Thesis Portfolio

Digitization of Surgical Flowsheets
(Technical Report)

Big Data and Inequity in Healthcare
(STS Research Paper)

An Undergraduate Thesis

Presented to the Faculty of the School of Engineering and Applied Sciences
University of Virginia - Charlottesville, Virginia

In Fulfillment of the of the Requirements for the Degree
Bachelor of Science, Department of Engineering Systems and Environment

Sarah Rambo
Spring, 2020

Table of Contents

Sociotechnical Synthesis

Data can be an incredibly important tool in improving medical care and better developing medical research and knowledge. But what happens when no infrastructure is in place to collect data? Or on the other hand, how do we ensure that data is used in ways that are fair and equitable? Through my technical project and STS research I look at data in both developed and developing data infrastructures in Rwanda and the United States and examine not only the importance of utilizing data in healthcare, but also the importance of understanding and working with the potential equity issues in the data.

My technical project focuses on optimizing and improving a data digitization system built for low and middle income countries (LMIC). Due to various barriers such as cost and existing infrastructure many healthcare systems in LMICs record medical data manually on papers. This method for data collection does not allow for easy aggregation and analysis of data, and therefore makes it nearly impossible for these providers to gain insights from data. My project worked to improve an existing system at the University Teaching Hospital of Kigali (CHUK) to digitize surgical records. One of the team's main improvements was the development of a mobile upload application to replace an existing web application that not only had several technical issues, but was also very inefficient in uploading data. In addition to redevelopment of the upload application we made several improvements throughout the process including changes to hardware. With the upgraded system users should be able to upload surgical flowsheets at a much faster rate, ideally leading to more data uploaded.

Where my technical work focuses on creating a data infrastructure to allow for more efficient data analytics, my STS research moves into looking at how the use of sophisticated data analytics has the potential to further perpetuate inequities in healthcare, namely in regards to

inequities concerning people of color. As I discuss in my technical project, the use of data analytics can be a powerful tool in healthcare, and combined with big data tools can help provide powerful insights. However, without proper consideration, biases in outcomes and predictions may form that may be harmful and may disproportionately affect minority racial groups. In my research I explore possible root causes of biases in these analytical methods and then present some solutions to combat the given issues.

Both my technical project and STS research delve into the use of data analytics in a healthcare setting, from the creation of data infrastructures to allow for analysis, to ensuring equity once robust and complex data systems are created. With my team's work we hope that the increased efficiency to upload data will increase the amount of data and quality of data collected, so the hospital may have a higher capacity to use such data for further research and clinical care. Furthermore, through my STS research I've discussed several reasons as to why big data may perpetuate biases and inequity and some potential solutions.

Finally, I would like to thank my technical project team members, Mary Blankemeier, Johnny Radossich, and Charlie Thompson, as well as my advisor Professor Sean Ferguson, for their guidance and support.

Digitization of Surgical Flowsheets

A Technical Report submitted to the Department of Engineering Systems and Environment

Presented to the Faculty of the School of Engineering and Applied Science
University of Virginia - Charlottesville, Virginia

In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science, School of Engineering

Sarah Rambo
Spring, 2020

Technical Project Team Members
Mary Blankemeier
John Radossich
Charles Thompson

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments

Signature: _____ Date: _____
Sarah Rambo

Approved: _____ Date: _____
Donald Brown, Department of Engineering Systems and Environment

