Visualizing Clinical Data of Patients at the Child and Adolescent Psychiatric Emergency Unit:
Project Proposal

John-Jose Nunez, Mona Fadaviardakani

—————————— ✦ ——————————

1 INTRODUCTION

Our project seeks to visualize the clinical data of patients admitted to the Child and Adolescent Psychiatric Emergency Unit (CAPE) at BC Children's Hospital (BCCH). This will allow both exploring as well as answering specific questions that psychiatrists, managers, and researchers may have, which will in turn allow better understanding of our patients and ultimately, provide better care. From our understanding, this will be the first visualization project of its kind especially with the objective of seeking to visualize this type of data, likely due to the difficulty of obtaining such data due to privacy concerns.

Our project team includes a part-time computer science graduate student/resident physician specializing in psychiatry who has domain expertise in this data, and a full-time graduate student also from the computer science department. The resident physician, John-Jose has been involved with a larger project utilizing this data, as part of his general research interest in applying informatics and data science to psychiatric data. He has experience working on psychiatric wards similar to CAPE, and with dealing with this data in a clinical context. He also has experiencing using this sort of clinical data in a research context, including with applications utilizing machine learning and natural language processing. The computer scientist, Mona has experience working with a variety of data types, and programming languages like python, and R which are widely used for data analysis and visualization.

2 DATA AND TASK

2.1 Domain/Who

The domain of our study is Child and Adolescent Psychiatry, which is the sub-specialty of psychiatry working with patients generally younger than 18 years of age. The intended users of our visualization include:

- Unit/hospital managers: To explore patterns in data to better predict demand, staffing, and resources.

    E.g. Are there times of the year we may need more staffing to cope with increased hospital admission?

- Researchers: To investigate patterns and how this local example compares to the child and adolescent psychiatry literature, and to observe correlations to generate hypothesis of causal relationships.

    E.g. Does the CAPE unit follow established patterns of more mania in the spring, and depressive illness in the fall?

- Clinicians: To understand patient characteristics and treatment patterns in the unit to improve their own care.

—————————

- *John-Jose Nunez E-mail: jjnunez@cs.ubc.ca*
- *Mona Fadaviardakani E-mail: mfadavi@cs.ubc.ca*

E.g. Are my medication choices similar to what is used in the aggregate on the unit?

2.2 Task/Why/Actions

For our project, we wish to identify, summarize, and compare aggregated patient data. We want to:

- Analyze: Primarily the task will be to discover and analyze patient data from admissions over a period of 40 months. This includes both temporal and non-temporal trends. The tool me also be used to present data, though we envision it as more for discovery.

- Search: Our vis will need to facilitate three of the four search functions. Users may have a specific patient subset they wish to know more about (lookup). Or the may want to know more about patients with a certain temporal or other characteristic (browse). However, exploring may be one of the biggest contributions of the tool. Facilitating discovery will be the biggest contribution from a vis perspective, and may present the most value-adding if it allows users to discover new information that can be useful in the clinic or for research. The tool will not support displaying the data of a specific patient (locate) and in fact must not do this due to privacy requirements.

- Query: This will be an important part of the task. Users may want to identify groups of items (patients) including various attributes. Similarly, summarizing various attributes will be an integral feature of the task, allowing users to summarize various attributes of various groupings.

- Clustering: Comparing is also a desired task. This will include allowing users to compare attributes of different selections. However, clustering patients by attributes may also be useful. Clusters may make comparisons easier, and could also have possible uses for users, such as identifying which patient groups will benefit from certain interventions.

2.3 Data/What

- Items: We have 243 items in the data set, where each item is a patient.

- Attributes: Every patient will have a consistent set of attributes, including:

    - Time: This represents the associated time of admission for each patient, and it is likely at a week or month level. Time will be an important attribute to consider for noticing trends and variances. Note that an item does not change over time, but items have different times of presentation.

    - Gender: categorical

    - Age: quantitative, only integers

    - Postal code: position. At this time we will not have access to this attribute, but may later on in the project.

– Medications: This data would be categorical and hierarchical: medications are part of sets of medications of similar mechanisms, which are in turn parts of larger groups of medications with similar purposes. For example, Fluoxetine is an element of Selective Serotonin Reuptake Inhibitors (SSRI), which are elements of antidepressants.

– History: This includes Psychiatric history (prior diagnoses, previous admissions), Medical history (diagnoses, surgeries), Substance use history, Social history (family structure, foster care). This data is mostly categorical, but there is sometimes both ordering and hierarchy involved. For instance, "suicidal content" could contain various past events related to suicidality of increasing intensity such as passive suicidality (thoughts such as "I'd be better off dead"), active suicidality (thoughts such as "I have a plan to suicide") and suicide attempts. Some aspects may also be quantitative, for instance frequency or amount of alcohol consumed.

