Practical Challenges of Progressive Data Science in Healthcare

Faisal Zaki Roshan *
Carleton University

Abhishek Ahuja†
Carleton University

Fateme Rajabiyazdi‡
Bruyere Research Institute
Carleton University

ABSTRACT

The healthcare system collects extensive data, encompassing patient administrative information, clinical measurements, and home-monitored health metrics. To support informed decision-making in patient care and treatment management, it is essential to review and analyze these diverse data sources. Data visualization is a promising solution to navigate healthcare datasets, uncover hidden patterns, and derive actionable insights. However, the process of creating interactive data visualization can be rather challenging due to the size and complexity of these datasets. Progressive data science offers a potential solution, enabling interaction with intermediate results during data exploration. In this paper, we reflect on our experiences with three health data visualization projects employing a progressive data science approach.

We explore the practical implications and challenges faced at various stages, including data selection, pre-processing, data mining, transformation, and interpretation and evaluation. We highlighted unique challenges and opportunities for three projects, including visualizing surgical outcomes, tracking patient bed transfers, and integrating patient-generated data visualizations into the healthcare setting. We identified the following challenges: inconsistent data collection practices, the complexity of adapting to varying data completeness levels, and the need to modify designs for real-world deployment. Our findings underscore the need for careful consideration of using a progressive data science approach when designing visualizations for healthcare settings.

Index Terms: Progressive Data Science, Data visualization, Health Data, Healthcare.

1 INTRODUCTION

Every day, a large amount of health data is collected and stored as a part of the healthcare system. These data include patient administrative information (i.e., demographic, visit records, appointments), patient data measured in the clinic (i.e., medical imaging, laboratory results, vital signs), and patient health data collected at home (i.e., blood pressure, heart rate, blood glucose). To optimize the healthcare industry's operational efficiency and enhance the delivery of healthcare services, clinicians, policymakers and management teams need to actively analyze healthcare data. However, this data is often large, messy, and difficult to interpret.

Data visualization is a promising solution that can support analysts in uncovering unknown unknowns, i.e., questions that are only uncovered while visually browsing data, finding underlying patterns and relationships that otherwise would have remained hidden, and revealing patterns and discovering valuable insights that might have otherwise remained concealed [2].

Fortunately, the healthcare industry has had a growing interest in visually analyzing these datasets. However, despite the interest in visualizing these datasets, there are still challenges to visualizing

*e-mail: faisalzaki@cmail.carleton.ca

†e-mail: abhishekahuja@cmail.carleton.ca

‡e-mail: fateme.rajabiyazdi@carleton.ca

and analyzing these datasets as the data is often large, incomplete, and collected in real-time. Progressive data science (PDS) could be a solution for data exploration in such large health data visualization systems to enable interaction with intermediate results [11, 7].

In this paper, we reflect on the opportunities and challenges we faced with our three visualization development projects in healthcare using the progressive data science approach. We explore the practical implications and challenges faced at five stages of PDS, data selection, pre-processing, transformation, data mining, and interpretation and evaluation, as defined by Turkay et al. [11]. Each project, from visualizing surgical outcomes to tracking patient bed transfers and integrating patient-generated data, highlights unique challenges such as inconsistent data collection practices, adapting to varying data completeness levels, and modifying designs for real-world deployment. Future work should focus on developing standardized methods and tools to streamline data collection, pre-processing, and interpretation and evaluation when using a progressive data science approach.

2 PROJECT#1: SURGICAL OUTCOMES

Many hospitals track and monitor various operational and patient health outcomes and clinical indicators to increase operational efficiency, aiming to enhance patient care, optimize resource utilization, minimize costs, and improve overall performance. These indicators are essential for monitoring and evaluating hospital processes, identifying bottlenecks, and guiding evidence-based decision-making about enhancing the quality of care delivery.

2.1 Project Description and PDS

In this project, we were invited to join forces with a local hospital to enhance to quality of surgical care. One pillar of this effort was to actively monitor and visually analyze the data. Thus, our goal was to design and develop an interactive data visualization system displaying surgical outcome measures enabling healthcare providers and administrators to make informed decisions.

We were asked to design an interactive visualization displaying predefined Key Performance Indicators (KPIs) collected from pre-, per-, and post-surgical periods. We started designing low-fidelity and medium-fidelity data visualization prototypes using a sample dataset collected from two hospitals. Once we developed our designed interactive data visualization system, we deployed it into the healthcare setting.

