Cystic Fibrosis Data Integration and Front-End Database Development

Yiqiong Xiao

November 2019

1 Introduction

Cystic Fibrosis is a hereditary disease with a life-limiting autosomal recessive manner (Brian, 2009). This disease is commonly caused by the mutation in a gene called Cystic Fibrosis Conductance Transmembrane Regulator (CFTR), which is the gene that mainly functions as the chloride channel. Changes in CFTR lead to dysfunctional CFTR protein. This protein abnormality is followed by an exorbitant amount of vicious secretions production in the airway of human lungs and in the ducts of the pancreas, resulting in further inflammation and tissue damage (Brian, 2009). As the disease aggravates, it would also affect individuals' organ systems such as lung, pancreas, kidney or liver. External factors like diet, environmental conditions and infectious exposure could alter the severity of symptoms in humans (McCallum, 2000).Although the symptoms may vary from patient to patient, in the long term, symptoms normally are shown as difficulties in breathing and coughing.

The complex nature of Cystic Fibrosis increases potential challenges when diagnosing and applying treatments. Collecting and storing data via accurate and reliable channels are crucial. Recently, many researchers are conducting research related to Cystic Fibrosis in different aspects. Through time, data accumulates, increasing the importance to securely store data. Developing a systematic approach in CF research, integrating experimental and clinical data, and creating a database for further usages are imperative. Data coming from other sources might have different standards and formats, which is necessary to clean and handle missing values of the data before data integration. Some data might have overlaps with others and there might exist some inconsistencies in the data that might need to be cleaned before putting the database. This project is going to focus on combining clinical data from medical records and laboratory data from conventional and omics platforms to provide powerful and flexible access for statistical and modeling efforts by individuals and teams. After the completion of constructing a database, the main focus would shift to GUI application, graphical user interface. It is a form of user interface that allows people to interact with electronic devices through graphical icons.

2 Aims

The overall aim of the project is to build a database and incorporate high-level applications to support Cystic Fibrosis related scientific research. The database will host experimental and clinical data from multiple laboratory units from Georgia Tech and Emory University on Emory AWS, also known as Emory Amazon Web Services. It provides cloud computing platforms and APIs which can store data and launch servers. A GUI interface would support some basic features such as

uploading and downloading data and data filtration. To manage and maintain a cloud-base high-quality computational infrastructure that supports access to integrated clinical, lab and research data, we aim to preserve the confidentiality of all subject data and deploy our data to the secured cloud base data warehouse on AWS. We will combine clinical, laboratory and research data from patient records with all other Georgia Cystic Fibrosis Research and Translation (Georgia CF R&T) and develop a secure, user-friendly interface that allows all researchers to explore the data in the context of other program findings.

3 Data Sources

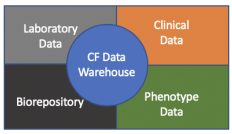

Figure 1: CF Data Warehouse

From figure 1, this project will capture data from different sources. The primary source of clinical data comes from CFF registry as well asa electronic medical records. But currently, the main source of the data in this project right now is from PortCF. It contains overall about 500 CF adult patients. In the near future, we would expect to receive more data from our many partnering CF-research labs. The data from these laboratories would mainly be experimental data. Although the data we used in the project are all adults, we might also obtain data in regards to children. And our focus is going to switch to integrating the data.

4 Future Plans

Brown lab has developed a database that is written in Python interface and a python library, SQLAlchemy, for SQL database. We used MySQL as the back-end for developing and debugging. The current back-end database structure building is completed. We have added the 77 patients information from PortCF to our database. I also plan to switch from MySQL to PostgreSQL for future database maintenance. My main focus for the next step would be integrating the data we would be obtaining in the future and developing a GUI application for user convenience. As the data increases, rather than working under command-line environment, it would be beneficial to have an interface for data visualization .

1) AWS instances setting

The database is now hosted under AWS. After opening HIPPAA compliant AWS account under Emory AWS and created private instance, we are able to deploy the existing database on EC2. In the next step, we would continue our focus on EC2. Since all the information of patients are sensitive, it is fundamental to include security actions, otherwise patients' personal information would be under a great risk.

