode: On

Click and Search: Off

Aggregate Mode: On

AND | OR | NOT

levetiracetam AND sodium valproate | Undo | Search

Query History
- ☑ levetiracetam AND lamotrigine
- ☑ levetiracetam AND sodium valproate

Focus View — letter-127, 48 YO, female, 5 AED(s) — View Document

36 2. Levetiracetam 500mgs bd

37 3. Lamotrigine 200mgs bd

66 She is currently on 3 antiepileptic drugs. Carbamazepine and Lamotrigine have been longstanding and

67 she is not sure whether either any significant benefit had. Levetiracetam has been added in more

68 recently without any clear benefit. She has previously tried Sodium Valproate.

Fig. 1 An overview of LetterVis. (A) shows the user options for searching, rendering and sorting. (B) illustrates the matrix view based on AED co-occurrences in the dataset, user-chosen cells are highlighted with a sequence number, which corresponds to their position in the history of queries. (D) depicts an overview of all super-imposed letters and their search result centroids in the dataset. (E) shows individual thumbnails of letters with local search result centroids joined by edges. (F) contains a detailed view of the letter in focus, lines without data-of-interest are collapsed by default.

nature of EHR data, the letters are manually anonymized by clinicians with patient identifiable information as well as other potential identifiable information manually replaced with similar but fictional text.

We first extract all text from Word files, the length of letters varies from 28-133 lines with an average of 98 characters per line. We then pre-process the letters using NLP with a list of curated extraction rules provided by domain experts to extract 12 text data categories. Our extraction rules started with numerical data as proof-of-concept. This was then expanded to include anti-epileptic prescription information based on our health data analysts' feedback. We also extract metadata for these categories, such as the position and length of matching text. The extracted data is then used to generate visualizations that match the position of data samples in the original letters.

In later iterations of the software development, we received a list of 26 generic antiepileptic drug (AED) names together with 24 equivalent trade names from our domain expert partners for the purpose of exploring patterns in drug co-prescriptions and effective combinations. Each AED is assigned a color from the colormap shown in Fig. 1A. We develop the matrix view (Section 4.2 - Matrix View) and drug chain view (Section 4.2 - Drug Chain View) based on this list and the user requirements in Sec. 3, with an additional pre-defined color legend for each AED shown in Fig. 1C.

3 Application Design Methodology

We describe our collaboration with three health data analyst experts in this section.

We collaborate closely with a consultant neurologist (E1) from the UK National Health Service (NHS), and a lecturer in Natural Language Processing who is also a senior health informatics research analyst in epilepsy-related research (E2) from a UK University. We also interview a health data scientist (E3) from a UK University Medical School during the initial design stage. Data for this application is provided by E1 and E2, described in Sec. 2.

Informing the Initial Design: Our initial design was informed by interviewing three health data analyst experts in EHR analysis. We follow the guidance of Hogan et al. [9] and constructed a set of interview questions involving 14 structured, semi-structured and open-ended questions. Each one-on-one interview session lasted around 25 minutes and is recorded and archived for post-analysis. We then analyzed the domain requirements to guide the development.

Informing Further Software Iterations: Over the course of development spanning over 12 months, we consulted E1 and E2 for feedback. E3 did not participate in the development as her specialization in injury prevention is not related to the letters on epilepsy that we worked on. We presented intermediate visualization prototypes to E1 and E2 in four separate feedback sessions. Each session lasted around 65 minutes and was video-recorded for post-analysis. We describe this feedback in detail in Sec. 5.2.

Our work can be easily extended and generalized to support other areas where letters are used systematically for communication, either currently or historically, for example, in the legal profession.

User Requirements and Design Goals

LetterVis was developed in collaboration with three health data analyst experts in clinic letter analysis. The following requirements are gathered from interviews and feedback sessions:

- **R1** An interactive tool that facilitates the exploration of EHR free text data,
- **R2** Software that supports the identification of patterns and outliers with respect to AED co-prescriptions in clinic letters,
- **R3** A design that supports the identification and exploration of AED co-prescriptions,
- **R4** An interface that supports analyzing several clinic letters simultaneously,
- **R5** Support for cross-referencing and linking visual representations with original letters,
- **R6** A tool that is compatible with experts' existing analytical workflow by supporting EHRs in JSON format.

Throughout the development process, we identified additional requirements from feedback:

- **R7** A query interface that is compatible with Apache Lucene syntax,
- **R8** An interface that conveys AED prescription evolution.

Brehmer and Munzner's multi-level typology of visualization tasks [2] provides a guidance on classifying and describing our visualization tasks.

Based on the topology, we derived six main tasks to meet the requirements above [2]:

- **T1** *Present* an overview of important text data with abstraction of user-chosen data-of-interest, that enables the user to *explore*, *identify* and *compare* patterns and outliers (**R1, R2, R4**). [present → explore → identify/compare]
- **T2** A coordinated visual interface that *presents* multiple levels of abstracted views to support the *exploration* of letters and *identification* of patterns and outliers (**R1, R2**). [present → explore → identify]
- **T3** A combination of customized visual designs for visualizing AED co-prescriptions and prescription progression, the interface enables the user to *lookup* and *compare* different letters and *select*, *arrange*, *change* and *filter* based on AED co-prescriptions (**R3, R4, R5, R8**). [lookup → compare → select/arrange/change/filter]
- **T4** Develop a visual query interface that is compatible with Apache Lucene syntax to support the existing analytical workflow which enables the user to *identify* outliers (**R6, R7**). [identify]
- **T5** Provide a range of interactive user options and their combinations, including *filtering* and *selection*, to support tasks **T1**, **T2** and **T3**. [select/filter]
- **T6** Provide a history of queries that supports the retrospective analysis through undo and redo functions. [record]

Our design follows the Visual Information-Seeking Mantra,"*overview first, zoom and filter, then details on demand*" [21], as a starting point. Shneiderman further proposes three essential tasks: *Relate* to view relationships between items, *History* to keep a list of actions performed and provide undo and redo functions, *Extract* to enable extraction of sub-collections. We believe the tasks established above are generic in most free text analysis projects, therefore our work can be extended to support other domains.