We used PDS to complete this project using PDS. We followed the five stages of data selection, pre-processing, transformation, data mining, and interpretation and evaluation.

2.2 Challenges

Below, we discuss the challenges we faced in this project using a progressive data science approach.

C1: Data Selection: We were invited to this project after parts of the data selection were done. This is a typical case in working with the healthcare industry, where the industry hires data scientists and data visualization experts to develop an interactive data visualization system for a dataset. These datasets are often collected in-house, in the hospital or by healthcare organizations.

However, when we started collaborating with hospitals for this project, only 2 hospitals out of over 50 hospitals started collecting
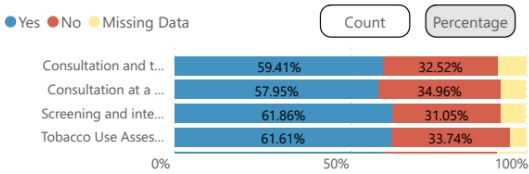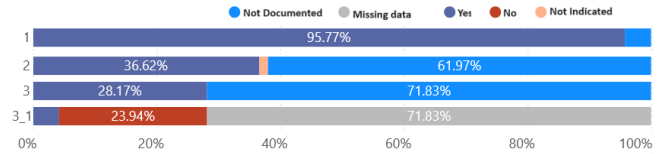
(a) Bar chart shows the distributions of 3 response options for several surgery quality check indicators in the selected hospitals.

(b) Bar chart shows the distributions of 5 response options for several surgery quality check indicators in the selected hospitals.

Figure 1: Changes to the response options for Project#1.

data. This introduced challenges in our process. Although we could work with parts of the data to start the design phase, it caused challenges in design later. Different centers collected data differently, which imposed design changes later in the process, consequently delaying the project.

C2: Pre-processing: As mentioned earlier, when started working on this project, we only had access to a small dataset collected from two hospitals. As these centers were the ones who initiated the project, they were extra careful in collecting data, leaving a very limited number of missing data in the dataset. This was great at first to work with a wholesome dataset; however, we faced challenges when other centers started collecting data. These centers had many missing data items that needed to be dealt with in the design.

At first, the originating hospitals collected three response options (yes, no, missing data) to a given question (See Figure 1(a)). When we gathered data from a few additional hospitals, we noticed other hospitals had two additional response options "not documented" and "not applicable", different from missing data.

This caused a challenge in both the design and development phases. We had a set colour scheme that was selected to be colour-blind safe and follow the institution's colour theme. However, to accommodate the two extra response options, we had to allocate two additional colours to represent all response options, which was challenging. The two additional colours needed to be easily distinguishable, colour-blind safe, and match the colour themes. This enforced some design changes in our visualization (See Figure 1(b)).

C3: Transformation: Consequently, we had to modify the database to transform the newly formatted data to allow for the collection of two extra responses. This necessitated our team to re-run the previously tested components to ensure they work as intended.

C4: Data Mining:

In this project, we employed a progressive data science approach and displayed the results of the calculations and visualizations using parts of the data. These included examples such as calculating the rate of patients who experienced complications after surgery.

The early partial computations enabled conversations with our medical collaborators on how to best display the calculation results. Our team involved administrative staff, medical professionals, programmers, and the information technology team. A challenge we faced was communicating a clear message that these results were not final. Expectations were set on the visualizations to show the final results. The team members mentioned that the numbers displayed would not match the numbers they ran when calculating results. Thus, we had to spend a great deal of time and resources explaining the progressive process to those who were not familiar with this process.

C5: Interpretation/Evaluation: Additionally, nearing the end of the project, since we used partial data and presented partial calculations, we realized some of the calculation formulas were not correctly used. For example, to calculate the days spent by patients in the hospital, one can count day surgeries as 0 or not count them in the overall calculation (sum or average). This considera-

tion changes the final results and the interpretation. Thus, we spent weeks revising such calculation formulas.

3 PROJECT#2: PATIENT BED TRANSFER STATES

A local hospital contacted us to find a solution for tracking and analyzing their patient smart bed transfers in the hospital to monitor and enhance the transfers.

3.1 Project Description and PDS

Our goal was to design and develop an interactive data visualization system of the beds' sensor log files generated by smart bed transfers to enhance delivery and operational efficiency.