Abstract

Five billion people, from disproportionately low and middle-income countries, are unable to access safe, timely, and affordable surgical and anesthesia care [1]. Patients in Africa are twice as likely to die after surgery when compared with the global average for postoperative deaths [2]. Most of this mortality happens after surgery, and it is therefore imperative to identify patients at high risk of complications and rapidly intervene. The perioperative mortality rate (POMR) has been identified by the World Health Organization as a global measure of the quality of surgical procedures. Perioperative data collected during surgery can predict adverse surgical outcomes. Access to such data is essential for decreasing perioperative mortality rates and improving medical treatment. In many low and middle-income countries, perioperative data is manually recorded on paper flowsheets, restricting the ability to discover medical trends. This method of data collection inhibits easy and efficient data aggregation and analysis. Thus, systems put in place to digitize these flowsheets are key in utilizing perioperative data to improve overall healthcare. By streamlining the digitalization of intraoperative flowsheets, more data will be collected while minimizing the time while optimizing the quality. In order to optimize the digitization process, the research team has made several improvements to the current system. One of the largest improvements includes a complete redesign of the digital upload process in the form of a mobile app that integrates scanner functionality and upload capability into one convenient and efficient step, thereby reducing devices and platforms needed for doctors and hospital staff to upload a flowsheet. This redesign also provided increased user feedback and corrected issues in which flowsheet uploads failed. In addition to creating a new mobile upload platform, improvements were also made to the SARA (Scanning Apparatus for Remote Access). SARA is a wooden box used to ensure the consistency of images captured. The design of the box

provides a standardized distance, lighting, and background for each scan, improving readability. Testing is currently being performed to improve the reliability of the current lighting setup. Possible replacement power supplies are being examined for durability, ease of repair, and functionality. Additionally, usability testing and evaluation is being completed to measure increases in successful task completion and decrease in time and steps required. The goal of this project is to design a system to digitize the information contained in surgical flowsheets at the University Teaching Hospital of Kigali in Rwanda in the most efficient and effective manner. To accomplish this goal, the research team reduced the time and devices needed to upload a surgical sheet by 78% and 50%, respectively. Hardware and software malfunctions were fixed, and the longevity of the system was improved as procedural checklists to upkeep and correctly utilize the system were implemented.

Introduction

Health metrics such as perioperative mortality rate (POMR) are important measures in determining the quality of surgical care and safety [3]. In order to determine important metrics such as POMR, medical data in digital form is ideal for analysis. However, low and middle income countries face several barriers in implementing electronic record systems that have become common in higher income countries for reasons including, but not limited to, lack of funding, lack of infrastructure, and legal hurdles. Therefore, the use of electronic health records and data is much lower in LMIC with only about 15% of low-income countries adopting electronic health records while over 50% of high-income countries have adopted EHRs [4]. Thus many hospitals in LMICs utilize paper records or flowsheets to record data. In order to transform data recorded on physical sheets into electronic data that is more readily accessible for analysis

there have been some attempts at creating digitization systems. However, while these systems may address many of the root issues preventing adoption of electronic health records, considering and optimizing their efficiency and ease of use is imperative in increasing user adoption and accurate data collection. In order to implement digitization systems, care must be taken to ensure that these systems operate efficiently and accurately. This paper details the improvements and optimization of the digitization system currently in place at the University Teaching Hospital (CHUK) in Kigali, Rwanda that was created through a collaboration between teams at the University of Virginia and providers at CHUK. While this system addresses many of the challenges in digitizing medical data in low and middle income countries, there are key issues relating to the system's efficiency, accuracy, and ease of use. In the past year approximately 375 sheets have been digitized through the system, but increased system efficiency would aid in digitizing CHUK's large collection of physical sheets, helping to provide more historical data for analysis. In order to optimize the digitization process we have replaced the web application with a mobile application that decreases the time and steps required to upload images and provides clearer user feedback. Additionally, while this paper focuses solely on the system in place at CHUK, this approach could be replicated at other hospitals and medical providers in LMIC facing similar obstacles in medical data collection and records.

Prior Work/Literature Review

In 2020 Rho et al. [5] at the University of Virginia implemented a system to digitize surgical flowsheets at University Teaching Hospital of Kigali, in Kigali, Rwanda. The system begins with the Scanning Apparatus for Remote Access (SARA), a wooden box with lighting that provides consistency in lighting, distance, and angle of images. Inside the SARA box is a

tray to hold surgical flowsheets at an appropriate distance, and a battery powered light source. The top of the SARA box has a small round hole for images to be taken through. The group chose to implement the SARA device due to concerns from doctors about resource availability and maintenance for a more traditional scanner. Images are then typically taken using a mobile phone or tablet with a third party scanning app, such as Tiny Scanner, and are then sent to a computer. From the computer the user accesses the image upload web application where images are sent to a UVA email address, where images are then downloaded and processed to extract data which is then populated into a PostgreSQL database.