4 LetterVis for Visualization of EHR Letters

Valuable patient information is recorded and exchanged in the form of clinic letters to deliver quality patient care. Although letter text is usually described as unstructured, our basic hypothesis and motivation is based on the implicit knowledge and hence structure hidden in letters. For example, postcodes do not appear at random positions in the letters. Their position is consistently in the top-left in letter-space. We believe that the position of numerical data in letter-space can provide important clues about the context hidden inside the unstructured text, likewise for drugs and prescriptions (For example, see Fig. 2). Another reason we focus on letter-space is because this is the space that clinicians and EHR analysts operate in and are used to. Any of the unfamiliar visual designs that we develop can be linked back to the familiar letter-space to facilitate interpretation by the analysts and/or clinicians that write them in the first place. This is crucial for any interdisciplinary project. Also, position of data on a page is important because it can reveal outliers (**R2, T1**). The order in which drugs appear is not random. The choice and order of prescriptions reflect important medical and pharmaceutical expertise held by clinicians (For example, see Fig. 3). Our visual approaches extract this information and leverage it to facilitate exploration and analysis of clinic letters.

We construct a letter-space by deriving the width (x-axis) from the clinic letters. The width is the average length of text, approximately 98 characters, derived from the full collection of letter bodies. We then use the standard letter aspect ratio to calculate the height (y-axis) for depicting a letter-space in all three letter abstraction views described in Sec. 4.1.

Our letter-space approach enables the real-time exploration of clinic letters that assists the decision-making process. This approach is intuitive to clinicians as it resembles part of their analytical workflow, thus requiring less cognitive load. This section describes the five customized visual interfaces we propose in detail. We also introduce support for visual queries and sorting options.
Visualization of Numerical Values: In the first few iterations of LetterVis, we focused on visualizing numerical data as proof-of-concept. Specifically we focus on both the type of numerical data and its position in letter-space. We extract and classify all numerical values (Fig. 1A left) using customized NLP extraction rules based in regular expression, and visualize the distribution of values to depict clusters in letter-space. As we anticipated extraction rules constantly evolve during our collaborative development lifecycle to meet new requirements proposed by our domain experts. We decided to use regular expressions for their flexibility to quickly prototype and refine our software. The numerical categories are used to support analytical tasks in all case studies in Sec. 5.1. We chose this approach as a starting point to demonstrate out idea to the experts in order to get feedback and inform future software features. One limitation of this categorization of data quickly identified by our domain expert partners was the ability to query data by keyword, which is essential to their analytical workflow.

Color Legend Interaction: We use Colorgorical [7] to create a discriminable and aesthetically guided colormap to represent the 12 data categories extracted. The legend is shown in Fig. 1A left. Any legend component can be clicked or dragged to the search bar to initiate a category search. Clicking on a legend item toggles the rendering (in-focus) of the corresponding category and individual samples in all three abstraction views, as shown in Fig. 2. Fig. 1A bottom shows a color legend for AEDs. On-mouse-over displays the AED name. This part of the interface can be hidden if the user would like to reduce the complexity of the interface.

4.1 Three Levels of Letter Abstraction

We provide three different views to represent the abstraction of letters from a top-down perspective. The first level enables the exploration at the cohort level for global analysis (Fig. 1D). The second level represents the abstraction of each individual letter by juxtaposition for closer observation (Fig. 1E). The third level links detailed visual elements with the original letter (Fig. 1F). This approach enables the user to explore the letters at different levels to identify patterns and outliers in a cohort of patients. All views are linked and coordinated, supported by interactive user options.

Visual Elements in the Global and Thumbnail Views: In the global and thumbnail views, we introduce three visual elements:

- *Centroid:* represents the arithmetic mean position of each text data category (numerical, AED and search term-based) in letter-space
- *Individual sample:* represents an individual text data sample in letter-space
- *Edge:* connects a centroid to its individual samples in the same data category

Global View: The global view (Fig. 1D) is the first and highest level of abstraction that shows all search term samples in letter-space and their corresponding category centroids extracted from the dataset in one superimposed letter (**T1**). This superposition approach enables the comparison of search terms in all of the letters simultaneously that is otherwise difficult or even impossible through juxtaposition or explicit alignment. Using juxtaposition or explicit alignment to obtain an overview or make comparisons is ineffective in this case. Clicking on a centroid triggers the rendering of edges to the corresponding individual search term samples (Fig. 2). For example, Fig. 2 top left shows the centroids of each data category listed in the color legend in the global view of letter-space (left). The greyscale points are rendered as context. They are the positions of the original search data samples in letter-space. Selecting a single search dimension, e.g, Drugs, causes edges to be rendered from the search data dimension centroid to individual samples (Fig. 2 bottom left). We render edges because they convey the area covered by a category of values in letter-space. Also, search term samples located further from the centroid often have a higher chance of depicting outliers. A convex hull could have been used, however it does not show the variation and density of the original search term samples. Clicking on an individual sample in the global view shows the corresponding letter in the focus view (**T2**). On-mouse-over details are provided for every visual element.

Thumbnail View: As the second level of abstraction (**T1**, **T2**), each juxtaposed thumbnail in the thumbnail view (Fig. 1E) represents an individual letter. Similar to the global view, a user clicks on a centroid to show each connection to individual samples of the query. Clicking on the title shows the corresponding letter in the focus view. On-mouse-over information shows the original data for every visual element.

Focus View: The focus view is the third level of abstraction (**T2**) that shows a summarized version of letters (Fig. 1F). Individual samples are highlighted. Lines with no text data of interest are collapsed by default and can be expanded interactively via clicking on any arrow glyph. At any stage, clicking on the 'View Document' button brings the original letter, with all individual samples highlighted, into focus. An example is shown in Fig. 4.

