Data Science in Practice: Querying and Visualization Final Report

Tanya Dimitrov
Leonardo Oliveri
Pushpa Yadav
Cheyenne Heath

Project report for the Data Science in Practice course 2021

LIACS
Leiden University
January 15th 2022

Contents

Data Science in Practice: Querying and Visualization Final Report

Pushpa Yadav [*1] **Cheyenne Heath** [*1] **Tanya Dimitrov** [*1] **Leonardo Oliveri** [*1]

1. Introduction

Nowadays, data is the most valuable and expensive resource in the world, data is the fuel of several businesses, and it is also responsible for several of the latest improvements in the average quality of life. Moreover, data collection, in the last decades, shifted to a digital-based process which consists of an automated and continuous flow of information. This process has slowly become the norm in our lives thanks to the devices with which we interface every day. However, there are some countries where digitization is still in an early phase and the lack of it is beginning to create issues.

This paper addresses one of these problems which has become even more important after the pandemics outbreak. Digitization and analysis of medical data in Africa is still at a starting point, due to the lack of resources and infrastructure. The absence of digital data makes it difficult to track useful information to assess relevant metrics, like the size of a disease outbreak or the level of infant mortality. For example, in the last few years digitization helped keeping track of COVID cases, deaths and vaccines. However, in Africa this tracking has been more challenging also due to the absence of a digital infrastructure.

Hospital records often collect patient information through several different formats, using a system unique to the hospital. This is convenient in the short-term as it allows personnel to rapidly fill in information in customized forms, but it is impossible to make comparisons with other facilities. In addition, paper forms often end up in filling entire closets and back rooms of hospital buildings, largely unused after collection. The use of paper records in hospitals causes problems in large scale analysis which cannot be performed without allocating a lot of time and personnel. However, if these records were to become available to be analysed through the process of digitization of the records, information from these records could be extracted and used in analysis.

The following work has been developed over the FieldLab proposal "Querying and Visualization" presented for the course Data Science in Practice as part of the VODAN project of the GO FAIR foundation. In the context of this paper, an approach that uses an interactive dashboard to visualize the relevant metrics will be investigated. A dashboard provides the opportunity to create meaningful analytics using aggregated data, which allows to perform analysis relying on real and updated data. Therefore, the team had to take into consideration different final users and different types of analytics to suit their needs.

The goal of this project is not about creating a ready-to-use product for the healthcare facilities in Africa, but to reflect on what could be a viable and useful solution. This is achieved by following the steps that will be described in the next chapters, where challenges will be analysed and setbacks will be discussed.

2. Team Expertise

2.1. *Tanya Dimitrov*

Tanya Dimitrov is a second year master's student at Leiden University studying Data Science at the Leiden Institute of Advanced Computing. Previously, Tanya received her Bachelor of Science Degree from the University of Connecticut, United States of America, in Biomedical Engineering with a minor in Computer Science. She has an interest in medicine, computing and the intersection of both topics. Tanya Dimitrov was previously employed by Brigham and Women's Hospital in Boston, United States of America as a Programmer/Researcher. At Brigham and Women's Hospital Tanya worked in the Sleep Medicine department investigating spindle brain activity during sleep.

Tanya's background makes her adept for research in the medical space. Her skills lie in computing and processing of medical data, however her background has also provided her with experience in patient privacy and data sharing that is essential in medical research and hospital settings. In addition, her computing background has provided her with python skills that will be used for data creation and analysis. Finally, she has experience in pipeline creation for large projects which will be important for a multi-scale project.

[*]Equal contribution [1]Leiden Institute of Advanced Computer Science, The Netherlands. Correspondence to: Pushpa Yadav <p.yadav@umail.leidenuniv.nl>.

2.2. *Cheyenne Heath*

Cheyenne Heath is in her second year of the master Data Science at the Leiden University. She has experience in programming with different languages as well as performing analysis on the gathered data. During this project she will assist with coding both the front and back end of the dashboard, as well as doing the internal and external communication when needed.

Cheyenne received her Bachelor of Science Degree from the University of Leiden as well as a Bachelor of Arts Degree in Film and Literature Studies. This combination of both natural science based courses and humanities courses makes for a varied point of view and can be a useful asset when finding different solutions for problems. During her Bachelor Computer Science she worked on creating a pipeline to predict possible depression signs of people using Twitter Data. This gained her some knowledge in working with a pipeline structure which will be useful for this project. Additionally working as a student ambassador for both her studies she has some experience with organize projects, which can be helpful in the planning stage of this project.

2.3. *Leonardo Oliveri*

Leonardo Oliveri is in his second year of the master program ICT in Business at Leiden University. During his bachelor in Information and Business Organisation Engineering at the University of Trento, he developed his business and information technology knowledge, which have been further strengthen during the master. For his bachelor thesis he developed a software to measure performance of the factory line operators using computer vision. The software provided a dashboard that uses the collected data to deliver relevant insight on how to improve processes.

Although Leonardo never went deep into programming, he knows the basics of coding, in addition he has a background in economics. His contribution to this project will be to provide support in the analysis of the data and translate needs and requirements of the stakeholders into a concrete solution. Thanks to his experience in the creation of a dashboard and in requirement analysis complement the range of skills available in the team.

2.4. *Pushpa Yadav*

Pushpa Yadav is a second year master's student at Leiden University studying Artificial Intelligence. She has few years of industry experience where she has used data science algorithms and analytic to gain insights and, analyse strategy and fund's performance. Her skills will be helpful in bridging the gap between front-end(user data visualization) and backend (querying the data).

3. Exploring the Problem

3.1. Related Work

From Report 3 Big data in health care is a field that has big potential (Raghupathi & Raghupathi, 2014). It can provide insights that would otherwise not have shown, while also reducing the cost of extracting this information. Health care facilities would greatly benefit from these techniques and the extracted insights. This data can be visualised and improve hospital scheduling and patient flow, research shows (Anneke Fitzgerald & Dadich, 2009). Using standardized electronic based Health Information Systems in healthcare can also improve the interoperability between health facilities (Olaronke et al., 2013).

While both paper and digital health record can be used to store medical data, research showed that a combination of both paper and electronic health record has advantaged over both individually (Stausberg et al., 2003). The advantage of paper health records could also be present in electronic health records if the advanced features of EHRs would be used (Price et al., 2013).

In order to provide the stakeholders with visualizations that give insight into the health facilities, the data needs to be queried. The data that is provided is either in JSON format or in RDF format. For JSON data, there has been research into structural ways to query this type of data (Petkovic, 2017). One structured way of querying JSON data is called J-Logic (Hidders et al., 2020). J-logic is an approach that focuses on paths. These paths consist of keys and can be used to access the tree structure that is typical for nested JSON objects.

To visualize the data, there are important steps to take to make sure that the visualizations is as transparent as can be. This is important as the stakeholders need to interpret the visualization correctly. An issue a lot of visualizations have is that they don't explain how to read them (Weissgerber et al., 2019). By keeping in mind colorblindness, adding a flowchart and using dot plots instead of bar graphs the visualization will be a lot more transparent.

3.2. Interviews preparations

In order to execute the interview in an efficient way, a few things need to be prepared. The first thing is to create a small introduction of our problem that we can read to the people being interviewed. In the previous assignment an introduction and problem statement were created. These paragraphs are the basis for the introduction we give to the interviewees. This introduction is put in a shared document, so we all have access to it, later we also created a presentation to better introduce the topic to the interviewees that were not familiar with it.

The next thing to prepare is the interview topic list, which is the list containing all the topics that we want to cover during the interview. These topics are organised to ensure the collection of insights from different perspectives, touching themes that have different relevance for different stakeholders. The mentioned topics are then expanded into questions, as can be seen more in detail in the annexes. The topic list is divided into five different topics, each with their own purpose:

- **Introduction** has the goal to get to know the person being interviewed and what their role is, consists of understanding what are the main challenges he/she faces in the daily work.
- **State of the art** is an analysis to understand what the main issues are and what we should expect while dealing with healthcare facilities in Africa. Additionally, we tried to investigate what is already existing in the current state of the art in terms of dashboards and implementations.
- **Needs and requirements** section consists of a set of questions that makes sure that the project team clearly knows what is being asked from the stakeholder and helps to give more clarity to the final goal. Therefore, these questions are used by the team to understand who the final user of the products is and what is valuable to him.
- **Implementation challenges** topic helps the project team to get insight in the pitfalls and risks that the interviewees know of or have experienced. For example, the treatment of personal data can raise privacy issues.
- **Data and technical tools** section helps the team to get to know more about the forms that they will use to collect data and the technical tools to query and to visualize the dashboard.

The team was helped by the project supervisor to get in contact and schedule the interviews with the stakeholders. Thanks to Mariam (the project supervisor), it was easily manageable to schedule the interview with our three main stakeholders: Mariam Basajja, Aliya Aleka Aktau and Nobert Sebbuggwawo. The interview was scheduled for Wednesday October 27th at 12 PM. However, due to technical problems due to an unstable internet connection Nobert was not able to join the call, so the team rescheduled an additional meeting on Monday November 1st.