– Diagnoses: similar to above, categorical data entailing the final diagnosis upon discharge. Again, can be hierarchical.

– Symptoms: there are multiple attributes pertaining to symptoms in a specific diagnostic subfield, for instance depressive symptoms. Each attribute could be a set of categorical possibilities. We may want to display this granularity, or treat these as a binary or possibly even ordinarily if we can rank the categorical symptoms in terms of severity.

– Various binary attributes: examples include previous admission, previous suicide attempt.

– Discharge location: categorical, e.g. home, other hospital ward, other institution.

• Notes: as we have a lot of data, we may end up not visualizing everything, or may combine or simplify many features. Currently, much of the categorical data is in free text, but are encoded by SNOMED clinical terms, a hierarchical medical ontology. This will need to be converted into readable terms. But, it also presents an interesting aspect of the data, as the hierarchical nature of SNOMED might lend itself to interesting visualizations.

3 PROPOSED INFOVIS SOLUTION

3.1 Potential Idioms and Design Choices

Idiom: Line graph

• This is used to show trends over time, such trends can include hospital admission rates by each month over the span of 40 months, suicidal rates by each month over the span of 40 months, or specific illness-related admission rates (for example depression admission rates)

Idiom: Pie chart

• This will be used to show aggregated data, instead of over time. Pie charts are commonly used in medicine, so users may expect them. Additionally, much of the data being visualized is often interpreted as percentages of a whole, leading to a natural fit with this idiom.

Idiom: Bar chart

• This is used to show trend similarities and/or differences between patient groups, how hospital admission rates vary with seasons or school terms

Idiom: Nest Tree Diagram(s)

• This is used to toggle the patient sub-types we want to visualize data for or want to be visualizing data of

There could also likely be some interesting uses of color as a form of channel because a lot of our data is categorical and have some hierarchy. For example, consider the example where we use a pie chart to show different psychiatric diagnoses and their severity:

• Hue can be used for the diagnosis

– blue could stand for depressive disorders
– red could stand for bipolar disorders (both types 1 and 2)

• Luminance can be used for distinguishing between the types of an illness that is within the same category

– a lighter shade of red could represent the less severe bipolar disorder type 2
– a darker shade of red could represent more severe bipolar disorder type 1

• Manipulate view

– An essential feature of the vis will be to allow users to interact to manipulate view. This includes selecting items temporally, selecting items based on attributes (e.g. diagnosis, demographics), and selecting which attributes are being displayed. Semantic zooming, allowing users to focus in on a subset of patients or zoom out to view data more broadly, will likely be a helpful design choice.

into Multiple Views

– Supporting multiple views will also be essential. Multiform views will allow users to view data both temporally (such as via a line graph) but also in aggregate (such as with a pie chart). Small multiple views may help users investigate characteristics of different data sub-types. Linked highlighting between all views may reduce the cognitive load of users. Superimposition may also be useful, to how different attributes at the same time, e.g. the gender and diagnoses.

Since one of our goals is performing clustering, we might benefit from derived data like cluster hierarchy which shows trends for different clusters over time. In order to reach this goal, we can use juxtapose multiple views of calendar and superimposed 2D curves.

3.2 Sample Scenarios

• A psychiatrist may want to know why there are a spike of admissions in May and December. She selects only looking at admissions in these months, then selects to look at how different school grades make up this population, visible in a pie chart. She notices that grade 12 students make up a disproportionate amount, and confirms a suspicion that it may have something to do with high stakes exams.

• A suicide researcher knows the various demographics of who attempts suicide according to large population studies, but wants to know how things are in the CAPE unit. She selects only looking at suicidal patients and then selects various attributes to visualizing, looking at how age, gender, diagnoses make up this population. She wants to investigate various subsets, and finds out what of male patients presenting with suicidality, a large proportion suffer from a psychotic illness; this inspires a new research project investigating why males with psychotic illnesses have so much suicidality in our population.

3.3 Obfuscation

• The purpose of this project is to view general, aggregated data for the patients, not to view the details of one specific patient. It will be essential that the vis cannot be used for the latter, as that would breach our ethics and privacy agreement. However, interactivity allowing users to search for patients with many attributes could allow a single user to be viewed. As such, we will need to obfuscate when viewing a very small subset of patients (such as less than 4). This will require additional reading and planning, but potential solutions are to slightly randomize data when viewing only a few, or showing a certain glyph when a resulting value is small, such as "< 5"

4 PROPOSED IMPLEMENTATION

The raw data is being provided in a CSV format. It will be loaded into a pandas data frame [8] in Python 3.7. Some data processing will need to be done in order to be readily accessible, such as converting brand names of drugs to their generic names. We will use pandas' powerful data processing tools, in addition to other Python libraries, to do this efficiently in a modular, reproducible fashion.