Rendering Options: We provide three rendering options for all three visual elements (**T5**), *centroid*, *individual sample* and *edge* in both the global and thumbnail views. *Focus* shows the data in color (see the user-chosen centroid and its individual samples in Fig. 2), *Context* shows the data in greyscale (see context centroids and individual samples in Fig. 2) and *Hide* removes the data samples and edges. In the drug chain view (Section 4.2), an option is provided to highlight the data in context.

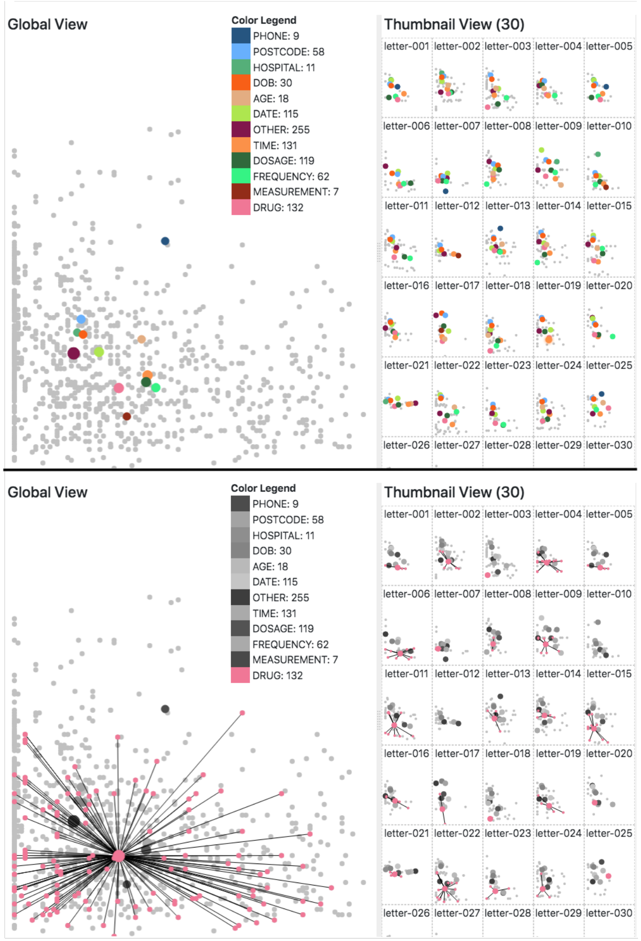

Global View
Color Legend
PHONE: 9
POSTCODE: 58
HOSPITAL: 11
DOB: 30
AGE: 18
DATE: 115
OTHER: 255
TIME: 131
DOSAGE: 119
FREQUENCY: 62
MEASUREMENT: 7
DRUG: 132
Thumbnail View (30)
letter-001 letter-002 letter-003 letter-004 letter-005
letter-006 letter-007 letter-008 letter-009 letter-010
letter-011 letter-012 letter-013 letter-014 letter-015
letter-016 letter-017 letter-018 letter-019 letter-020
letter-021 letter-022 letter-023 letter-024 letter-025
letter-026 letter-027 letter-028 letter-029 letter-030

Fig. 2 An illustration of centroid exploration in the global and thumbnail views. The top shows the default search and rendering when loading 30 letters. By default, the global view searches and renders 12 text data categories in focus and renders individual data samples in greyscale, a classical focus+context approach. A thumbnail is also presented with the same default rendering options for individual letters. The bottom shows the edges connecting the user-chosen(focus) centroid, *DRUG*, and individual samples in both views. Other centroids and individual samples are rendered as context. Edges can also be hidden as an option.

4.2 Advanced Visual Filtering and Selection

LetterVis facilities data exploration via a visual querying interface combined with rendering and sorting options. We also color-mapped visual elements and provide further interaction via color legends.

Matrix View: We include a co-occurrence matrix specifically for visualizing AED co-prescriptions, as a special requirement requested by our domain expert partners to support the exploration of common and unusual AED co-prescriptions (**T3**). Co-occurring AEDs appear as color-coded cells where a row and a column intersects in the matrix view (Fig. 1B). We extract AEDs in real time. By exploring common and unusual AED co-prescriptions visually, they can potentially reduce the number of trials needed for finding optimal AED co-prescriptions for patients. The matrix view (Fig. 1B) is automatically rendered when letters are loaded. Popular co-prescriptions are trivially observed when the matrix view is sorted by co-occurrence frequency. Co-prescriptions with a higher frequency are also rendered in more distinct colors than others. Hovering over a cell initiates an arrow from the y-axis to the x-axis connecting the corresponding pair of AEDs and also highlights both AEDs (**T5**). See Fig. 1B.

Drug Chain View: The matrix view only indicates co-occurrence of pairwise drugs. When a query involves multiple AEDs, the drug chain view (Fig. 1C) is rendered. In the drug chain view, blocks representing multiple AEDs are linked in order of appearance in the corresponding letter. This view aims to provide a visual representation of prescription progression and may unveil unique insight into epilepsy progression as well(**T3**). User-chosen AEDs are shown in color(focus) while the remaining AEDs are rendered as context. Chains can be interactively aligned between letters and sorted as shown in Fig. 3 (**T5**). Unusual chains immediately stand out. In Fig. 3B, a user selects Lamotrigine to specify it as the base AED. All remaining chains are then aligned by the first appearance of the base AED. Chains with no matching AED are rendered as context. See Fig. 3E.

Advanced Visual Queries: LetterVis supports a subset of Apache Lucene [23] syntax, namely boolean operators and grouping. The incorporated boolean operators provide users with the flexibility to include or exclude keywords from the result.The color legend is then updated to include random colors assigned to each keyword in the query history, an example is shown in Fig. 5. The implementation aims to provide the visual means for the user to query and filter letters (**T4**). A query can be constructed multiple ways:

- Clicking on cells in the matrix view will execute a query formed as ‘AED on the y-axis *AND* AED on the x-axis’
- The user can drag categories from the color legend to the search bar
- Under ‘Click and Search’ mode, the user can select any centroid from the global view to populate the search bar

We store a list of user-specified search queries to support undo and redo functions (**T6**. See Fig. 1A right). The user can use the checkbox located in front of each query to toggle the visibility of the corresponding query and its results.