3.3. Interviewees

The three people selected to be interviewed are the main stakeholders of this project:

- *Aliya Aleka Aktau* a PhD student working with Dr. Mirjam van Reisen on the VODAN Africa project as a Technical Support Group Leader. Aliya works on the architectural components of VODAN, specifically the software tools and the interaction they require between one another. Aliya is one of the interviewees because she works as a Techincal Support Group Leader. She can provide the project team with insight in the technical side of querying data, but also can share her experiences she gained while working in this role for VODAN.
- *Mariam Basajja* a PhD student working with Dr. Mirjam van Reisen on the VODAN Africa project as a Technical Support Group Leader. Mariam shares similar responsibilities as Aliya, working on the technical coordination of the individual components of the project. In addition, she is in close contact with many of the stakeholders found in Africa Mariam is one of the interviewees because as the supervisor of this project, she has valuable knowledge in both working with the hospitals in Africa and working with CEDAR to digitize the data.
- *Kampala International University Teaching Hospital* Our last stakeholder to interview will be Kampala International University Teaching Hospital Uganda. Due to technical difficulties, we were not yet able to meet with this stakeholders. This will be a vital stakeholder as it will help determine the final dashboard implementation for the project.
- *Case Clinic in Uganda* Finally, we will be in contact with Case Clinic in Uganda who will provide us on insight on the dashboard. We will be in contact with Nobert Sebbuggwawo, a Data Management Personnel at case clinic. Norbert is an interviewee because of mostly the same reasons as Raymond and Enjamine. However, because he works for another hospital his independent thoughts and requirements for the goal of this project are very valuable to the project. Norbert works in the IT department of the Case Clinic hospital in Uganda and will be able to assist us during the project with the technical challenges of creating the dashboard.

3.4. Findings

Interview 1: Group Interview with Mariam Bassajja and Aliya Aleka Aktau

In the first meeting the team interviewed Mariam Basajja and Aliya Aleka Aktau, PhD students with Dr Mirjam van Reisen and Technical Support Group members of the VODAN Africa Project, participated in a joint interview on October 27th, 2021.

The interviewees indicated that the main challenges facing healthcare facilities in Africa are data organization and data management, with an emphasis on paper-based organization. Paper data management accumulates large amounts of data that are not integrated. The interviewees mentioned that some hospitals with internet access have access to a DHS software, however this is not a widespread tool present across Africa. In addition, the facilities that do have DHS access do not often have access to large scale data in DHS to complete analysis. The DHS tool to better succeed in its goal must be integrated inside a strong digital infrastructure, comprehensive of several facilities. Only when this is achieved the derived analysis will be able to show relevant metrics to study trends in African healthcare.

In addition to paper-based records and problems with data management, many of the hospitals in Africa lack the technical infrastructure to support the foundations for this project. There can be internet connectivity issues due to a lack of coverage especially in the rural areas. Additionally, a lot of facilities are encountering a shortage of computers to store digital data, and no data stewards to help with data integration.

Although digitizing paper records will be a large hurdle to overcome, once completed, a dashboard can be beneficial to stakeholders. As the users are different in jobs and needs, the dashboard must be highly customizable. Therefore, it can be adjusted to an individual stakeholder need, for instance increasing and decreasing the timeslot of when the data are generated or including only the hospitals in a specific area. Although it is not clear yet what characteristics the mock dashboard for this project will have (this will be determined during future interviews with hospitals and when the mock data available has been defined), it is clear that there are many different avenues that can be reached with such a project.

Other important aspects that should be considered when dealing with data are privacy and the intended audience. It is likely that doctors and nurses may want to use the dashboard, but the access should be granted at different levels based on privacy concerns. Also, it is possible that the dashboard may be designed for business analysis and may only be able to display aggregate patient data due to privacy concerns.

The team will be provided with a form to collect the data to create a first dashboard with some visualizations. However, due to privacy concerns the team must create their own set of data to show in the dashboard. The populated forms on CEDAR will provide a set of json files with the information filled in the template, those are the data that the team will query. When asked if there are any existing dashboards to use as a model for our design both interviewees agreed that there is not an existing platform that they are aware of.

Interview 2: Interview with Norbert Sebbuggwawo

In the second meeting the team interviewed Norbert Sebbuggwawo, IT manager of the Case Clinic hospital in Uganda on November 1st, 2021.

The interview was very insightful. He informed us about the current process of data management, where analytics is done manually by exporting data from the database into excel sheets. Our project aims to query data and provide visualizations which will help them to draw different insights using the current data like predicting yearly trends for a specific disease. The existing DHS software is a government reporting system that provides aggregated data on metrics such as vaccination rates and cases of HIV in the facilities that are included in the program.

Regarding the current use of data in his clinic Norbert says: "We have a lot of data, but we are not using that data to make informed business decisions". He highlighted that our work could be used by management employees to track patient numbers, drugs and vaccines count, turnaround time and to attract more patients to the facility. In this case we can see that the information of the patients is hidden, and the metrics are crucial for resource management of the hospital.

We also discussed how FAIR policy could be adopted in our project. In order to maintain privacy patients' details (for e.g., name) will not be displayed. Only aggregated data (without confidential data) will be shown in the dashboard. In addition, regarding other challenges that may be encountered he was optimistic. He stated that there will not be any infrastructure challenge to overcome, because the facility is ready to welcome such a tool, and people are eager to learn this kind of tool.

Norbert has also agreed to join our biweekly meetings where he will track our progress and give us feedback.

3.5. Setbacks from the interviews

- *Data*:
  Before the interviews we were expecting real data from stakeholders in order to start this project. But now we have been instructed to work with mock data. We would like to know more about the method that we can use to emulate real world data and what is a fair number of mock data to be generated.

- *Fields to be displayed*:
  Hospital data contains confidential information. We need some more clarity related to fields to be displayed on the dashboard. Are we allowed to treat personal data? Nobert suggested to displaying general information where personal details will be hidden.

- *Future Interviews*:
  It was not possible to schedule all of our desired in-

terviews due to time restrictions, so there is still more information to be gained in future conversations. In the next interviews there should be the perspective of a hospital staff member with another role, like a doctor, and also some patients. In addition, other hospitals that are less furnished should be investigated to understand what are the challenges to be overcome there.

3.6. Problem Statement

Given the information collected during the interviews, the team was able to create a problem statement that takes into consideration the highlighted challenges and the value creation objectives. The overall objective of this project is to create a pipeline that can extract information from images of paper patient records and create live visualizations and analysis of the data on an interactive dashboard.

The interviews raised some issues for such an analysis and visualization tool applied in the context of African healthcare data. The main ones are paper hospital record format, the type of form that the hospital wants to analyse, and the method of analysis desired. This also creates a challenge in the querying of the data to perform necessary analysis. In addition, hospital records are typically confidential, creating an obstacle in the collection of data to be used for the testing of the prototype.

To overcome some of these initial hurdles, the pipeline prototype will be tailored to an example of a stakeholders needs. The stakeholder will provide example forms from the hospital, which will be used to simulate patient records to be used in the creation of the prototype. The stakeholder will also provide specific analysis needs that the team will leverage to create relevant metrics to include in the prototype. The prototype will be adjusted to be inter-operable amongst different hospitals to be able to query data with different hospital form formats.

CEDAR (Musen et al., 2015) is a tool that can be used to transform paper forms, to machine readable data. By using a structured way of creating fields in templates, CEDAR makes sure that the metadata of the data is on the same level everywhere. CEDAR can then produce JSON data in which all the information of the digitized forms is contained. This project starts there, using this JSON data (which will be synthetic data) the pipeline should send queries to extract the desired information. This information will then be displayed on an interactive dashboard for the researchers and other health care personnel to use for analysis.

4. Plan

This section describes the plan for this project. To make the plan visual we also included a project canvas and GANTT chart to make sure everyone (including the stakeholders) is on the same page. At the beginning of our orientating and planning phase we had roughly 8 weeks to allocate for this project. The planning was a high priority as every team member had a busy schedule and we wanted to make sure that there was enough time allocated for each tasks even with possible roadblocks we may encountered. Therefore we did spent a bit on time thinking out our plan, planning in interviews to orientate ourselves more on the problem and dividing the tasks among the group members.

Some considerations to take into account when planning out the project plan was the amount of time that is needed to create mock data. Because this mock data had to be made in order to fill out the dashboard using queries. A bit of time was therefore allocated to this specific task.

The priorities that are taken into account when planning the project were understanding the problem and determining what the desired output had to be. Most of these priorities would be taken care of during the interviews that we did with the stakeholders of our project. Another list of priorities that we made in advance to planning this project dealt with the demands for the final product. This included:

- Acquiring different types of forms
- Digitize these forms
- Create enough mock data to show some numbers on the dashboard
- Find a way to query the data from the generated json files
- Create a user friendly dashboard

For most of these points some research is needed before these elements could be implemented or executed. Therefore the time it takes to research should also be kept in mind when planning this project. These priorities were finalized after the interviews with the different stakeholders. The stakeholders from the hospitals had different ideas about the final product and its usage. However, after several meetings we concluded that the following elements should be included in the final Dashboard:

- Daily patient visits (real-time)
- Admissions (real-time)
- Discharge (real-time)
- Frequency of selected diseases (real-time)
- Frequency of morbidity (real-time)
- Frequency of mortality (real-time)

These elements will form the 6 KPIs(Key performance indicators) that have to be displayed on this project's dashboard. These visualizations will give the management of the hospitals more insight into the behaviour of their hospital. They can also see at what times there are more admissions or more discharges. All this information is a lot easier to view on a dashboard where you can interact with the graphs than static tables.