While the use of electronic medical records in Africa falls behind world averages, there has been a marketable increase in usage since the early 2000s. This increase has been driven by increases in computer ownership and increased internet access, which increased 2,357.3% between 2000 and 2010 in Africa [6] . In many cases the adoption of electronic medical records in Africa is driven by research collaborations between African health institutions and international institutions, namely regarding HIV and AIDS research [7]. A 2017 review of the adoption of electronic health records in sub-saharan Africa outlines four key barriers to adoption: high implementation and maintenance costs, limited computer skills, lack of constant internet connectivity, and lack of prioritization of EHRs [8]. While these barriers may not consistently have significance in all LMIC, one or several are likely to stand as an obstacle to adoption.

Researchers at Vanderbilt University developed and deployed electronic data collection systems in a Kenyan tertiary hospital. The system was built upon the Research Electronic Data Capture (REDCap), a free, internet-based data collection tool. Due to constraints on internet access the researchers created an asynchronous version of the system to operate while internet connection was unavailable. This system allowed a shift from manual to electronic data

collection, allowing for additional data analysis and reporting. While this approach was successful, resources vary greatly in different LMIC that may act as a barrier to implementing a similar solution [9].

In a 2012 paper from Amity University in India, researchers created a system to scan and digitize electrocardiogram (ECG) graphs. ECG results are typically printed onto thermal paper, these records are then typically kept in storage for future reference. In this study the researchers note using the camera in a mobile phone for scanning the images. Once scanned the images were further processed using Laplacian filtering, a method to reduce background noise in the image, and by color based segmentation to create a binarized image of only black and white pixels [10].

While many hospitals and providers move toward completely digital records, the Khayelitsha Hospital in Western Cape, South Africa, has created a dual physical and digital system. Due to legal restrictions from the South African government the hospital currently requires “hardcopy” documentation, thus has implemented a large scale digitization system [11]. The hospital system includes a system where handwritten notes or documentation are placed in folders, these documents are then transcribed by clerks into the electronic record system, and then are finally scanned and stored. The folders of records are then stored as well, therefore, three separate records of each document exist, the physical copy, the transcription, and the scanned image [12] While this system has been successful in this hospital it relies on large investments in not only software and technical infrastructure, but also human resources to scan and transcribe all documents.

System Design

The current system incurred software malfunctions, insufficient understanding and communication of the current system between stakeholders tied to the University of Virginia and CHUK, inefficiencies for medical personnel in CHUK, and hardware failures.

Within the existing applications, there were issues within both the backend of the code and within the frontend's user experience. One of the most significant issues on the backend involved sheets uploaded through the web application were often "lost" in that they were uploaded into the web application's internal storage and were encrypted, but the process failed when sending the file through the email. Sheets were able to be manually recovered through the back end of the application but could not be connected to a patient identification number. In October 2020, 24% of uploaded sheets to the web app never successfully sent through email. Significant changes were needed to avoid losing nearly a quarter of the data uploaded.

A lack of storage within the existing web application resulted in a manual reset approximately every four months. The application only had 512 MB allowed on the free level, and the design of the application stored all sheets uploaded within the app. When the storage maximum was met, no more sheets could be successfully uploaded, and application administrators had to manually delete files from the application, which was time-consuming and required continual monitoring and attention. With these issues in mind, the mobile application was designed so that the images are not saved locally within the application after upload, creating no strain on available storage.

Besides constant monitoring required to ensure storage space was available, the free application host required that the application be refreshed every three months to keep the website active. This again is resolved with the move to a mobile application. While many of these

backend issues could be resolved with revisions to the web application, several other overarching issues involving user experience still existed from incomplete tabs and pages within the web application. To address these challenges of efficiency, intuitiveness, and effectiveness of the current digitization upload process, a new Android mobile Android app was developed.