3.2 Challenges

Below, we discuss the challenges we faced in this project using a progressive data science approach.

C1: Data Selection: A small sample retrospective dataset was shared with our team to start our design steps. The dataset was a collection of bed sensor logs when transferring a patient between units in a hospital. Using this small dataset, we sketched and designed several visualization representations. Upon sharing the designs with the healthcare provider team, we received feedback on the design and modified the design in an iterative process. Upon extensive discussions with healthcare providers and stakeholders, we selected key visualizations that best represent smart bed usage data for integration into the healthcare management system.

C2: Pre-processing: To transform our sample raw data into a structured format, we used a Python script and parsed the log files, extracting pivotal information while discarding irrelevant details. We then used the cleaned data as input to create a Microsoft Power BI data model.

C3: Transformation: In Power BI, we further transformed the data to enrich the data model and facilitate the calculation of KPIs critical to the project's objectives. When we shared these KPIs with our collaborators, we received feedback that the indicators did not display the information they anticipated. We then had to make changes to the calculation of KPIs.

C4: Data Mining: We started the project with a small dataset and designed the data visualizations accordingly. However, data at a large city hospital scale will exponentially grow, particularly in the case of bed transfer that happens multiple times a day for many patients.

Over the course of the project, more data was accumulated and displaying the final results was no longer possible. The hospital data recording systems are often old and slow. Thus, upon getting feedback from the hospital staff and administrators, we adopted our design to a progressive data representation. We will display a partial calculation of the KPIs (See Figure 2(a)) using a Gauge chart (i.e., a speedometer or dial chart) to show partial data and update the charts as data are collected. Once the numbers are finalized, we will display the final numbers (See Figure 2(b)). We are aware that

(a) Hospital bed transfer visualization dashboard while the calculations in progress.

(b) Hospital bed transfer visualization dashboard after the calculations are finalized.

Figure 2: Progressive demonstrations of the bed transfer reports for Project#2.

showing a partial number may be risky as it may be interpreted wrongfully. This needs to be clearly stated in the dashaboards.

C5: Interpretation/Evaluation:
This project started out with a loosely defined set of requirements and interpretation of data. The visualizations needed to be modified even at the final stages of the project, this caused extensive re-work on the calculations, which caused more changes in the backhand.

4 PROJECT#3: PATIENT-GENERATED DATA

Collecting patient-generated data is becoming increasingly common in chronic disease management [8]. Patients use various tracking tools to collect health and lifestyle data [1]. However, most of these tracking tools are not designed to fully meet patients' and healthcare providers' expectations [5] and do not support reviewing patient-generated data with healthcare providers during clinical visits. One way to support patients in discussing their data with healthcare providers is effective visualization of patient-generated data collections.

4.1 Project Description and PDS

We were approached by a group of healthcare providers from a local hospital who are involved in the care of chronic patients to explore if, and how, to design technology that can enhance the presenting and reviewing of patient-generated data during a clinical visit. To answer this question, we took an iterative approach with the involvement of patients and healthcare providers and designed various visualizations representing the patient-generated data collections [10, 9].

4.2 Challenges

Below, we discuss the challenges we faced in this project using a progressive data science approach.

C1: Data Selection: Both healthcare providers and patients agree that patient-generated data, could be used by patients to make informed decisions to improve their quality of life, and could aid providers in making decisions about patient ongoing care [12]. There are existing technologies for tracking and visualizing health data such as sleep (e.g., [4]), physical activity (e.g., [6]), and blood sugar level (e.g., [3]). We asked eight patients to bring a sample of their data and share it with us.

C2: Pre-processing: Some patients recorded and maintained their data on paper mixed with other personal information. Thus, as researchers, we needed to clean and transform hand-collected data into digital forms. Some of the patients who used apps or tools to collect data did not have easy access to their data as some of these tools did not provide an easy way to export or share their data. Thus, for some of the patients, we only had access to partial data.

C3: Transformation: Data collected by patients came in different formats, chosen personally by each patient. To include the data collected from patients in the electronic medical record, a certain format needs to be followed. We needed to reformat the data input to acceptable forms that are readable by medical record systems.

C4: Data Mining: We designed various individually tailored visualizations based on patient health data they collected at home. Upon extensive discussions with the healthcare providers, as a group we selected four visualizations that best represent patient data for incorporation into the provincial care plan platform.