To make the project feasible, we are creating a prototype that works with local data. This makes it easier to get access to the data without having to write an API that links to CEDAR, the platform where the digitized forms are being stored. Another request from the stakeholders was that all data had to be accessed in real time. However working with local data makes this impossible as it is not linked to some kind of server. However, if files from the same form are being added or changed in the local folder, the result will be visible on the dashboard. Another consideration we had was to instead of manually populating the digitized forms, a program would be written that can populate hundreds of forms automatically. A downside of this is that the randomly generated results might clash or contradict each other and are not a good representation of real life data, but it did make it feasible for us to show more data and thus in turn demonstrate the capabilities of the dashboard. All these small adjustments made it much more likely that it would be feasible to deliver a good prototype in the given time.

4.1. Project Canvas

This section will explain the project canvas that was created before starting the project. The project canvas is divided into different boxes which each indicate a key point of every project. These boxes are visualized in Figure 1 and will be explained individually.

Data: Every dashboards needs data to visualize. Depending on the wishes of the stakeholder this data has to be selected. In the case for this projects mock data was used. This removed a lot of the privacy concerns and allowed us to easily generate more data without having to manually fill in each form. The data is based on different forms provided to us by Mariam Basajja. These forms included a lot of fields like: Health Facility Name, total number of cases of malaria, animal bites, cholera, measles etc. per week, total number of people who passed away from these diseases and causes, open fields about other diseases and causes not listed in the form, usage of different medicines and testing kits etc. Only the data needed to answer the 6 KPI's is selected during the querying.

Skills: To be able to complete this project a number of different skills is needed. We need skills in programming in python to build the dashboard. We rely heavily on python because it is easy to use and easy to understand. To be able to query the information we need from the different forms we need some knowledge on the process of querying. Even past knowledge of SQL could be useful as that provides some insight into the workings of querying. However for this project we will do the querying in Python as well. To work with the digitized forms. knowledge of CEDAR is needed. During the course we all got some hands on experience working with CEDAR as we digitized forms, filled in forms and created our own vocabularies. This experience is also needed to understand why CEDAR is used and how it is used in large institutions like hospitals. It explained the way the data is structured. In line with the reasons we need to understand CEADR, we also have to have some understanding on vocabulary building and BioPortal. For the actual dashboard creation part we need to be able to make nice visualizations and a nice user interface to make working with the dashboard as present of an experience as possible. Finally to be able to keep in touch with the different group members we all need some skills on communication, as communication is key with every project.

Output: The desired output of this project is a dashboard which can display the 6 KPI's as described above: Daily patient visits (real-time), Admissions (real-time), Discharge (real-time), Frequency of selected diseases (real-time), Frequency of morbidity (real-time) and Frequency of mortality (real-time). This dashboard needs to be nice to look at, easy to understand and easy to work with. It should also be able to work with different forms from different hospitals.

Value proposition: Dashboards can add a lot of value to institutions if done right. In the case of this project we want to show that there a several points in which the usage of a dashboard will add value. The most obvious value a dashboard might add is it will provide an easy way for both hospital staff and researchers to utilize the hospital analytics. This because the dashboard should provide some insight on the data which can then be used to perform analytics on. The dashboard can also provide statistics about the current and past hospital population, incidents of disease, usage of medicine and kits etc. This is valuable information that can be very hard to uncover if you don't have an easy and centralized visual way to view the data. In theory, a dashboard could be used to improve hospital efficiency and save money. This has several reasons: for one it is possible to see if there are any patterns in when people visit the hospitals the most, or when certain test kits and medicine are used most to avoid ordering too much or too little medicine and in turn waste less money. Finally the dashboard will bring value to both medical staff and business management staff alike in addition to benefiting to

patient care. When a hospital is better prepared for certain situations because they can use the dashboard to gain knowledge from the past, this will benefit all previously motioned groups by either having clarity what to do, have insight into the costs and have a better prepared hospital.

Integration: The project we are working on for this course will be a prototype and thus no actual integration into the stakeholder's data system will be provided. We will attempt to create a model hat is compatible with the stakeholders needs for possible future integration. However it was out of the scope for the amount of time and skills we have to be able to do this. The main purpose of this prototype is to show what value a dashboard like this can bring and why it would be word investing in such kind of tools.

Stakeholders: The different stakeholders and their roles are also described in Section 3. In total we have five distinct stakeholders: Case Clinic, Marian Basajja, Aliya Aktau, VODAN Africa and KUI. These stakeholders all have some interest in this project.

Customers: Possible customers for the result of this project are healthcare facilities in Africa that have the infrastructure to support a dashboard for data analytics and querying. This includes a lot of different things like a working stable internet connection, the storage space to store thousands of digitized forms of the hospital, a server to host a dashboard etc. In case a hospitals has the means for all these different needs, it is possible for them to implement such a dashboard structure. It would benefit all hospitals who don't have a similar systems because of the reasons described in earlier sections.

Costs: As a prototype, there is no cost associated with this projects other than time and labor. To implement the prototype to hospital environments may incur costs. The means that are necessary to implement such a product can be quite costly, especially in countries in Africa where it is not yet the norm. It also depends greatly on the country where the hospital is situated, therefore making it hard to define precise hypothetical costs.

Revenue: The dashboard does not directly generate revenue. However, if implemented by hospital business management to analyse different parts of the institution it could be used to improve the hospital management and spending. This because if the management has a better overview of when certain kits and drugs are used, or when the hospital is most populated during the year, they can anticipate these events and plan appropriately. This can be done for example by adjusting the amount of kits and drugs ordered or hiring more or less staff depending on the season.

FAIR maturity: The data that is used during this project is mock data. Therefore there are not direct concerns to the FAIR maturity. However, this project will be implemented using CEDAR and will comply to the FAIR data standards and the VODAN Africa projects standards. This was made sure by communicating well with Mariam Basajja and Alina Aktau as they have a lot of experience using data within the standard of FAIR and VODAN Africa.

4.2. GANTT Chart

In order to prepare our time well while planning for this project a GANTT chart is made. A GANTT chart is commonly used in project management as it is a very useful way of showing activities against time. This makes is so everyone on the team can easily view what needs to be done in which period of time. The GANTT chart for this project is visualized in Figure 2. There are 13 tasks displayed on the right and the time for this project spanned roughly 10 weeks, of which the first few weeks were allocated to creating a plan. The paper writing task spanned all weeks as this is an ongoing task which we added to each week by completing homework for the course and documenting our work. Conducting the interviews with our stakeholders as described in Section 3.3 was planned in for just a week, however one of the interviews was a bit later than planned because of scheduling issues.Next on the list was to create the project canvas as described in the previous section and define our goals which had to be described in the project canvas. For this actually quite a bit of time was planned in because this is one of the most crucial steps of the project. Defining the goals defined also how the rest of our planning, and GANTT chart, would look like.
The first task of the implementation was digitizing the CEDAR form. Both because this had not been done yet, but also to gain experience using the CEDAR platform. Next the mock data needed to be created using this form, the other forms were supplied to use by Mariam Basajja. The generation of the mock data was a bit of a lengthy process as first research needed to be done to determine what normal distributions were for field such as pregnancy and age. By using a legitimate distribution we hoped to get more realistic data. Whent his mock data was generated it was time to explore the data and select the needed data for the 6 KPI's we had to display. For this only a week was assigned for as this was not super difficult because we only had a limited number of forms to work with. If this was done on a bigger scale this would have taken much more time.
With all the data ready, it would be time for the actual implementation of the dashboard. Research into creating a