We will then use the Plotly Python Open Source Graphing Library for the visualization. This library is built upon D3. However, its python implementation will help shallow the D3 learning curve, and will allow easy integration with the data processing part of the code. A purely python implementation will also better fit with the larger research group's current pipeline, which is also written in python. We also plan to use D3.js if we have enough time to make our visualization approach more interactive or to utilizie more complex features.

5 MILESTONES AND SCHEDULE

October 28 Preliminary Work

- (4 hours JJ) Prepare project pitches and initial discussion of data
- (5 hours JJ) At least six meetings to discuss project and parameters of using data
- (4 hours Both) Completing required ethics/privacy tutorials and

November 4: Project proposal

- (10 hours Both) Develop and edit project proposal
- (5 hours Both) Initial literature review and developing previous work section

November 12: Preliminary data exploration and vis/pipeline development

- (4 hours JJ) Preparation of git, and deployment of simple demonstration using our data with pandas and python plotly library
- (3 hours JJ) Developing a privacy-compliant dataset to work with, including discussion with stake-holders how to optimize making a data-set that will be sufficient for development but minimize privacy concerns.
- (3 hours Mona) Review current project, get up to speed, obtain some understanding of domain and prior work.
- (10 hours Mona) Further exploration of vis capabilities, deploying working demo featuring possible idioms (bar chart, line graphs, pie chart) with some interactivity to change data being viewed.

November 19: Create first version of vis software using finalized idioms

- (2 hours Both) Meet to go over prototype so far and finalize idioms/design
- (6 hours JJ) Further development of data processing pipeline to convert ontology codes to readable data, deal with missing and erroneous data.
- (4 hours JJ) Begin development of capacity to show the hierarchical and ordinal nature of our data. Will include vis/data processing interfacing, and vis demoing of capability.
- (10 hours Mona) Begin developing other features of the vis, such as deployment of multiple views together with chosen idioms, and interactivity to support selection/filtering.
- This week, priority will be to supporting vis capabilities of only a select set of the attributes.

November 26: Iterate and improve vis software

- (4 hour Both) Meet with potential users and review initial vis, ready demo for same
- (8 hours mostly JJN) Improve vis: JJ's focus will be on allowing vis of more of the attributes, and functions that will allow the same such as superimposion.
- (8 hours mostly Mona) Improve vis: Mona's focus will be on adding/improving interactivity and multiple view capabilities, such as different zooming and linked highlighting.
- By this week, vis should support all attributes that will be included in this project.

December 3: Finalize vis software and possible extensions

- (6 hours Both) Further refinements, bug fixes, etc.
- (8 hours Mona) Develop and implement different clustering capabilities
- (8 hours JJ) Extend vis with features desired by broader CAPE research group. This includes a vis to compare the data generated by the NLP pipeline vs our hand-generated dataset, which will need to vis only a select set of attributes. Another requested feature was that our vis/processing could vis a new set of data which is generally the same, but may have some missing features.
- Depending on how long things are taking, during this week or the prior we may instead spend time implementing the vis in D3.

December 10: Final presentation

- (8 hours Both) Slides, rehearsal, demo

December 13 Final paper

- (20 hours Both) Preparation and editing. Likely some division, such as writing the methods section for what each of us implemented.

6 PREVIOUS WORK

Information visualization has a long history of application to clinical data, which we distinguish from other data used in biology/medicine such as omics (RNA, DNA...) and imaging (MRI, CT, xrays). Early examples, dating from mid-19th century, include Charles Mindard's graph showcasing the losses of Napolean's army as they marched to and from Moscow [2], and Florence Nightingale's radial bar chart showing causes of death during the Crimean War [10]. The amount of clinical data generated and potentially available to visualize is steadily increasing, owning to widespread adoption of Electronic Medical Records (EMRs) and the digitization of insurance records [7].

Early examples of visualizing these new data sources focused on showing the time series data of a single patient, such as the adoption of *LifeLines* [9]. This application uses the horizontal axis to represent time, while the vertical axis to fit labelled bars representing distinct events which happen over periods of time. Similiarly, the *KNAVE* system facilitate selection of a patient's attributes to be viwed in time-series small-multiple views [11]. Refinement of such work, in the *LifeLines2* [12] and *VISITORS* [6] systems respectively, allowed data from multiple patients to be viewed. *LifeLines2* was primarily designed to visualize temporal categorical data, of single or double-digit number of both subjects and attributes. Vertical space is divided into spaces for each patients, whose data is then shown through horizontal distance which codes data, with the color channel used on labelled glpyhs to the time of a number of categorical events. The *VISITORS* system used different views to allow both more subjects and more attributes. Users make selections of subjects, attributes, and time periods, and different views are used accordingly. For examples, a quantitative time-series of only one attribute uses a scatter plot, while parralel coordinates are

used to show multiple attributes from multiple subjects at discrete time points.