Sorting Options: We provide 11 options to sort individual letters in the thumbnail view and the abstraction of AEDs in the drug chain view (**T5**). Sorting letters by a user-chosen dimension gives users the control they need to find patterns and outliers quickly. For example, outliers will stand out with long edges in the global view and thumbnail views. In addition, cells in the matrix view can be sorted alphabetically or by co-occurrence. We demonstrate the benefit of having a wide selection of sorting options in our case studies in Sec. 5.

5 Evaluation

Our evaluation comprises of three case studies, described in detail in Sec. 5.1 and feedback from domain experts in Sec. 5.2. Section 5.3 includes reviews written by two health data analysts.

5.1 Case Studies

Each case study is motivated based on the discussions with domain experts in EHR analysis. The first case study aims to identify commonly and rarely co-prescribed AED combinations. The second case study tests the ability to find patient outliers. The third case study explores the relationship between pregnancy and AEDs. All case studies are based on 200 anonymized clinic letters described in Sec. 2.

Case Study 1 - Identifying Common and Unusual AED Co-prescriptions: Clinicians are generally confident in prescribing the most suitable first drug, however, co-prescribing a second drug is always challenging. Visualizing the common and unusual co-prescriptions may help the clinician with the decision and potentially reduce unnecessary co-prescription trials needed on patients.

After loading all letters, a matrix view is automatically generated with 18 AEDs, shown in Fig. 1B. The matrix by default is sorted alphabetically by AED co-occurrence. We then sort the matrix view by frequency on both axes from top-left to bottom-right. We immediately observe the top two AED co-occurrences near the cluster at the upper-left corner, Levetiracetam-Lamotrigine (Fig. $1B^1$, 45 co-occurrences) and Levetiracetam-Sodium Valproate (Fig. $1B^2$, 37 co-occurrences). We select these two pairs of AEDs and obtain 51 letters. This effectively filters the data with the following query: (Levetiracetam *AND* Lamotrigine) *OR* (Levetiracetam *AND* Sodium Valproate). Fig. 1D and E show the corresponding AEDs in the global and thumbnail views. Centroids

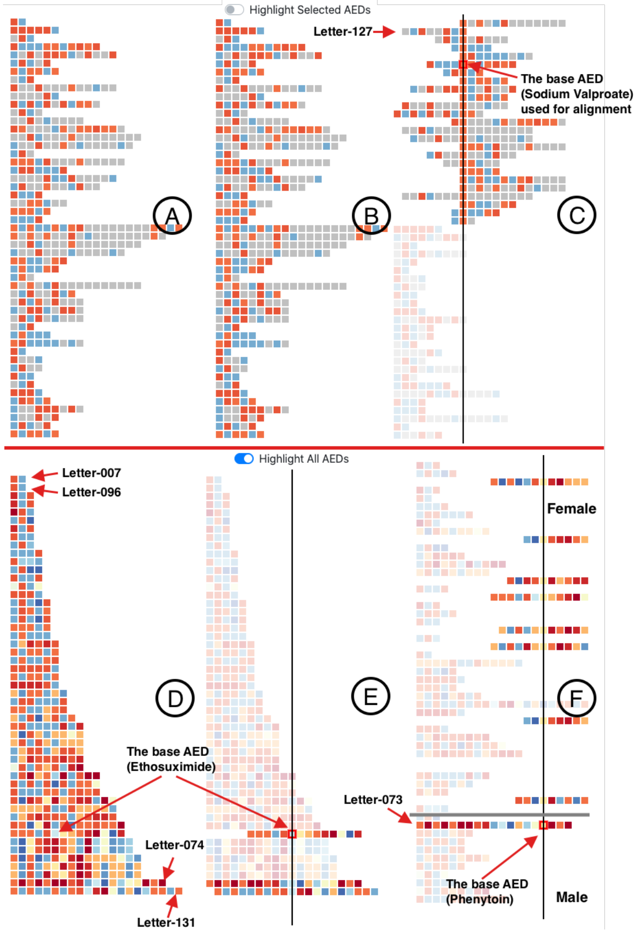
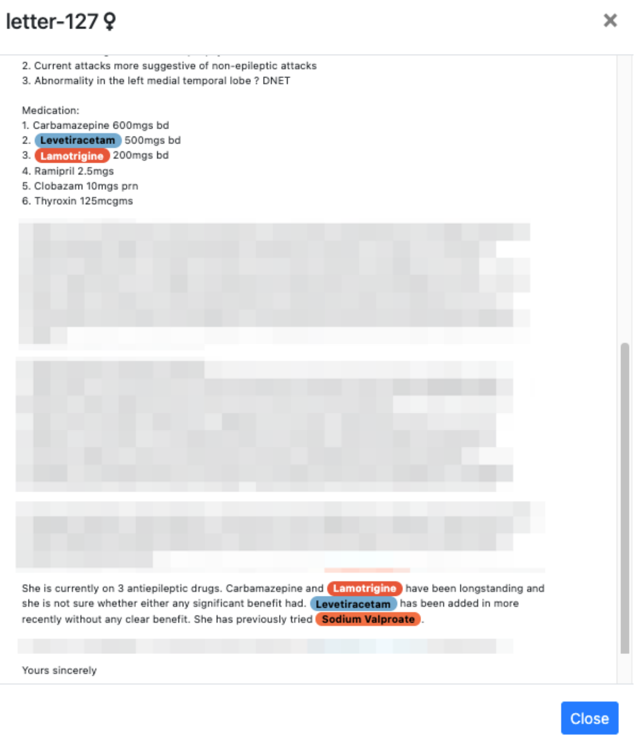


Fig. 3 Illustrations of letter alignment in the drug chain view. A) The initial layout of the drug chain view. B) Letters are aligned via clicking on a base AED, Lamotrigine, highlighted with a red border in the first letter. Letters without the AED are shown in context with reduced opacity. C) Letters are sorted by alignment, with context letters being shifted to the end of the queue. Hovering over any block will display a tooltip containing the letter title and the AED name. D) All AEDs are put in focus mode (in color) via a toggle. Chains are then sorted by the number of AEDs. E) Ethosuximide, a rare prescription in the dataset, is selected as the base AED for alignment. F) Chains are sorted by gender, with a horizontal grey bar as the separator. The same subset of letters is used in this figure.

representing each AED are connected by an edge if they are connected by an *AND* operator in the query.