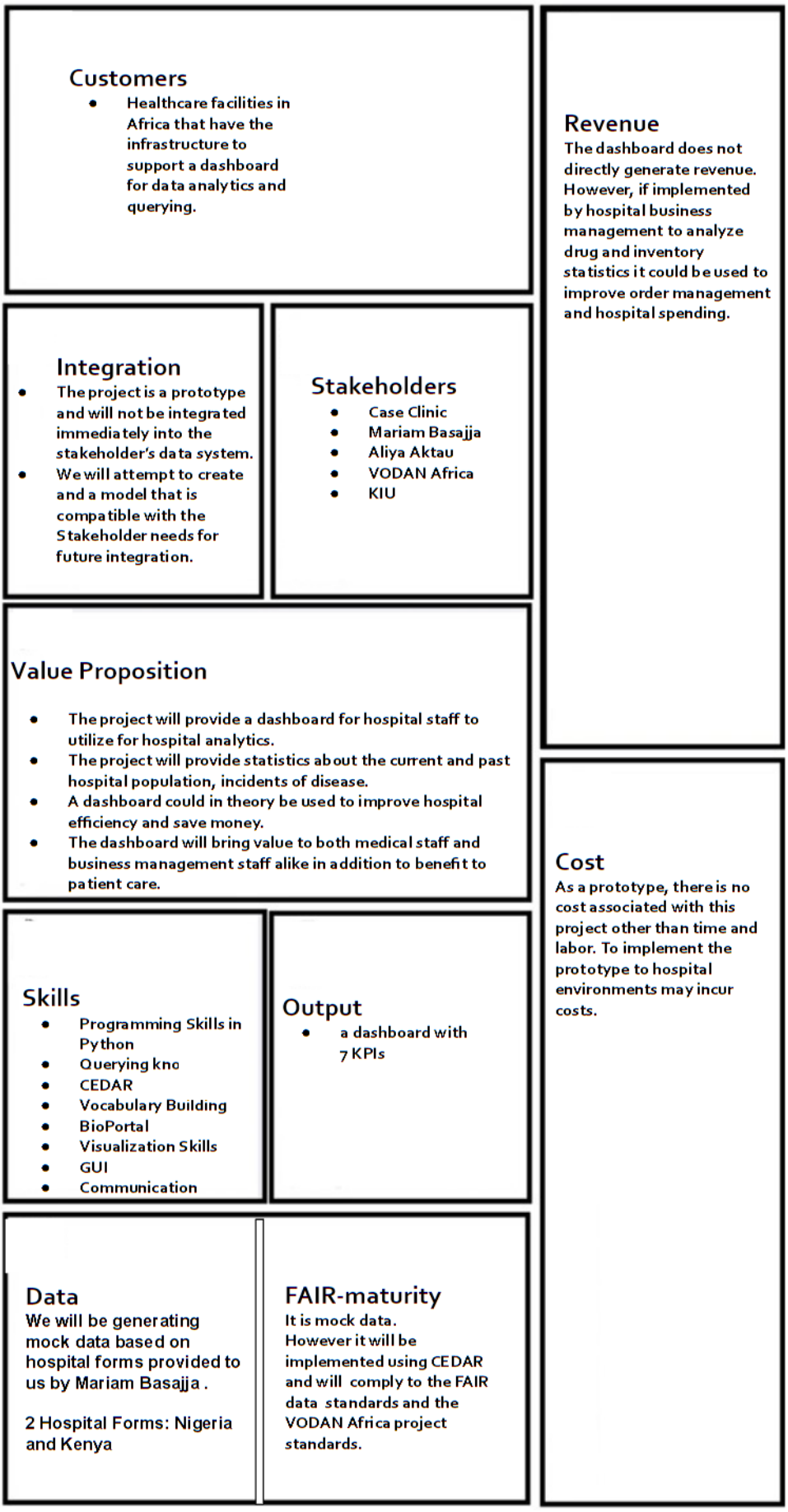

Figure 1. Canvas for this project

dashboard and querying was already planned in for during the creation of the mock data so that when that task was finished the querying could start immediately. But before the querying was done, a framework for the dashboard should already be in place to increase efficiency. The tasks were divided in such a way so that a number of people could work simultaneously without having to wait too much on other work to complete.
The final stint of the project would be dedicated to refining the dashboard and finishing the paper. In each of the tasks there was a little bit room for error, but especially with defining the goals, generating the mock data and creation of the dashboard, more time was allocated than for other tasks. Because these tasks all required coding or in case of the creating of the canvas, careful planning.

4.3. Division of labour

Looking at the GANTT chart described in the previous section, there is a blue and purple bar on the right. This indicated the two distinct categories of tasks we were able to define. the purple bar indicates the data and planning part, whereas the blue bar indicates the querying and programming of the dashboard part. Therefore it was very easy to split our team up in two groups: Leonardo and Tanya were in charge of the puprle bar and Puspha and Cheyenne were in charge of the blue bar. To make the tasks more specific the following distribution of tasks was made:

Leonardo: Paper writing, digitizing the forms, conduct interviews, communication, presenting.

Tanya: Paper writing, digitizing the forms, generating the mock data, communication, presenting.

Pushpa: Paper writing, research querying, querying the JSON data, creating the dashboard, communication, presenting.

Cheyenne: Paper writing, researching the dashboard creation, creating the dashboard, making the dashboard visually appealing, communication, presenting.

4.4. Methodology

The approach for this project is well visualized in the GANTT chart 2. After creating a plan the work can be divided and everyone has something to do. Weekly meetings and communication via a group chat makes sure that everyone is up to date with the work so far. Weekly meetings were held on Tuesday mornings unless we moved it specifically. The meeting always consisted of the four group members and were sometime accompanied by Mariam Basajja.

To keep the data and code organized we used a shared GitHub repository which we all have access to and can update individually. The advantage of using GitHub is that it also manages previous versions of the code. So for example, in the worst case if a mistake is made in the code and therefore makes a file unusable you can use GitHub's version control to get a previous working version back.

5. Implementation

5.1. CEDAR Form Generation

Although the templates of the analysed forms have been provided by the supervisor, we created the template of the previous form using CEDAR. We thought that this procedure was worth mentioning in this report. The "HMIS FORM 033b" is a weekly epidemiological surveillance report for a specific health unit. It focuses on the main diseases, like Malaria, dysentery, or SARI, identifying weekly cases and deaths. The form comprehends ten sections, we can see an overview in Figure 3.

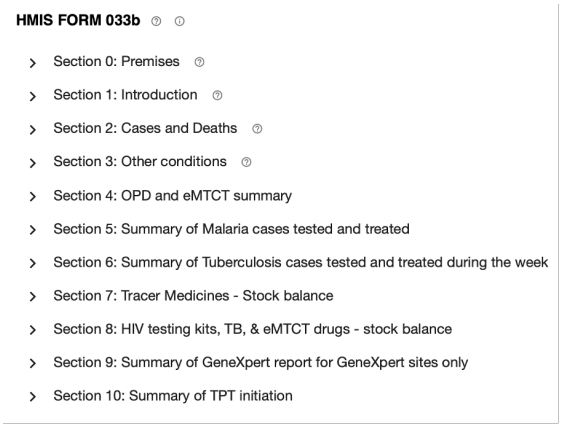

Figure 3. Overview of the template sections of the digitized form.

The creation of the form is a long but straightforward process, for each entry in the paper form you have to create one in CEDAR. Then every question must be defined to identify its format, which could be a multiple-choice question, or it may be an open question that requires a date, a number, or a string. For example, in section 2 regarding cases and deaths the entries are all numbers as it is shown in Figure 4.

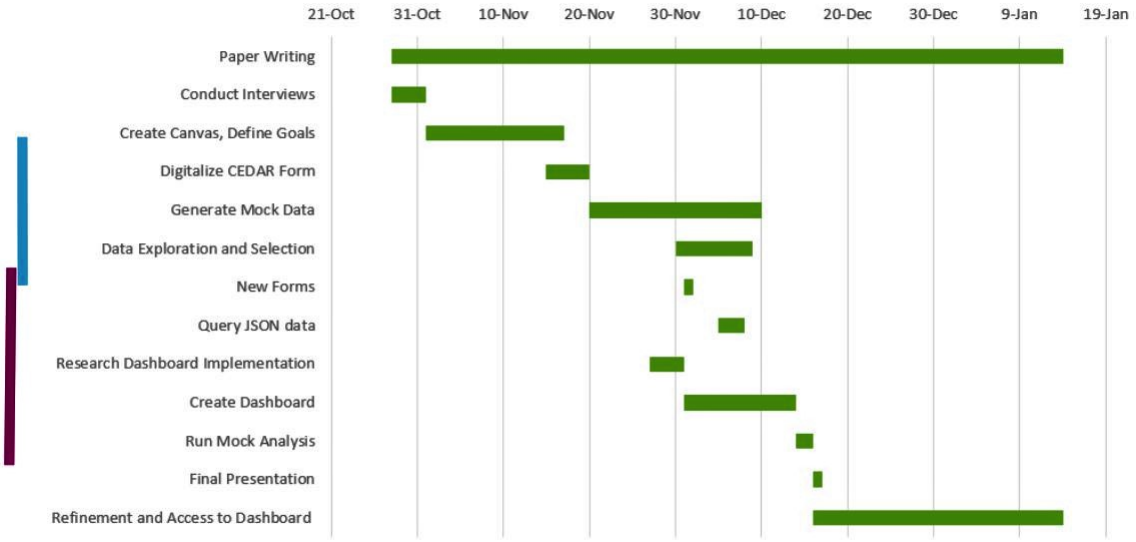

Figure 2. GANTT chart for this project, indicating the time span for different tasks.

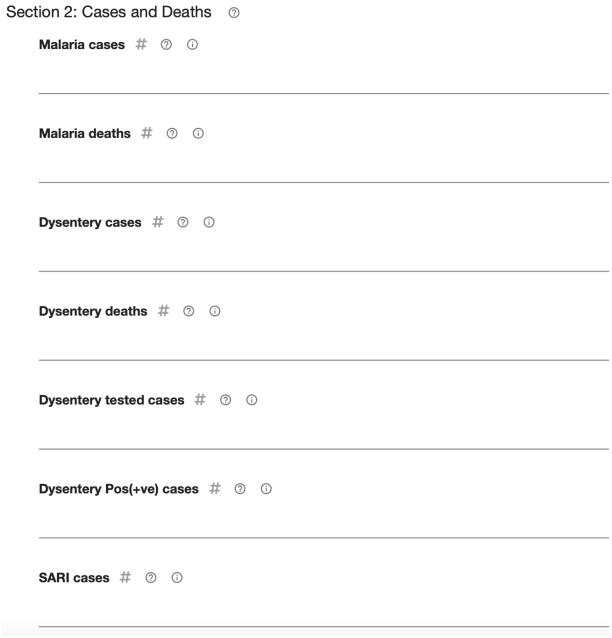

Figure 4. Example of form question formatting.

Finally, the template questions are connected to the BioPortal, which is a tool that provides anthologies and allows an unambiguous identification of the field.

5.2. Mock Data Generation

Two existing forms in CEDAR were provided for the project. One form is labeled the "Kenya Outpatient Register", and the second form is the "Nigeria HMIS Outpatient Register". Both forms were connected to the bioportal and use a controlled vocabulary. An example of the "Nigeria HMIS Outpatient Register" form can be seen in Figure 5. The drop down menu for the field "Type of Attendance" shows an example of the controlled vocabulary.