A. Mobile App Development

Incorporating the entire upload process into a singular app will significantly decrease the time and steps required for each uploaded patient medical record. The upload process begins with a login screen. The user must log into the application with a provided username and password that is authenticated by Google's Firebase authentication service. Users must request login credentials from the research team as a measure to ensure higher security standards. The login screen is just one of several security measures that were implemented in order to uphold patient confidentiality regarding their medical data.

Once the user is successfully logged in they are directed to a minimalist home screen where they can begin the upload process. The design of the home screen serves to reduce the amount of time and clicks required to submit a sheet and to create an easy and intuitive interface for any user. The previous web application included several tabs, most of which were not functional. Furthermore, the web application required the image to be manually sent to an email account prior to being manually uploaded to the application, creating an extra step in the process. For a standard upload, the process only requires entering one identifier, taking one image, and clicking three buttons. To upload a flowsheet image, the user must first enter the patient's identification number. This number is different from the patient's MRNO, as a measure to protect patient privacy. Patient identification numbers are unique numbers that are randomly assigned

using an excel function by the users in Rwanda. To ensure data privacy, the true MRNOs are not known by anyone further in the process.

After the user has entered the patient identification number they must indicate the side of the sheet that is being uploaded using the radio buttons. These buttons concatenate an identifier which labels the surgical sheet as Intraoperative or Anesthesia. The inclusion of these buttons help to reduce issues in the previous system where there was no way to differentiate between sheets. In the future this function could be expanded to include options for other classifications of sheets as well. Unlike the previous system, this mobile application does not allow the user to incorrectly advance without filling out the patient identifier and surgical sheet side.

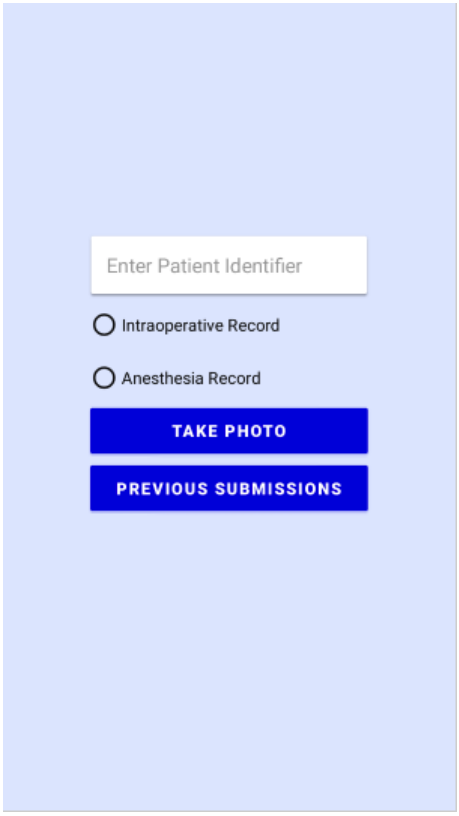

Figure 1. *Mobile App Login Page*

Figure 2. Mobile App Upload Page

Once the information is entered, the user can click the “take photo” button which will launch the device’s camera. Using the camera functionality that is integrated into the application the user can “scan” the flowsheet using the app and the SARA box. Once the user clicks *send*, the image is instantaneously encrypted to a dedicated University of Virginia email address. The subject line of each email is the patient identification number that the user enters prior to opening the camera. Once the image is taken the user has the option to discard the image and take another, go back to the home page, or to continue. In terms of the encryption method chosen, the app utilizes the Advanced Encryption Standard (AES). AES encryption is a symmetric-key algorithm that utilizes block cipher to achieve data encryption. The algorithm takes a bitmap data type and converts the image into an array containing each byte of the image [13]. Byte by byte, these pieces of information are encrypted to ensure that should the image fall into the wrong hands, it could not be viewed without the private key which is shared only with collaborators at CHUK and the University of Virginia.