We then sort the letters by total edge length, the sum of distances between edges in each letter. In the thumbnail view, we observe that the centroids are more sparsely placed in letters such as letter-022 and 054 than their peers. According to domain experts, this often indicates that AEDs appearing in them are not co-prescriptions but previous medications or new recommendations. Centroids in letters such as letter-073, 133 and 174, are closely co-located, which often represents co-prescriptions. We inspect these letters manually in the focus view. The finding confirms these hypotheses.



letter-127 ♀

2. Current attacks more suggestive of non-epileptic attacks
3. Abnormality in the left medial temporal lobe ? DNET

Medication:
1. Carbamazepine 600mgs bd
2. Levetiracetam 500mgs bd
3. Lamotrigine 200mgs bd
4. Ramipril 2.5mgs
5. Clobazam 10mgs prn
6. Thyroxin 125mcgms

[redacted]

[redacted]

[redacted]

She is currently on 3 antiepileptic drugs. Carbamazepine and Lamotrigine have been longstanding and she is not sure whether either any significant benefit had. Levetiracetam has been added in more recently without any clear benefit. She has previously tried Sodium Valproate.

[redacted]

Yours sincerely

Fig. 4 A screenshot of one letter's original view(cropped). Search query terms are highlighted in their corresponding colors, as described in Sec. 4.1.

Case Study 2 - 5 Different Ways to Find Outliers: In this case study, we examine LetterVis' ability to identify letter outliers. A letter with abnormal patterns often carries valuable information that requires the analysts' attention. An outlier may also indicate an error.

In Fig. 3B, we select Sodium Valproate, due to its popularity in co-prescriptions, in the first letter to align the drug chain view. This sets it as the base AED for alignment. All remaining letters are then aligned by the first appearance of the base AED. Letters without the AED are rendered as context. See Fig. 3B and Fig. 3E. We then sort the letters by alignment for further filtering in Fig. 3C. **(1)** We are able to identify one outlier immediately, letter-127, as it is the only chain ending with Sodium Valproate. We inspect letter-127 in its focus view (Fig. 1F) which indicates Sodium Valproate was recently prescribed. According to E1, *"this is abnormal as Sodium Valproate is ususally the first or second AED for epilepsy patients."* We expand the focus view to show the entire letter (Fig. 4). We discover that multiple AEDs have been prescribed to the female patient with no clear benefit.

We then sort the letters by the number of AEDs (Fig. 3D) in them and explore letters on both sides of the sort spectrum. **(2)** Letter-007 and 096 mention only two AEDs, we observe that the patients in the aforementioned letters are suspected cases awaiting further diagnosis. Whereas letter-074 (19 AEDs) and 131 (21 AEDs) represent confirmed patients with a history of epilepsy of more than 15 years.

Ethosuximide, shown in light blue in Fig. 3D, has significantly fewer appearances than others. **(3)** We align the drug chain view by Ethosuximide (Fig. 3E), discover that it is only co-prescribed to three patients (letter-060, 074 and 131) that are on Levetiracetam with Lamotrigine or Sodium Valproate. We view these three letters in the focus view and identify vomiting as an adverse effect caused by Ethosuximide for the patient in letter-074. The other two patients experience no benefit from Ethosuximide.

When chains are sorted by gender (Fig. 3F), **(4)** we find letter-073, the only male patient (24 years old with over 23 years history of epilepsy) in the cohort that has been prescribed to Phenytoin. The patient is also the only male who has been prescribed Lacosamide.

Individual samples in the global view can also be used for finding outliers.**(5)** We explore two outlier patients (see red arrow in Fig. 1D). While both patients prescribe to Lamotrigine, the patient in letter-186 (top-left corner) is tapering it off as it has no clear benefit in containing seizures. On the contrary, the patient in letter-104 (top-right corner) is building up the dosage as a replacement for Carbamazepine.

Case Study 3 - AEDs and Pregnancy: In this case study, we evaluate LetterVis' ability to identify and explore AEDs prescribed to patients with planned or ongoing pregnancy. Avoiding AED adverse effects is an important research topic. By using LetterVis' advanced visual interface, the user can combine any keywords to study their associations and unveil insightful patterns. The case study is inspired by the research on the effects of AEDs in pregnancy [12], the research indicates in-utero exposure to Sodium Valproate is likely to negatively affect a child's cognitive ability, while Lamotrigine and Carbamazepine have little or no impact.

We first construct the query '(pregnant *OR* pregnancy) *AND* DRUG' to filter out 182 letters. The query highlights both term *pregnant* and *pregnancy* with all of 50 AED names supplied by our domain expert partners. We then apply sorting by gender to obtain the thumbnail view shown in Fig. 5.

Letter-011 indicates a high risk case, a 52-year-old patient who suffered four seizures in eight months, is planning for pregnancy. In this special case, a higher than usual dose of Folic Acid is prescribed to help prevent birth defects. In letter-060, the physician proposed multiple AED co-prescriptions to gradually replace Sodium Valproate, in order to prepare the patient for pregnancy. In letter-144, Sodium Valproate is showing a remarkable effect in reducing seizure frequency for the patient. Because she has no pregnancy planned, the physician decided to increase the dosage.

Letter-093 contains a special case where a female is suspected to suffer from non-epileptic psychogenic seizures, hence common AEDs such as Levetiracetam, Lamotrigine and Sodium Valproate were never prescribed.