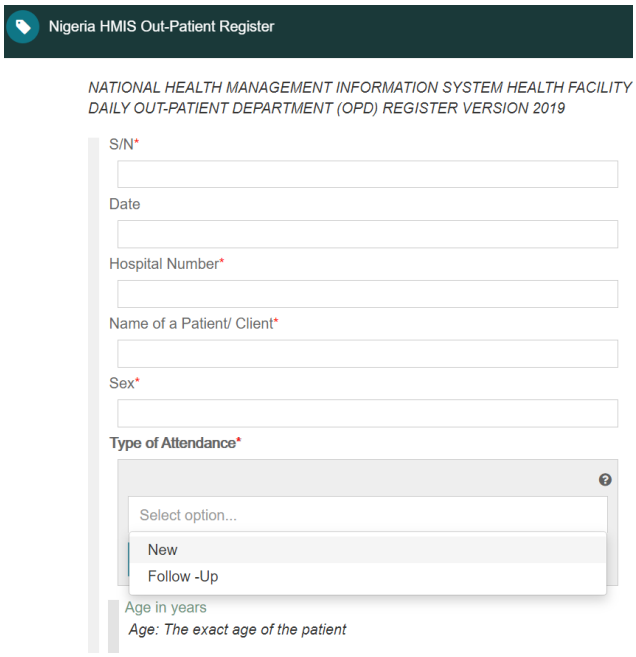

Figure 5. A portion of the "Nigeria HMIS Out Patient Register Form" that contains the field names. Select field names are linked to a bioportal controlled vocabulary as can be seen in the "Type of Attendance" field. For these fields, you select one of the provided options.

Due to privacy reasons it was not possible to use populated hospital forms. As a result, we needed to generate mock data. Mock data generation was expected to be completed by hand. This task would require using the CEDAR form to manually enter each value for a mock patient and download the individual .json file. Although this is a valid way to complete mock data generation, it would not have been a sustainable method in the generation of large data, as is needed to show a data continuation on our dashboard. As a result data generation needed to be an automated task.

To automate the data generation, we first we populated one form with mock data and downloaded the .json file. A .json file organizes data to have the data field name in quotations **" "** and the entry prefaced by an **@value** keyword. Using the bioportal controlled vocabulary for each data field, a separate .txt file was generated that had all the possible values for each data field. Using that document it was possible to edit the .json file looking for the proper field names organized with quotations **" "**, and then populating the **@value** field with a random choice of the suggested entries for that field name per the bioportal. The .json file was edited as one would edit a .txt file in python, but saved as a .json file at the end. For fields that are not linked to the bioportal (i.e. age) we generated appropriate values and randomly selected from that pool (for age this was 0 -100). Using these tools mock data was automatically created and gave access to hundreds of .json files without having to populate each form manually. An example of the json file can be seen in Figure 6.


```
  "Hepatitis": {
    "@context": {
      "Hepatitis B Screening": "https://schema.metadatacenter.org/properties/e4ebdd83-
d18a-4cdb-838b-8f197cd32fde",
      "Hepatitis C Screening": "https://schema.metadatacenter.org/properties/26dc8296-
37f9-4c93-a82c-31a08647aa90"
    },
    "@type": "http://data.bioontology.org/ontologies/NCIT/classes/http%3A%2F
%2Fncicb.nci.nih.gov%2Fxml%2Fowl%2FEVS%2FThesaurus.owl%23C3095",
    "Hepatitis B Screening": {
      "@context": {
        "Treated or Referred": "https://schema.metadatacenter.org/properties/2334177d-
bdd6-4c36-9caa-7b909acd511e",
        "Hepatitis B Screening":
"http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#C92806",
        "Hepatitis B Screening Result":
"https://schema.metadatacenter.org/properties/b5dafc62-7bfe-453c-99a9-7ce3ce2ab901"
      },
      "@type": "http://data.bioontology.org/ontologies/NCIT/classes/http%3A%2F
%2Fncicb.nci.nih.gov%2Fxml%2Fowl%2FEVS%2FThesaurus.owl%23C92806",
      "Treated or Referred": [
        {
          "@value": "T"
        }
      ],
      "Hepatitis B Screening": [
        {
          "@value": "No"
        }
      ],
      "Hepatitis B Screening Result": [
        {
          "@value": "Pos"
        }
      ]
    },
```


Figure 6. A mock generated .json file for the "Nigeria HMIS Outpatient Register" form. The field names are in quotations while the field values are prefaced with the **@value** keyword.

For the project demo, 100 forms of the "Nigeria HMIS Outpatient Register" were generated and 100 forms from the "Kenya Outpatient Register" were generated.

5.3. Dashboard Integration and Querying

A healthcare dashboard is a modern analytics tool to monitor healthcare KPIs in a dynamic and interactive way. A common example is a hospital KPI dashboard, that enables healthcare professionals to access important patient statistics in real-time.

The Dashboard for our use case was created using python DASH. Dash is a python framework created by Plotly for creating interactive web applications using python. Dash is open-source and the applications built using this framework are viewed on the web browser. Dashboard will display important information in the form of charts and graphs. This will provide quick information to the user.

After the mock data generation, the next step was to use the data to create the dashboard for querying and visualization. The data was generated in a JSON format. The data from mock forms, "Nigeria HMIS Out Patient Register Form" and "Kenya Outpatient Register" were loaded in a table using a pandas data frame. Two data frames were created, one for "Nigeria Form" and one for "Kenya Form". Each row in the

dataframe represents data related to one form. So, if there are 100 forms then there will 100 rows in the table. Each form consisted of many fields but not all were useful for the dashboard. Fields were converted into columns. Many fields were not in the right format to be used for querying. As part of pre-processing step, we converted such fields in the right format. One such field was ”Patient Service Date” in the Nigeria form and “Date of Service” in the Kenya form. Extra space and date format was modified for comparison purpose.

We are reporting certain KPIs in the dashboard. The KPIs can be filtered based on dates. Users can select one of the dates from the dropdown and data from those dates will be populated in the dashboard 7 . By default data for the last 25 days will be displayed on the dashboard. The functionality in the DASH was implemented using “DatePicker”.

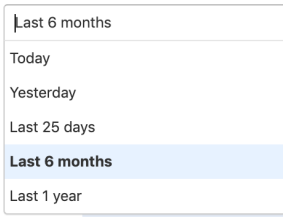

Figure 7. Dropdown List in the dashboard.For these fields, you select one of the provided options.

KPI’s are displayed on the dashboard. Query used for generating data pertaining to those KPI are discussed in following sections

- PATIENT VISITS:

  This section displays the number of patients who have visited the hospital. Users can select dates from the dropdown and the number of patients pertaining to those dates will be populated. The number of patients from both forms is aggregated. As described earlier, all forms are loaded in a data frame and each row represents data related to one form. So, 100 forms contain data of 100 patients and each patient record is a row in the table. Counting number for rows based on the date criteria gives us the number of patients who have visited the hospital on that particular day.

  Refer Figure 8 to view a sample figure to check number of patients who visited hospital on a particular day.

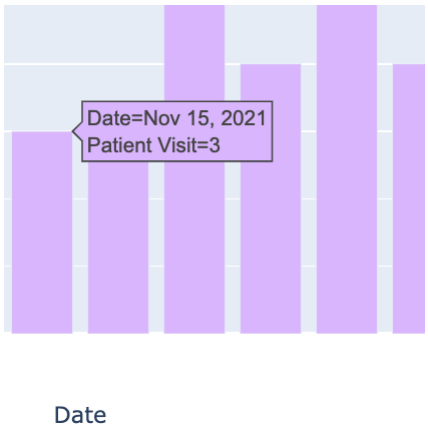

Figure 8. Patients Visit.

- DIAGNOSIS:

  This section displays the nature of an illness or other problem by examination of the symptoms. The form contains a field called “Diagnosis” and the data from that field is displayed in the dashboard. Users can select dates from the dropdown and diseases diagnosed during that period will be displayed in the dashboard. Users can hover over each bar graph and check how many patients were diagnosed with a particular kind of illness.

  Refer Figure 9 to view a sample figure to check number of patients suffering from malaria.

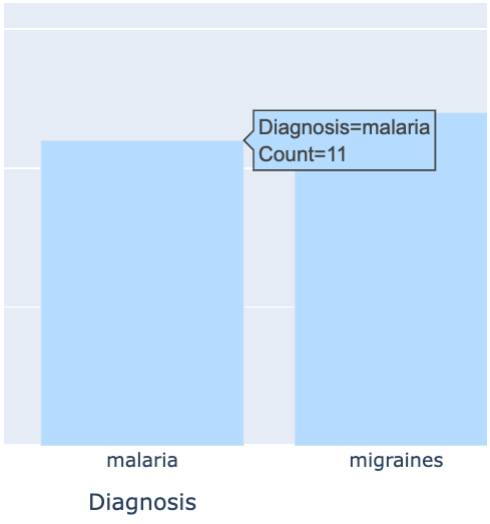

Figure 9. Diagnosis.

- OUTCOME OF HOSPITAL VISIT:

This section provides information related to the outcome of the hospital visits . Users can check the number of patients who were admitted to the hospital, referred to other hospitals, were treated and not treated by the hospital. All this information can be filtered based on the date. A python dictionary is created where the key is the outcome of the hospital visit and the value tracks the count.

Refer Figure 10 to view a sample figure to check outcome of a hospital visit. Figure shows that 22 patients were not treated.