Continuing through the flow of the system, the encrypted image will be decrypted at the start of the digitization process. Following this decryption, the image will be subject to several image processing tools in order to ensure the data will be translated with maximum accuracy. To start, edge detection methods will be passed in order to ensure that none of the data is cut off and missing in the database. Following this, the image will be converted to grayscale in order to maximize contrast and facilitate the natural language processing algorithms used to convert the images to hard data.

Results and Discussion

A. Upload Process Improvements

With the integration of the camera into the upload application, the need for a third-party app is removed, and the number of devices required for upload are halved. The previous system required users to use a third-party scanning app on a mobile phone or tablet and then send the scanned image to a computer for upload. The mobile application includes pop-up messages throughout the user's experiences, updating them on accurate login information, requiring all patient fields to be filled in, and giving instant feedback once the patient chart is successfully sent to the email. The user experience is dramatically improved by decreasing the number of steps required to upload a photo.

Comprehensive user testing of both systems was performed to compare processes and test the systems' efficiency, intuition, and effectiveness. Participants were randomly assigned a system to test first to remove learning bias. The directions of each system were given to the participant, and the time, fatal errors, and post experimental survey were recorded. As this was only a test of the scan and upload process, the sheet remained in the box across all tests. This eliminated possible error from users having varying comfort with using the box itself.

The directions for each system are listed below:

Web Application

1. Open an external scanning app on phone
2. Scan sheet
3. Send scan to specified email
4. Download image on computer
5. Upload image to web application
6. Enter patient information

7. Send

Mobile Application

1. Enter patient information
2. Take photo
3. Approve photo (Sends automatically upon approval)

Compared to the old system, the average time to upload a sheet is reduced by 79% from 3 minutes to 40 seconds when using the newer, mobile application. The number of fatal errors where the user is unable to proceed without being prompted reduces from 89% to 0%. 100% of the users preferred the mobile application. In a post-usage survey, the two most common answers to the question "What do you think was the hardest/worst part of the process when using the web app?" related to the manual emailing of the file and the upload process. These two actions of that system are performed automatically in our app, thereby alleviating the core user complaints. Without the handicap of having to send the files manually, the team believes the upload time following a quick learning curve will lower the average upload time per sheet. In the medical field, any amount of time saved is essential to ensure all patients are being attended to and other work is getting done. This was the driving force behind the design of the system - to decrease the amount of time to upload a single medical record in order to save the doctor's time and incentivize thorough data collection.

B. Hardware Improvements

Beyond the implementation of a new upload application, other changes were made to the system as a whole and to the SARA device. One limitation of the original construction is the use of advanced cutting technology, namely waterjet cutting, to build precise pieces that were sent to Rwanda. This results in a high transport cost and a lack of reproducibility. As such, we sought to

develop a version of the SARA box that was reproducible with tools and materials that were easily obtainable in Rwanda. To do so, we built a box by hand with the same dimensions and lower-quality plywood, using basic power tools. We retained the original lighting system for consistency. Even though this box was built to much higher tolerances than the previous model, the image quality was the same as the previous box. The scanner app also appeared to have better cropping capabilities in this box, which led us to investigate the effect of contrast on cropping.

The most significant limitation of the physical scanning process was consistently incorrect cropping. We were able to identify two suspected sources of this error. The first was an incorrect distance from the camera lens to the sheet being scanned. On some phones being used, the scope of the image either did not include the full sheet, or filled to the exact edge. This prevented the image from being properly scanned and led to data loss. The initial design of the box utilized a tray to raise and lower the sheet, with screw mounts being used to adjust the height. In examinations of usage practices in Rwanda, we noticed that the screw mounts were not being used at all, and the tray laid on the bottom of the box for maximum distance. This led us to test the efficacy of eliminating the tray altogether. While this ensured the entire sheet was contained in each image, the cropping was still imperfect. This led us to our next correction for cropping error.