During the process, we also discover two identical letters (letter-103 and 107) using different pseudo names, this is likely due to human error during the manual

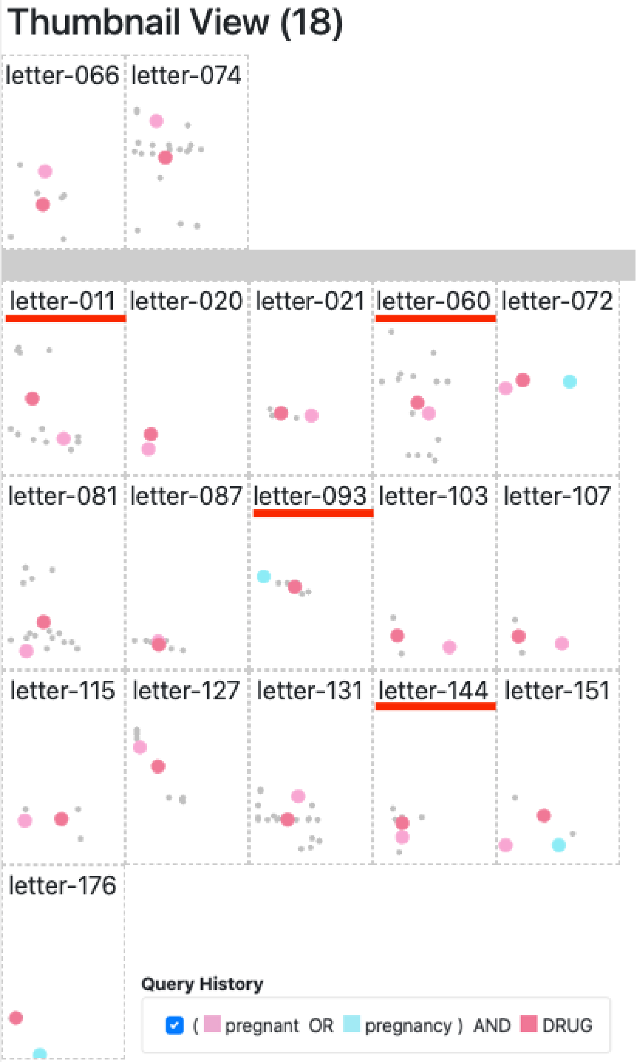


Fig. 5 We execute the query '(pregnant *OR* pregnancy) *AND* DRUG' and sort letters by gender to focus on pregnancy. The thumbnail view shows two male patients are separated from the rest by a grey bar. High risk cases discovered in Case Study 3 are highlighted with a red line.

anonymization process. Please see the accompanying video for a demonstration of these case studies.

5.2 Domain Expert Feedback

We regularly demonstrate LetterVis to our domain expert partners (E1 and E2) to guide the development and present intermediate results. We also provide a live version available online for them to explore. We provide excerpts of their feedback below. In general, our collaboration process adheres to the Visual Information-Seeking Mantra [21]: 1) we demonstrated the global view that shows an *overview* of the data to experts, 2) they then requested to *zoom and filter* the outliers found in the global view. This is fulfilled by the thumbnail view, 3) eventually *details were demanded* to verify any findings from previous stages, through both the focus view and chain view.

During our first demonstration of LetterVis with E2, we demonstrated the coordination between the global, thumbnail and focus views, the expert immediately commented, *"That's interesting, if you can spot pregnancy and certain drugs are in close proximity, you immediately want to read that letter because something is not right. I can see it's been really useful"* (**R2**). E2 pointed out the limitation of our simple numerical approach described in Sec. 4, that the user is unable to query for words. The expert was also particularly interested in seeing a view that's specifically tailored for visualizing AEDs, with the ability to use syntax-based search queries to improve exploration of items of interest (**R4**). We implement his recommendation, Apache Lucene, into our next version (**R7**, **T5**).

In our second feedback meeting with E1 and E2, immediately after feature introduction, both experts were able to picture a use case for identifying outliers by using deviated centroids in the global and thumbnail views, E2 stated that, *"If you are able to see centroids representing AEDs in one letter deviated far more than the global trend, we don't trust this letter, we might need to investigate the prescriptions in that particular letter"* (**R2**). Both experts were keen on visualizing AED co-prescriptions, the demonstration spent 30 minutes on discussing this topic. E1 pointed out that, *"One useful use case I can imagine is to visualize combinations of AEDs for different patients and even how often one AED is mentioned together with another. It's really hard to conduct clinical trials for more than one AED, clinicians usually know what the best first AED is, but not the second. Designing a trial for that is nearly impossible"* (**R8**). E2 further elaborated that, *"Some patients on multiple AEDs might have multiple seizures per week, but when they are given only one AED, their seizure frequency might be reduced to one per week. One existing tool we are using still relies on the command line to operate, so this dashboard-like visualization tool can be really helpful"*.

We introduce the matrix (Fig. 1B) and drug chain view (Fig. 1C and Fig. 3) for visualizing AED co-prescriptions and prescription evolutions (**T2**). We demonstrated these two visual designs to E1 and E2 during our third and fourth feedback sessions. E2 commented that, *"This is definitely useful, the way we are currently doing is really laborious, we have to go through the patient's EHRs and sequentilly look at what AEDs were given at each visit. The drug chain view is looking at the problem we need to solve"* (**R8**). When the drug chain view was sorted by the number of AEDs and aligned the chains by a rare AED, E1 was immediately able to identify an outlier patient that is on an unusual AED co-prescription, *"Pregabalin and Retigabine are a very strange combination, I never thought of searching for that, I'm definitely going to look at that patient"*.

5.3 Domain Expert Review

The following written feedback was provided directly by the expert health data analysts (E1 and E2).

"LetterVis presents a novel way to visualize trends and potential outliers across sets of clinic letters. Unstructured texts have not been traditionally used as a data source for analysis in healthcare, as the data is not available in a readily parse-able format such as structured data. Recent advances in the field of Natural Language Processing have yielded NLP methods to extract structured data from clinical prose [17], where more traditional analyses can take place. LetterVis employs NLP and data visualization techniques to help isolate and communicate important trends to the user.