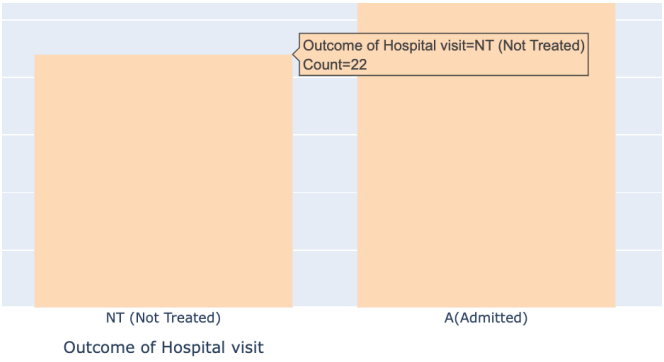

Figure 10. Outcome of hospital Visit.

- MORTALITY and MORBIDITY:

  A pie chart is created to display mortality and morbidity rate . The "Outcome of the hospital visit" field is used to find the frequency of mortality and morbidity.

  Refer Figure 10 to view a sample figure to know more details about the mortality and morbidity rate. Figure shows that 87 percentage of patients were suffering from some illness. There are 100 patients out of which 87 people are suffering from some illness.

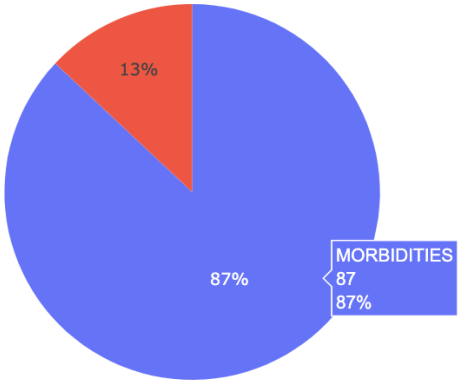

Figure 11. Mortality and Morbidity.

5.4. Dashboard

Like stated in the previous section, the dashboard created for this project is made using Python and specifically the package Dash (das, 2017). Dash provides a structure that is easy to use and expand on using simple lines of code. This enables the user to create attractive looking dashboards that are easy to understand without having to dive into HTML and JavaScript too deep.

With the data queried and the figures created it was very easy to implement these in the dashboard. Using dash a canvas is created that spanned the whole web page. On top of that the different images and accompanying text were posted so that each figure has some explanation.

The final result is a dashboard with a few figures to demonstrate the possibilities that having such a dashboard can offer. Figures 12, 13 and 14 show the final images on the Dashboard. The current final dashboard can be viewed in the users own browser by going to their local host webpage.

The repository has been made public and can be viewed on: `https://github.com/CheyenneH/VodanDashboard`.

6. Discussion

Within the scope of one semester the following tasks were completed: the team generated forms in CEDAR, created mock data for 200 patients, created a dashboard and integrated it with the mock data to perform data visualization. The end product was a prototype interactive dashboard. The prototype dashboard can be used to visualize 6 KPI's across time. Although the team was able to succeed in their goals there were many obstacles that were overcome. In addition, there are many opportunities for product improvement in the

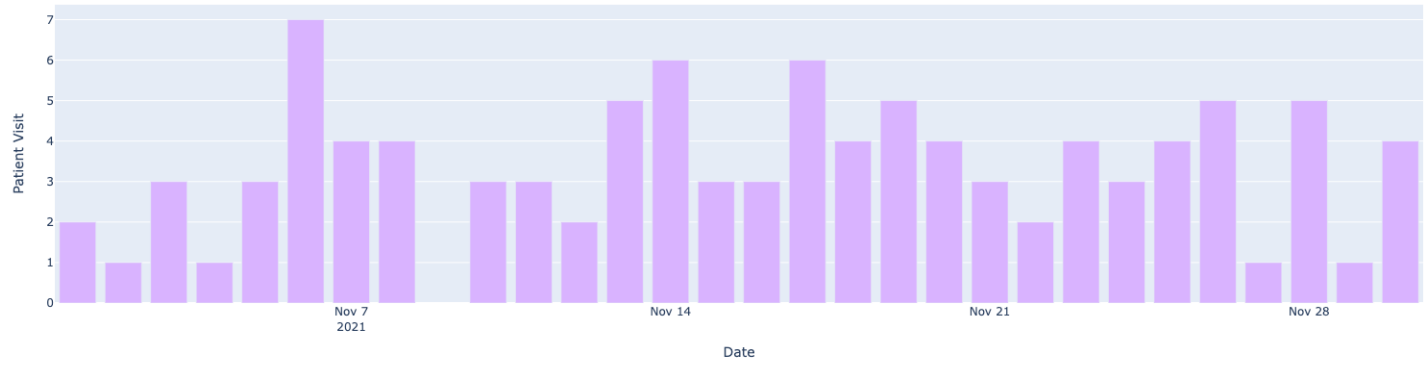

Figure 12. PATIENTS VISIT

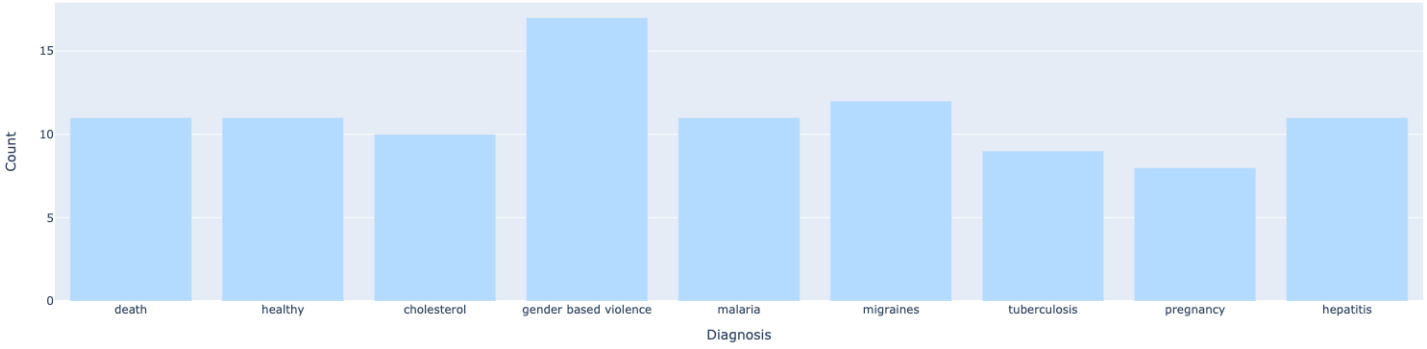

Figure 13. DIAGNOSIS

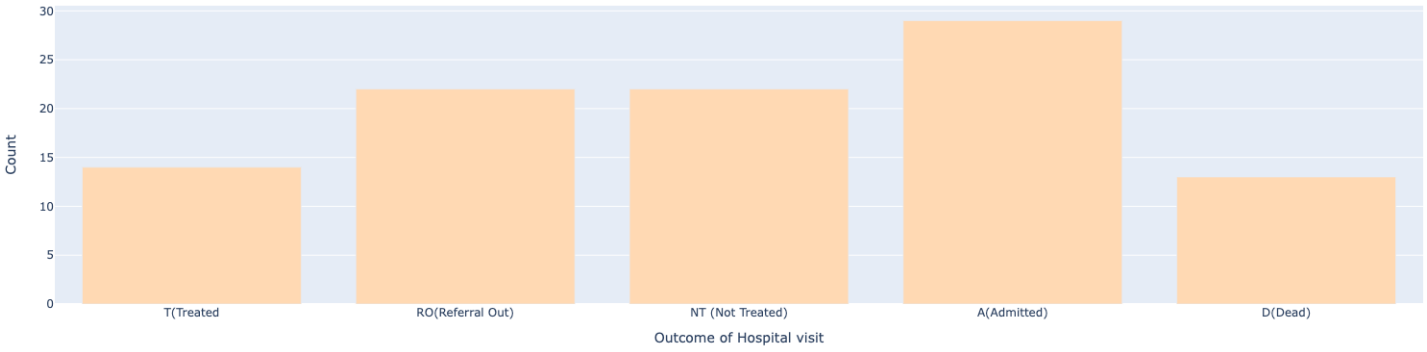

Figure 14. OUTCOME OF HOSPITAL VISIT

future. Finally, the project has great potential in industry.

As with any project, there were many hurdles to overcome. There were limitations in the data generation portion of the project. One limitation was a technical difficulty in the connection between the bioportal and CEDAR for the "Kenya Outpatient Register" form. The connection linking the two websites is not working and cannot provide the controlled vocabulary for each field. This meant that we could not generate mock data with the controlled bioportal fields. This did not affect the "Nigeria HMIS Outpatient Register" as it may have used different bioportal links, however the mock data for the"Kenya Outpatient Register" form was affected. This resulted in the team generating a controlled variable rather than using the specified values. In addition, the team wanted to generate mock data in a manner that was reflective of real world case numbers of disease. However, this proved difficult as there is not easily accessible information on disease incident averages with standard deviation across multiple different countries. As a result, the generated data was random and not representative of the real world case incidents. However, this did not affect the dashboard creation and visualization.

Due to the fact that data generation was random, this also meant that there was little control over data incongruity in mock forms. For example, the "Kenya Outpatient Register" has two fields: "Primary Diagnosis" and "Secondary Diagnosis". Using random data generation could result in the following situation: The 'Primary Diagnosis" of a patient could be "tuberculosis" and the "Secondary Diagnosis" could be "healthy". This is an obvious contradiction, as a patient cannot be healthy if they have tuberculosis. This type of data incongruity was present in the mock data generation. In the scope of the field lab this did not present a major difficulty. Both of the tasks (querying and visualization) were possible regardless of data incongruity and since it is mock data it is not meant to be interpreted as a real result. Data incongruity is likely to happen at lesser rates in the real world where forms will be filled in manually rather than generated randomly so it is not a concern that there are no such checks in place. Although there are many possibilities for error in manual data entry, it is less likely to have data incongruity at the same scale as with random data generation.