Recent work has continued explore different visualization techniques, though most applications remain centered around visualizing data as a time-series. Instead of a time series, the *DICON* system [3] is centered around visualization clusters of similiar patients. Each cluster is shown by a tree map, with color coding of the regions inside corresponding different diseases, which are in turn divided into regions corresponding to an individual patient having a certain disease. Another application eschews time-series to instead display the current state of an ICU patient [5]. Unsupervised machine learning is used to produce a quantitative measure for a set of attributes which are then visualized using a radial line graph.Another visualization was prototype a web-based pictorial visualization system [4] which allows spatial interactivity through representations of human body images (front and back views) and temporal interactivity through interconnected time axes. The end user for this system are both patients and doctors.

Our work seeks to build on this prior work in both technique and domain. While prior work has visualized clinical data, the focus has almost always been on the quantitative data easily accessed from an EMR, such as lab results or the timing of specific events. Our visualization will instead seek to visualize data that is more categorical, and more focused on the characteristics of the patients rather than what happened to them in hospital - a broader view of who our patients are. This, coupled with effective, interactive vis, will broaden the use of clinical data vis both clinically and in research. It will allow a broader range of clinical questions to be answered and generated. It will also be the first vis project of this sort in psychiatry, with less use for the common data previously visualized such as lab results and vital signs. The nature of the data we will be visualizing will also present technical contribute to clinical vis. We will explore new uses of various channels, interactivity, and idioms to effectively vis our data that is categorical but hierarchical, and will do so by utilizing medical ontologies. We will also contribute by using modern tools, such as D3 [1] and plotly, and by interfacing our vis with an NLP pipeline, helping to show what may soon be possible as NLP pipelines using clinical texts improve.

Note: the above is only an initial discsussion of prior work. We will be adding additional citations and discussion, such as of some more recent papers and a relevant MSc thesis on EMR vis from 2017.

REFERENCES

[1] M. Bostock, V. Ogievetsky, and J. Heer. D³ Data-Driven Documents. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2301–2309, Dec. 2011. doi: 10.1109/TVCG.2011.185

[2] P. Costigan-Eaves. Edward R. <I>Tufte The visual display of quantitative information</I>. *Information Design Journal*, 4(3):235–236, Jan. 1984. doi: 10.1075/idj.4.3.12cos

[3] D. Gotz, J. Sun, N. Cao, and S. Ebadollahi. Visual Cluster Analysis in Support of Clinical Decision Intelligence. *AMIA Annual Symposium Proceedings*, 2011:481–490, 2011.

[4] S. Jiaren. Pictorial visualization system with patient portal for problem-based electronic medical record. *A thesis submitted at university of Ottawa*, 2017.

[5] R. Joshi and P. Szolovits. Prognostic physiology: modeling patient severity in Intensive Care Units using radial domain folding. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, 2012:1276–1283, 2012.

[6] D. Klimov and Y. Shahar. A Framework for Intelligent Visualization of Multiple Time-Oriented Medical Records. *AMIA Annual Symposium Proceedings*, 2005:405–409, 2005.

[7] I. Ko and H. Chang. Interactive Visualization of Healthcare Data Using Tableau. *Healthcare Informatics Research*, 23(4):349–354, Oct. 2017. doi: 10.4258/hir.2017.23.4.349

[8] W. McKinney. Pandas: A foundational python library for data analysis and statistics. *Python for High Performance and Scientific Computing*, pp. 1–9, 2011.

[9] B. Milash, C. Plaisant, and A. Rose. LifeLines: Visualizing Personal Histories. In *Conference Companion on Human Factors in Computing Systems*, CHI '96, pp. 392–393. ACM, New York, NY, USA, 1996. event-place: Vancouver, British Columbia, Canada. doi: 10.1145/257089.257391

[10] F. Nightingale. *Notes on matters affecting the health, efficiency, and hospital administration of the British Army [electronic resource] : founded chiefly on the experience of the late war.* London : printed by Harrison, 1858.

[11] Y. Shahar and C. Cheng. Intelligent visualization and exploration of time-oriented clinical data. In *Proceedings of the 32nd Annual Hawaii International Conference on Systems Sciences. 1999. HICSS-32. Abstracts and CD-ROM of Full Papers*, vol. Track4, pp. 12 pp.–, Jan. 1999. doi: 10.1109/HICSS.1999.773019

[12] T. D. Wang, K. Wongsuphasawat, C. Plaisant, and B. Shneiderman. Visual Information Seeking in Multiple Electronic Health Records: Design Recommendations and a Process Model. In *Proceedings of the 1st ACM International Health Informatics Symposium*, IHI '10, pp. 46–55. ACM, New York, NY, USA, 2010. event-place: Arlington, Virginia, USA. doi: 10.1145/1882992.1883001