Many clinical decisions can be improved by the analyses that LetterVis offers. For example, there is limited evidence on the best anti-epileptic drug (AED) combination to use for patients with severe epilepsy [18]. The co-occurrence matrix can visualize AED combinations across a range of frequencies to potentially identify the most common and/or stable AED combinations. Less frequently used combinations can also be easily identified. A limitation of the co-occurrence matrix is addressed in the drug chain view – namely the ability to visualize patients that are prescribed more than two AEDs, and it is immediately clear which patients may be problematic based on the number of different AEDs appearing in the chain.

By loading AEDs vs side effects into the co-occurrence matrix it is possible to view associations between AED and side effects. The thumbnail view adds more granularity to this analysis by presenting the proximity of side effects to the AED in question. In 2018 the Medicines & Healthcare products Regulation Agency strengthened their position on the avoidance of valproate to be used in women and girls, especially during pregnancy [15]. It is therefore important to determine the likelihood that letters mentioning both pregnancy and valproate fall into categories around the education of valproate in pregnancy, or if valproate is still being used during pregnancy. The thumbnail view is an ideal tool to hone in on letters that may show the latter case because currently prescribed AEDs are usually found at the beginning of the letter, where pregnancies will be mentioned later on in the document.

Some future work might include the integration of medical ontologies to rapidly build code lists of interest that can be used in the already highly configurable Lucene-based query capabilities. Given that context is very important when using search terms i.e. negation or hypothetical discussions around AED side effects, adopting more advanced NLP techniques or integrating with existing technology such as GATE [3] or cTAKEs [24] would help increase user confidence in any trends that are presented to them. LetterVis could potentially expand to present entire timelines for individual patients and not be constrained to analyses within one letter. For example, the chain view could be used to align newly diagnosed patients and their first AED to determine popular first line drugs, and how that changes over time, or monitoring side effects and seizure frequency when a new AED is detected in subsequent letters.

LetterVis is well positioned to take advantage of the emergence of unstructured data being used for healthcare research, and its methods will offer clinicians the vital tools they need to see the big picture across potentially millions of clinic letters."

Supplementary Video: We demonstrate LetterVis and the case studies in a supplementary video: `https://youtu.be/jSVzhCjLi_U`.

6 Limitations and Future Work

Our work is limited by the dataset size. We are currently pending approval to utilize a much larger dataset (with thousands of letters) from a NHS Healthboard. We then plan to address scalability issues. Our current approach depends on the local computational power and the resources are restricted by the browser. Performance may be slower when the data exceeds a certain size. A cloud-based implementation with more processing power than a local browser can address this. However, due to the sensitive nature of EHR data, namely privacy and security, this may impose impediments for a cloud-based approach.

Our current NLP approach is based on a list of rules supplied by domain experts. One of the next research goals is incorporating advanced NLP techniques. This may enhance the accuracy of extraction as well as enrich the text categories. The current approach requires manual querying for analysing AED side effects and interactions. Extracting more text categories can potentially automate this process.

The drug chain view only shows the sequence of AEDs mentioned in a letter. This is not necessarily a co-prescription. Human intervention is required to verify if a co-prescription is present. Future NLP algorithms may be able to assist with this, however, human verification will always be a requirement (**R3, R5, T3**).

7 Conclusions

In this paper, we present a novel visualization tool, LetterVis, to support the analysis of clinic letters through advanced interactive visual designs and queries. The work aims to support EHR researchers to explore free text EHR data and address their research hypotheses in a transparent and explainable manner. The strength of this work is the novel concept of letter-space and how it is applied to a real-world problem. Through our collaboration with three domain experts, we identify and address a selection of important tasks via customized letter-space designs and interactive user options. We incorporate NLP techniques to pre-process clinic letters with a list of extraction rules curated together with our domain expert partners. We then develop an advanced visual query interface including five customized visual designs to support the analysis of EHR free text data. We demonstrate LetterVis with three empirical use cases inspired by real world scenarios. In depth evaluations are also conducted with domain experts.

References

1. Bernard, J., Sessler, D., Bannach, A., May, T., Kohlhammer, J.: A visual active learning system for the assessment of patient well-being in prostate cancer research. In: Proceedings of the 2015 Workshop on Visual Analytics in Healthcare - VAHC '15, vol. 25-October, pp. 1–8. ACM Press, New York, New York, USA (2015). DOI 10.1145/2836034.2836035
2. Brehmer, M., Munzner, T.: A Multi-Level Typology of Abstract Visualization Tasks. IEEE Transactions on Visualization and Computer Graphics **19**(12), 2376–2385 (2013). DOI 10.1109/TVCG.2013.124
3. Cunningham, H., Tablan, V., Roberts, A., Bontcheva, K.: Getting More Out of Biomedical Documents with GATE's Full Lifecycle Open Source Text Analytics. PLoS

Computational Biology **9**(2), e1002,854 (2013). DOI 10.1371/journal.pcbi.1002854