Once data generation was accomplished, there was another obstacle in using the data. Although both of the forms implemented bioportal controlled vocabularies, meant to improve data integration between platforms, the vocabulary is not identical between forms. For example, a test result in the Kenya form could be "Yes" or "No", while a test result in the Nigeria form could be "Y" or "N" although they represent the same responses. This applied to field names as well. This meant that querying from both forms required processing to integrate field names and field values. There were only two forms in this project, however the processing required would increase rapidly with the addition of more forms.

Our final limitation was in the dashboard. Although the dashboard is interactive and can query and visualize data it still has many improvements that need to be made before it can be implemented in an industry setting. The dashboard needs to be able to query multiple forms and be integrated with a hospital's data management system. For this project, since we had a small data set we assumed all the files were in one folder. It was possibel to query all the .json files with python from that folder without needing a larger data management system. However, this is not plausible for a hospital organization that has a large database of patient information. For an organization such as a hospital, there would need to be a more complex data querying system, which we did not implement in our querying due to the fact that we did not have access to such a system. In addition, the dashboard can be developed to be more advanced, encompassing more hospital statistics but further implementing statistical analysis and machine learning. This was not completed in our analysis due to time constraints.

Although we created a working prototype for the querying and visualization of 6 KPI's, it would be difficult task to extrapolate the prototype into a fully functional dashboard. There are two possibilities for a dashboard: A dashboard can be implemented for a a specific hospital establishment or a dashboard could be made to be interoperable between multiple hospital systems potentially across borders. The former is a more straightforward task. With a focus on a single hospital system it would be possible to design the hospital querying to be specialized for that hospital's forms and language. Regardless of the bioportal controlled vocabulary, which reduces variation in data considerably, integration between multiple hospital forms is still complex. With manual oversight, it would be possible to integrate hospital forms to query in a congruent manner. However, this task would become exponentially more difficult with the addition of more hospitals, or hospitals in different countries as was the latter option. It would become a necessity to have a strictly controlled vocabulary across borders. This may also pose problems as different languages are spoken in different countries which would add another layer of complexity to the task at hand.

The interoperable hospital dashboard system also adds a layer of complexity to the task at hand due to the need to respect FAIR data principles. If a dashboard was to query information about multiple hospitals, or multiple countries it would need to practice data visiting in order to comply to FAIR standards. This is essential to ensure that data stays in its location, however it would require time to create a

querying system to comply to the standards. One option is to create a centralized database where we can load data from multiple forms from different countries. We can create three tables. In the first table, we will assign a unique id to each form. So if we have 10 different forms, 10 unique IDs will be assigned to it. The second table will have information about the country and a unique id will be assigned to each country. The third table will load the data from all forms and will have correct mapping based on form and countries. This will be our central repository which will be used by UI. We can control access based on the countries and forms.

Finally, there would need to be security measures in order to bring the project into fruition. To create a working dashboard that could help doctors, administration staff and staff for hospital business staff there would need to be several levels of access. This would be a non intuitive task and would require user sign-ins, granting users access to different information. In a multi-hospital framework, this would require security checks between hospitals. In addition a multi-hospital system would need to have the same security measure compliance in order to integrate the systems. This was not implemented in our prototype as it was beyond the scope of this project.

In general, the creation of such a dashboard holds great promise. If it can be integrated into a hospital setting it would be an asset to the hospital. The administration staff would have access to a database of organized and digitalized hospital forms. Many hospitals in Africa currently rely on paper forms and closets dedicated to paper storage. This does not provide for an easy organization system and requires more time to locate specific hospital forms. A digitalized data management system and the ability to query forms, would provide the administration staff with a time advantage and convenience. In addition, administrative staff would have access to hospital metrics such as the number of available beds and the number of nurses/doctors. This would aid in hospital oversight.

Doctors and nurses would have several uses for the interactive dashboard. Doctors and nurses could use the interactive dashboard for information about individual patients. This would aid medical workers by offering convenience and ease of access to patient information. Doctors and nurses could also use the dashboard to interpret information about a population of people visiting the hospital. From such a dashboard, medical workers could do several of the following tasks: plot incidents of disease, visualize trends in data, understand community , reach conclusions about the general population's health. Finally, if the dashboard has more complex statistical analysis methods integrated and machine learning possibilities then it may be a huge asset in understanding underlying information about the population of people or in predicting future disease occurrences. With such a prediction tool the medical staff can better understand how to treat its patients.

Finally, this tool would be very valuable for the hospital business division. It is possible that hospital business staff could use the trend in patient data and disease to aid in the ordering and stocking of the hospital stock. A tool such as this could be used to understand what the expected amount of medicine would be necessary for a period of time. It could also let hospital business staff about when would be the necessary time to order supplies based on incidents of the population disease in order to be prepared. Finally hospital business staff could use the dashboard to understand whether there is need for hospital expansion or expansion of a branch of the hospital based on the rates of patients.

The dashboard has huge potential in statistics and machine learning. Access to large datasets of digitilized data provides with the opportunity to build models of the population. The models could be predictive of future incidents of disease magnitude. The models could also be used to unearth underlying factors in disease that may have been previously looked over by medical professionals. Patient disease histories could be summarized to help predict other patients progression in disease. Models could be run in the interactive dashboard and the results visualized in an interpretive manner. Although this would require considerable medical fluency and computer science knowledge it could be a major tool in hospital environment. We hope to see this in the future.

There is much value added for medical establishments in the creation of such a dashboard. With an investment of time and resources it is an achievable task. We hope to one day see it in use in hospitals and hear about the way it has changed hospital function.

7. Conclusions

As part of this field lab, we created a dashboard for querying and visualizing the hospital data. A dashboard is a visual representation of a hospital's key performance indicators (KPIs) in the form of charts and graphs. They are faster to read and provide a good overall view of hospital insights. We used Plotly DASH to create a dashboard instead of an excel dashboard which requires manual labor. The implementation was purely in python and could help future developers to easily work with it.

The project was a real-life use case to work upon. There were various steps involved in the project such as requirement gathering, planning, data handling, UI development, regular meetings, etc. In all of these steps, there were multiple iterations. The importance of team and having regular meetings helped us to understand the problem

statement and complete our project on time.

Working on this project was challenging because of the involvement of multiple vendors and changing requirements but rewarding at the same time as we were able to develop a basic version of the dashboard that can be used by hospitals. After working on this product and after presentation, we got a bigger picture of this project. We learned how different teams can support each other to create a final product.

The project can be summarized in below steps.

- Requirement gathering
- Conducting Interviews
- Form digitization
- Vocabulary creation
- Mock data generation
- Data analysis
- Dashboard creation

This prototype dashboard can be used by hospital management and researchers to track information related to diseases and vaccines. The current implementation is a very basic version and can be modified further. The current dashboard provides information related to patient visits, disease distribution, frequency of mortality, and frequency of morbidity. This product will save a lot of time and effort for employees. They do not have to engage with manual analysis after using this product.

The field lab gave us an opportunity to work with clients. We collaborated with vendors and conducted interviews to get a broader picture of the problem. The interviews provided a lot of useful and essential information for the project. From the interviews with Mariam and Aliya, the main challenges facing healthcare facilities in Africa were clarified, as well as the problems in data organization and data management. Mariam provided us with the sample form which was used for creating the dashboard. The interview with Norbert was very valuable for giving the project a clearer goal. He provided the key elements that the management of hospitals would be interested to see in a dashboard concerning the data of a hospital. The requirement was finalized after we received the final two forms.

The forms were digitized and mock data was generated for two forms which were further loaded in a table and displayed in the dashboard. As part of this project, we created vocabulary. Working with Cedar, bioportal, controlled vocabulary, and dashboard helped us to know the importance of controlled vocabulary while querying information from the database. A controlled vocabulary and structured data can help us to retrieve information from multiple forms and can be shown very clearly to the user. This also helped us to include **FAIR** principles.

The data and controlled vocabularies play an integral role while loading the data in the database and writing queries. We are showing aggregated data and not displaying any sensitive information in the dashboard. A vast amount of time was spent on the data analysis to map various fields as many field names were misleading. We used mock data to populate information in the dashboard and therefore it may not be very representative of real life. Testing with real data would have given more insights to understand the hospital data.

The scope of this project is huge and it can be scaled massively with the right resources and further planning.

8. Recommendations

We have created a basic version of the dashboard in which we have integrated two forms and displayed basic KPIs in the form of bar graphs and charts.

We have implemented the below points in the current project

- *Acquiring forms*:
  We gathered requirements from vendors to understand the importance of dashboards for hospitals. We were given two forms, ”Nigeria HMIS Outpatient Register” and the “Kenya Outpatient Register” to start our development.
- *Digitization of the form*:
  Forms were digitized using CEDAR and BIOPORTAL.
- *Mock data generation*:
  We generated 200 mock forms for data in the JSON format.
- *Dashboard*:
  We have created dashboard using plotly DASH. The implementation was done in python and tables were created using dataframe. Right not we are displaying limited KPIs. The KPIs implemented are listed below:
  - Daily patient visits
  - Admissions
  - Discharge
  - Frequency of diseases

  - Frequency of morbidity
  - Frequency of mortality

We have implemented all the requirements that we had received during the first week of December. The dashboard can be improved further by implementing some more features to make it more user-friendly and informative.