2) GUI application

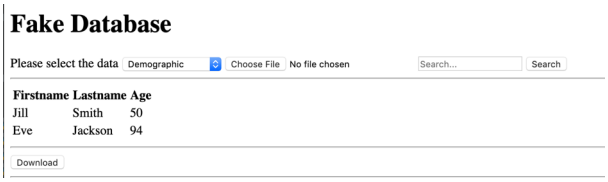

Figure 2: GUI Prototype

Currently, I am trying to build a GUI for the database. As the main goal of developing such

database is to help and save time for researchers, it is necessary to develop an interface that allows them to visualize the data in a webpage instead of working in the command line environment. Here I plan to use Django, HTML and Javascript together to write the whole GUI application. Ideally, in the future, this GUI application should provide basic features such as user login, view data, filter data and download their data of interest, as well as more advance features such as uploading data to the database. According to figure 2 above, it is a simple webpage that we created. This "Fake Database" is used as our prototype for mainly testing results. Right now, it is able to render data from database to the front-end, as well as allow users to upload data from the page, parse the data and send it to back-end database. It would update the page and add the new data to the front end.

3) Data cleaning and data integration

Although the construction of the basics of the database is completed, it would be more optimal if there is a function that could clean and handle missing data prior to putting the data into the database. By pre-processing the data, fewer error would be detected when we update the database. It would generate more uniformed data and increase the quality and the overall productivity. With that being accomplished, achieving the next data-integration step would require less effort. Ideally, we would like to combine data coming from multiple labs at different institutions, Emory hospital and all Georgia CF core center cores and etc. These data would be linked together by unique IDs in regards to the study, model or patient that they are corresponded to. Due to the complexity of data, a unique global identifier is required to help and link everything together. As I mentioned above, if this could be achieved, then the database would continuously grow along with more and more input data.

4) User login access

Another important functionality in GUI application is the access control. A secured database needs to verify that only authorized users are allowed to view all the data while others are limited to certain public data. In order to achieve this, I am going to create a desensitized replicated database for general view and securely store the user information to determine the what can a user see. As for now, a basic structured webpage that is able to show the data in the database and allows the user to upload to data is written.

5) Machine Learning and Statistical analysis

These days, with a large amount of data later integrated to the database, clinical prognostic models can be derived or inferred. These models could be useful resources to base on and can inform critical diagnostic and therapeutic decisions. With that being said, it would be ideal if we could provide API to machine learning tools like AWS Sage Maker, and let the user train and adjust the model or let users download the model itself and run it directly on our server. Additionally, basic statistical tools will be included as well. It would be helpful for simple data analysis. These tools could be used to provide users a summary of the data without downloading the whole dataset. This part would be embedded in the GUI application after the base of GUI is completed.

5 Timeline

1) 12/01/2020 — 03/15/2020: GUI & Data Integration With the simple basic webpage existing, it would potentially take another 10-12 weeks to comprehensively finish developing the GUI application along with allowing the data integration. This would be taking the most time, since it includes displaying data and also storing user information and a series of related security issues.
2) 03/16/2020 — 05/31/2020: Tools development After the 2-3 months of GUI development, our focus would shift to developing statistical tools and machine learning of the data that we hold. That would take about 4-8 weeks to complete.

6 References

[**1**] Alaa, A. M., & van der Schaar, M. (2018). Prognostication and Risk Factors for Cystic Fibrosis via Automated Machine Learning. *Scientific Reports*, 1-19.

[**2**] Brian, O. (2009). Cystic fibrosis. *Lancet*, 1891-1904.

[**3**] Bruce, M. (2017). Cystic Fibrosis Foundation Patient Registry 2017 Annual Data Report.

[**4**] Bethesda, Maryland: 2018 Cystic Fibrosis Foundation.

[**5**] Buzzetti, R., Maffeis, P. (2014). Cystic Fibrosis Database (CFDB): A new web-based tool for cystic fibrosis specialists. *Pediatric Pulmonology*, 938–940.

[**6**] Castellani, C. (2013). CFTR2: How will it help care? *Paediatric Respiratory Reviews*, 2-5. John, M. (2013). Cystic Fibrosis Carrier Screening. *Paediatric Respiratory Reviews*, 270-275.

[**7**] Zhao, C. Y., Hao, Y., Wang, Y., Varga, J. J., Stecenko, A. A., Goldberg, J. B., Brown, S. P. (2019). Microbiome data enhances predictive models of lung function in people with CF. bioRxiv, 656066.