4. Glueck, M., Gvozdik, A., Chevalier, F., Khan, A., Brudno, M., Wigdor, D.: PhenoStacks: Cross-Sectional Cohort Phenotype Comparison Visualizations. IEEE Transactions on Visualization and Computer Graphics **23**(1), 191–200 (2017). DOI 10.1109/TVCG.2016.2598469
5. Glueck, M., Hamilton, P., Chevalier, F., Breslav, S., Khan, A., Wigdor, D., Brudno, M.: PhenoBlocks: Phenotype Comparison Visualizations. IEEE Transactions on Visualization and Computer Graphics **22**(1), 101–110 (2016). DOI 10.1109/TVCG.2015.2467733
6. Glueck, M., Naeini, M.P., Doshi-Velez, F., Chevalier, F., Khan, A., Wigdor, D., Brudno, M.: PhenoLines: Phenotype Comparison Visualizations for Disease Subtyping via Topic Models. IEEE Transactions on Visualization and Computer Graphics **24**(1), 371–381 (2018). DOI 10.1109/TVCG.2017.2745118
7. Gramazio, C.C., Laidlaw, D.H., Schloss, K.B.: Colorgorical: Creating discriminable and preferable color palettes for information visualization. IEEE Transactions on Visualization and Computer Graphics **23**(1), 521–530 (2017). DOI 10.1109/TVCG.2016.2598918
8. Gunter, T.D., Terry, N.P.: The Emergence of National Electronic Health Record Architectures in the United States and Australia: Models, Costs, and Questions. Journal of Medical Internet Research **7**(1), e3 (2005). DOI 10.2196/jmir.7.1.e3
9. Hogan, T., Hinrichs, U., Hornecker, E.: The Elicitation Interview Technique: Capturing People's Experiences of Data Representations. IEEE Transactions on Visualization and Computer Graphics **22**(12), 2579–2593 (2016). DOI 10.1109/TVCG.2015.2511718
10. Iakovidis, I.: Towards personal health record: current situation, obstacles and trends in implementation of electronic healthcare record in Europe. International Journal of Medical Informatics **52**(1-3), 105–115 (1998). DOI 10.1016/s1386-5056(98)00129-4
11. Koleck, T.A., Dreisbach, C., Bourne, P.E., Bakken, S.: Natural language processing of symptoms documented in free-text narratives of electronic health records: A systematic review. Journal of the American Medical Informatics Association **26**(4), 364–379 (2019). DOI 10.1093/jamia/ocy173
12. Lacey, A.S., Pickrell, W.O., Thomas, R.H., Kerr, M.P., White, C.P., Rees, M.I.: Educational attainment of children born to mothers with epilepsy. Journal of Neurology, Neurosurgery & Psychiatry **89**(7), 736–740 (2018). DOI 10.1136/jnnp-2017-317515
13. Liddy, E.: Natural Language Processing (2001)
14. McNabb, L., Laramee, R.S.: Survey of Surveys (SoS) - Mapping The Landscape of Survey Papers in Information Visualization. Computer Graphics Forum **36**(3), 589–617 (2017). DOI 10.1111/cgf.13212
15. Medicines and Healthcare products Regulatory Agency: New measures to avoid valproate exposure in pregnancy endorsed. Tech. rep. (2018). URL `https://www.gov.uk/government/news/new-measures-to-avoid-valproate-exposure-in-pregnancy`
16. MIT Critical Data: Secondary Analysis of Electronic Health Records. Springer International Publishing, Cham (2016). DOI 10.1007/978-3-319-43742-2
17. Névéol, A., Zweigenbaum, P.: Clinical Natural Language Processing in 2014: Foundational Methods Supporting Efficient Healthcare. Yearbook of Medical Informatics **24**(01), 194–198 (2015). DOI 10.15265/IY-2015-035
18. Pickrell, W.O., Lacey, A.S., Thomas, R.H., Lyons, R.A., Smith, P.E., Rees, M.I.: Trends in the first antiepileptic drug prescribed for epilepsy between 2000 and 2010. Seizure **23**(1), 77–80 (2014). DOI 10.1016/j.seizure.2013.09.007
19. Rind, A., Wang, T.D., Aigner, W., Miksch, S., Wongsuphasawat, K., Plaisant, C., Shneiderman, B.: Interactive Information Visualization to Explore and Query Electronic Health Records. Foundations and Trends® in Human–Computer Interaction **5**(3), 207–298 (2013). DOI 10.1561/1100000039
20. Rostamzadeh, N., Abdullah, S.S., Sedig, K.: Visual Analytics for Electronic Health Records: A Review. Informatics **8**(1), 12 (2021). DOI 10.3390/informatics8010012
21. Shneiderman, B.: The eyes have it: a task by data type taxonomy for information visualizations. In: Proceedings 1996 IEEE Symposium on Visual Languages, pp. 336–343. IEEE Comput. Soc. Press (1996). DOI 10.1109/VL.1996.545307
22. Sultanum, N., Singh, D., Brudno, M., Chevalier, F.: Doccurate: A Curation-Based Approach for Clinical Text Visualization. IEEE Transactions on Visualization and Computer Graphics **25**(1), 142–151 (2019). DOI 10.1109/TVCG.2018.2864905
23. The Apache Software Foundation: Apache Lucene (2015). URL `https://lucene.apache.org/`
24. The Apache Software Foundation: Apache cTAKES™ - clinical Text Analysis Knowledge Extraction System (2018)
25. Trivedi, G., Pham, P., Chapman, W.W., Hwa, R., Wiebe, J., Hochheiser, H.: NLPReViz: an interactive tool for natural language processing on clinical text. Journal of the American Medical Informatics Association : JAMIA **25**(1), 81–87 (2018). DOI 10.1093/jamia/ocx070
26. Vellido, A.: The importance of interpretability and visualization in machine learning for applications in medicine and health care. Neural Computing and Applications **0123456789** (2019). DOI 10.1007/s00521-019-04051-w
27. Vollmer, S., Mateen, B.A., Bohner, G., Király, F.J., Ghani, R., Jonsson, P., Cumbers, S., Jonas, A., McAllister, K.S.L., Myles, P., Grainger, D., Birse, M., Branson, R., Moons, K.G.M., Collins, G.S., Ioannidis, J.P.A., Holmes, C., Hemingway, H.: Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness. BMJ **368**, l6927 (2020). DOI 10.1136/bmj.l6927
28. West, V.L., Borland, D., Hammond, W.E.: Innovative information visualization of electronic health record data: A systematic review. Journal of the American Medical Informatics Association **22**(2), 330–339 (2015). DOI 10.1136/amiajnl-2014-002955
29. Zhang, Z., Ahmed, F., Ramakrishnan, A.M.I.V., Zhao, R., Viccellio, A., Mueller, K.: AnamneVis: a framework for the visualization of patient history and medical diagnostics chains. IEEE VAHC Workshop (January), 1–4 (2011)