Below functionalities can be added in the next version:

- *Form field*:
  Right now we are displaying aggregated information from two forms. We have added a filter on dates. Similarly, we can create a dropdown for forms. Users can select a particular form and get more details about that form.
- *More features*:
  In this project, we implemented only a few KPIs but in the next version we can add additional KPIs and show more information in the form of graph and charts.
- *More Filters*:
  A user-friendly UI should help users to retrieve the exact information in an efficient manner. For example, if a person needs information related to a particular disease from a particular hospital, they should be able to retrieve that information by selecting appropriate criteria. In the future, we can implement additional filters in the form of dropdowns and textbox to query the data more easily. Right now data can be filtered based on the date criteria.
- *Real data*:
  Mock data was generated to populate information in the dashboard. Mock data is not representative of real-life data. We can try to gather real-life data by collaborating with African hospitals. They can mask sensitive personal information and pass on other details. So, one of the goals of the future is to test the application with real data to get more insights and then deploy it in production.
- *Forms*:
  The current dashboard was built upon the two forms "Kenya form" and "Nigeria form". Having access to more forms can help us to have more variety of data. Testing with more data will also push this project to have additional servers for data management. In future, it would be very interesting to acquire more forms from the hospitals and integrate them with the dashboard.
- *Centralized Repository for data management*:
  We can create a central repository to manage the data from different forms and different countries. We can assign different access to different country.
- *Integrating machine learning algorithms*:
  The data from the dashboard can be used by machine learning algorithms to find interesting patterns in the data. Models can be trained to find the relation between various features. We can also apply LSTM and RNN models to predict the death of an ailing patient. This would require a lot of data for training.
- *Security*:
  Security is an important aspect of any application. Access level and authentication protocols will help to secure data and will give the right access to the right people. Hospital data contains sensitive information and they need to be dealt with extra care. We are already following the FAIR policy but in the future creating a login page for user authentication and assigning different access to users based on their roles will make data and application more secure.

A prototype of the dashboard has been developed by our team and there are still many places where functionalities and features can be added to the dashboard to make it more efficient and user-friendly.

9. Annexes

9.1. Interview Tools

All interviews were computed using zoom. There were no other tools needed for the interviews.

9.2. Interview questions

1. Introduction
   - Can you tell me something about your role?
   - How many years of experience do you have?
2. State of the art
   - What do you think are the main issues in the healthcare facilities in Africa?
   - What does exist now in terms of dashboard and data collected?
3. Needs and requirements
   - Who will be more interested to see data in the dashboard in your opinion?
   - What is information you think is useful to see in the dashboard?
   - What differences can be found in different countries regarding data and needs?
4. Implementation challenges
   - What do you think can be the risks of implementing this solution?

5. Data and technical tools
    - What are in your experience the most common methods/tools used for querying data from Json format?
    - What is important to keep in mind when creating synthetic data?
    - What are in your experience, the most common tools to create a dashboard?

9.3. Interview Transcripts

Both interviews were conducted were not recorded. Therefore, there are no transcripts available for the interviews. However, a group member completed note-taking and the notes can be found below.

9.3.1. Interview 1: Aliya Aktau and Mariam Basajja

1. Responsibilities
    - Aliya Aktau: architecture components to create software and tools that need to speak to each other. Making sure that all the components are in place and interacting with each other. Training data stewards for machine readible data.
    - Mariam Basajja:similar responsibilities as Aliya. Architecture responsibiliteis. Implementing technical coordination.
2. Main Challenges and Issues for Healthcare Facilities?
    - data organization and management
    - Data is mainly paper based.
    - Data is not integrated .
    - Facilities in the different African countries have a standard tool DHS software. This is available for some facilities with internet access they have a limited amount of time to log in.
    - Facilities do not have access to data analysis .
    - Some health facilities are lacking technical infrastructure, example: no computers in place , wifi, etc Data stewards need to be trained
    - Creating semantic vocabularies for analytics.
3. Who will be most interested in this product? Doctors, nurses management?
    - Stakeholders, aka the management , they have specific goals they are interested in for example: to handle the different medical restocking
    - Data management personnel .
    - Beneficial for doctors themselves to see a picture of a patient. Note: an internal dashboard is more wise.
4. Envisioned product and use of Digitilized data ?
    - Most facilities are using HMS systems.
    - Data is entered into the system and the ministry of health is the one that manages the system .
    - Analytics must be completed .
    - There is no system currently being used , mostly just for storage . Enter the data , but don't complete any analysis .
    - retrieval of files.
    - This kind of system is not currently implemented anywhere in the world.
5. Who should have access to which data and at what level ?
    - Forms approved by the ministry of health.
    - Some structure and information might differentiate between forms and people who need access.
    - This solution will be the basis for a more scaled system for different countries, data, forms and personnel.
6. What are the challenges we will encounter to create something that will be scaled in different countries and data?
    - There is no data. We need mock data.
7. Should the dashboard be internal or external ?
    - Internal dashboards can have different permission levels for doctors and nurses, where not all personnel can get an entire history of the patient.
    - External dashboard needs different access levels where no information is provided about individuals. All information must be aggregated data. External dashboards can have numbers on diagnoses, data about a service date, (i.e. for october we have this number of patients with certain diseases , Number of patients that have facilities in a day)
8. Implementation
    - Synthetic data or mock data - it is nice to have some kind of real distribution, normal distribution
9. Querying .json data, What kind of tools and methods are commonly used ?
    - CEDAR provides some documentation on how to query certain templates
    - Jupyter notebook file where you send a request to a certain template with a key
    - Get all the medical forms that have been generated from that template then you can get number of patients , etc, number of diagnosis

9.3.2. Interview 2: Nobert Sebbuggwawo

1. What is your role in the hospital ? What are your responsibilities?
    - Norbert is an IT manager.
2. What is your current patient filing method? What does your data storage and management look like ?
    - There is already a reporting tool that uses aggregate data. Basic hospital management system is already in place.They have a server and a database integration .
    - Extract data using excel and upload it in different analytics.
    - Looking for querying system. DHA is a government reporting system that has aggregate data such as vaccination rates, cases of HIV .
3. Who will use the tool?
    - Management will be interested in knowing how the business is doing Patient numbers, turnaround time, what services we are offering , where we can improve, attract more patients to the facility.
    - Case clinic and KIT are private hospital and profit oriented.
4. What level of privacy would be required?
    - Depends on the kind of data in the dashboard.
    - Patient diagnosis (names and diagnosis) breaches privacy issues.
    - Diagnosis with no names is okay , no identifying characteristics.
    - Aggregate data will not raise any privacy concerns .
5. What would be the necessary features? What metrics would be useful to display or calculate?
    - Blood Pressure
    - Number of diseased patients per week
6. What do you foresee are some challenges of implementing such a tool in your current system?
    - There will be no challenges in infrastructure .
    - Training users will not be a problem , people will be able to play around with different visualizations and learn the software.

9.4. Tools developed for or used in the project

The tools and programs that were used in this project are the following:

- CEDAR
- Python
- Jupyter notebooks
- Dash package in python

References

2017. URL `https://dash.plotly.com/`.

Anneke Fitzgerald, J. and Dadich, A. Using visual analytics to improve hospital scheduling and patient flow. *Journal of theoretical and applied electronic commerce research*, 4(2), 2009. doi: 10.4067/s0718-18762009000200003.

Hidders, J., Paredaens, J., and den Bussche, J. V. J-logic: a logic for querying json, 2020.

Musen, M. A., Bean, C. A., Cheung, K.-H., Dumontier, M., Durante, K. A., Gevaert, O., Gonzalez-Beltran, A., Khatri, P., Kleinstein, S. H., and OConnor, M. J. e. a. The center for expanded data annotation and retrieval. *Journal of the American Medical Informatics Association*, 22(6): 1148–1152, 2015. doi: 10.1093/jamia/ocv048.

Olaronke, I., Abimbola, S., Ishaya, G., and Janet, O. Interoperability in healthcare: Benefits, challenges and resolutions. *International Journal of Innovation and Applied Studies*, 3(1), 2013.

Petkovic, D. Json integration in relational database systems. *International Journal of Computer Applications*, 168(5): 14–19, 2017. doi: 10.5120/ijca2017914389.

Price, M., Singer, A., and Kim, J. Adopting electronic medical records: are they just electronic paper records?. *Canadian family physician Medecin de famille canadien*, 59(7):322–329, 2013.

Raghupathi, W. and Raghupathi, V. Big data analytics in healthcare: promise and potential. *Health Information Science and Systems*, 2(1), 2014. doi: 10.1186/2047-2501-2-3.

Stausberg, J., Koch, D., Ingenerf, J., and Betzler, M. Comparing paper-based with electronic patient records: Lessons learned during a study on diagnosis and procedure codes. *Journal of the American Medical Informatics Association*, 10(5):470–477, 2003. doi: 10.1197/jamia.m1290.

Weissgerber, T. L., Winham, S. J., Heinzen, E. P., Milin-Lazovic, J. S., Garcia-Valencia, O., Bukumiric, Z., Savic, M. D., Garovic, V. D., and Milic, N. M. Reveal, dont conceal. *Circulation*, 140(18):1506–1518, 2019. doi: 10.1161/circulationaha.118.037777.