One design was dedicated to showing blood pressure data, one design to representing stress level data, and two designs to displaying blood glucose data (one for patients who use an insulin pump and one targeting diabetes patients who use an insulin pump).

One design displays blood glucose from a Continuous Glucose Monitor for patients using an insulin pump machine. The continuous measures of glucose can be accumulated, making a large dataset when collected over months. Displaying this large dataset visually in a graph can take a large computational effort. Each data point needs to be placed on the graph in the correct placement and take an associated colour code demonstration (red: too high or too low glucose; yellow: borderline; and blue: normal glucose). From our discussions with healthcare providers, we gathered they sometimes do not need to review all this data in detail particularly during a short clinical visit. Thus, one proposed solution is to use PDS and demonstrate parts of the data (See Figure 3(a)) before finalizing all data placement on the chart (See Figure 3(b)). If healthcare providers have a concern and need to further examine the data, they may wait for the full data to be processed.

C5: Interpretation/Evaluation: Once we confirmed the design, the next step was to deploy them. However, we had to make changes to our original designs in preparation for developing and integrating the designs into the provincial care plan. Some of the libraries we used for rendering the visualizations were not in the accepted list of libraries for deploying into the healthcare system. This challenge was not raised when we showed the design and development in a progressive manner to the stakeholders. At this stage, we had to change some of the designs to accommodate the available visualization and development libraries, which caused further delays in the project.

5 CHALLENGES AND OPPORTUNITIES

Here we discuss the practical challenges and opportunities we identified when employing progressiveness in designing and developing data visualization dashboards in the healthcare setting. Although taking on a progressive approach was instrumental in the rapid progress of the visualization projects, it introduced some challenges. Particularly in the Data collecting and cleaning, Data Mining, and Interpretation/Evaluation phases. Often these challenges introduced issues in rendering and deploying the data visualizations. To solve these challenges, we needed to make changes to the design of the visualization, a step backward from development or deployment to the design process.
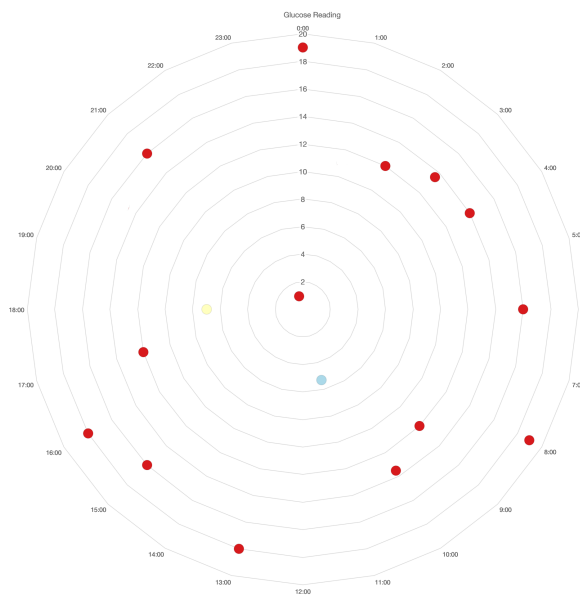
(a) Glucose data visualizations in progress.
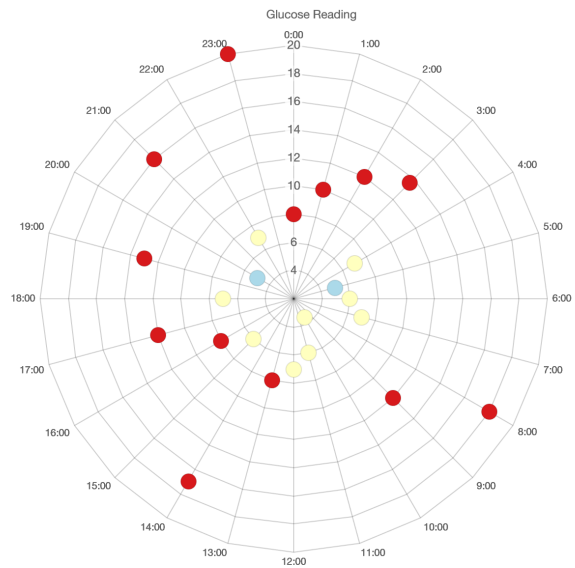
(b) Glucose data visualizations finalized.

Figure 3: Glucose data visualizations progression for Project#3.

6 CONCLUSION

Our exploration of progressive data science in healthcare reveals several challenges and potentials. By enabling real-time interaction with intermediate results, progressive data science can speed up the progression of such projects. However, this approach also introduces unique challenges, particularly in data selection, pre-processing, and consistent communication with stakeholders.

Each project we undertook—visualizing surgical outcomes, tracking patient bed transfers, and integrating patient-generated data—highlighted the need for flexibility and iterative design. Inconsistent data collection practices, varying completeness levels, and adapting designs for deployment in real-world settings necessitated continuous adjustments and stakeholder engagement.

Future work should focus on developing standardized methods and tools to streamline data collection, pre-processing, and visualization when using a progressive data science approach. Additionally, fostering strong collaboration and clear communication among all stakeholders is crucial for successful implementation. By addressing these challenges, we can harness the potential of progressive data science to enhance healthcare delivery and outcomes.

ACKNOWLEDGMENTS

REFERENCES

[1] J. S. Ancker, H. O. Witteman, B. Hafeez, T. Provencher, M. Van de Graaf, and E. Wei. The invisible work of personal health information management among people with multiple chronic conditions: Qualitative interview study among patients and providers. *J Med Internet Res*, 17(6):e137, Jun 2015. doi: 10.2196/jmir.4381 3

[2] E. Bertini and D. Lalanne. Investigating and reflecting on the integration of automatic data analysis and visualization in knowledge discovery. *SIGKDD Explor. Newsl.*, 11(2):9–18, may 2010. doi: 10.1145/1809400.1809404 1

[3] P. P. Brzan, E. Rotman, M. Pajnkihar, and P. Klanjsek. Mobile applications for control and self management of diabetes: A systematic review. *Journal of Medical Systems*, 40(9):210, Aug 2016. doi: 10.1007/s10916-016-0564-8 3

[4] E. K. Choe, B. Lee, M. Kay, W. Pratt, and J. A. Kientz. Sleeptight: Low-burden, self-monitoring technology for capturing and reflecting on sleep behaviors. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '15, pp. 121–132. ACM, 2015. doi: 10.1145/2750858.2804266 3

[5] E. K. Choe, N. B. Lee, B. Lee, W. Pratt, and J. A. Kientz. Understanding quantified-selfers' practices in collecting and exploring personal data. CHI '14, pp. 1143–1152. ACM, 2014. doi: 10.1145/2556288. 2557372 3

[6] S. Consolvo, D. W. McDonald, T. Toscos, M. Y. Chen, J. Froehlich, B. Harrison, P. Klasnja, A. LaMarca, L. LeGrand, R. Libby, I. Smith, and J. A. Landay. Activity sensing in the wild: A field trial of ubifit garden. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, pp. 1797–1806. ACM, New York, NY, USA, 2008. doi: 10.1145/1357054.1357335 3

[7] J.-D. Fekete, D. Fisher, A. Nandi, and M. Sedlmair. Progressive data analysis and visualization, 2019. 1

[8] S. Fox and M. Duggan. *Tracking for health*. Pew Research Center's Internet & American Life Project, 2013. 3

[9] F. Rajabiyazdi, C. Perin, L. Oehlberg, and S. Carpendale. Exploring the Design of Patient-Generated Data Visualizations. In *Graphics Interface Conference (GI)*. Canadian Information Processing Society, Toronto, Canada, 2020. 3

[10] F. Rajabiyazdi, C. Perin, L. Oehlberg, and S. Carpendale. Communicating patient health data: A wicked problem. *IEEE Computer Graphics and Applications*, 41(6):179–186, 2021. doi: 10.1109/MCG.2021 .3112845 3

[11] C. Turkay, N. Pezzotti, C. Binnig, H. Strobelt, B. Hammer, D. A. Keim, J.-D. Fekete, T. Palpanas, Y. Wang, and F. Rusu. Progressive data science: Potential and challenges, 2019. 1

[12] H. Zhu, J. Colgan, M. Reddy, and E. K. Choe. Sharing patient-generated data in clinical practices: An interview study. In *AMIA Annual Symposium Proceedings*, vol. 2016, pp. 1303–1312. American Medical Informatics Association, 2016. 3