This second error was the contrast between the sheet being scanned and the wood of the tray. The type of plywood used was very lightly colored and may have hindered the scanning app's ability to identify the sheet outline. When constructing the second variation of the box, a darker shade of plywood was used, and the cropping was consistently correct. This led us to hypothesize that the darker frame color would improve cropping capabilities, and as a result we used black tape to border the sheets in our full-app testing. Using an exact replica of the box

being used in Rwanda, with a border of black tape and the tray removed, we saw zero crops that were too small and/or resulted in data loss.

While the intention of the SARA device was to ensure consistency in lighting, image angle, and cropping, there was great variability in these metrics in images that were taken using SARA. In order to address these problems, we incorporated a number of smaller changes with the lighting, alignment guides, and training materials.

Dr. Christian and his team at CHUK encountered a lighting failure within SARA and were unable to fix it with a lack of instructions and understanding of the current lighting system. The importance of clear documentation and communication of the system highlighted the necessity and creation of checklists for every part of the process: mobile application, web application, SARA box, and lighting system.

Using both the pictures of the lighting system in Rwanda and materials left from the previous Capstone team, we found typical 1.5 Volt C batteries were connected in series producing 3 Volts from lead to lead. For a white LED battery, the required voltage ranges from 3 Volts to 5 Volts. Battery voltage diminishes with remaining energy, so 1.5 Volt output on each C battery does not last long and therefore may barely reach the full 3 Volts minimum before decreasing. Voltage can also decay in storage and other conditions. This could be the sole reason for the lighting issue, given LED light bulbs won't operate unless you exceed their required voltage. As a result of the findings the lighting system was redesigned. A short and long term solution was developed. In the short term two LED strip lights, velcro, and a small tube was shipped to Rwanda. This is a rechargeable light that uses the same USB cords as the standard Android phones used throughout CHUK. Sold in a pack of two, one can be in the box as the

other one is charging. This was tested and proved to ensure the quality of light is not overexposing nor underlighting the box.

The long-term lighting solution is leveraging a 12 Volt motorcycle battery to power the LED bulb. This has several advantages. First off, motorcycles and mopeds are common in Africa. Assuming the majority of these vehicles contain a battery, this system should be replicable anywhere. The socket and 12 Volt LED bulb draw 3 Amperes, so under ideal scenarios, we should have 10 hours of use. Furthermore, this system could be a potential solution for repeatable scanning in remote regions. Since a motorcycle battery charges while the engine is on, if a remote clinic has a camera, they can take the battery out of a motorcycle or moped, connect to the system, scan the documents they need to scan, and put the battery back in their vehicle.

Conclusion and Future Work

After evaluating the descriptive scenario of the current medical record upload system, there were several areas of improvement that the team decided to study and optimize. Through this research, we were able to develop and implement a functional application for the digitization and encryption of documents, specifically intraoperative flowsheets for the purpose of the study. We succeeded in improving upon the web application constructed by Rho et al. by reducing the average process time by 79%, and average occurrence of inhibiting errors by 89%. This drastically improves the efficiency of the digitization process. Additionally, by removing the web application altogether and developing our system on the Android operating system, we have removed the limiting need for a computer. This allows for a higher number of users in the hospital and easier integration into standard operating procedure.

Usability of the box was also improved through the identification of the sources of cropping errors. Analyzing the medical records that have been scanned and uploaded via the web application, the team found insight into the reasons for images with poor quality (these typically came in the form of poor contrast, dark images, or sheets with cut-off edges. Overcropping and imperfect scanning resulted in data loss on many sheets examined through the previous system, and preventing this will improve the value of our system.

Future work will consist of the maintenance and improvement of the application, as well as the continued improvement and simplification of the box design. The adoption and longevity of this system will rely on the ease of understanding for a wide range of users, and the repeatability of box construction. Further research on the SARA box can explore the accessibility of certain components across developing nations, and the continuation of a construction manual that requires only locally procurable parts.

Further research on the application can examine adding multiple language support to expand the user base to non-english speakers. Additionally, research can improve upon the user interface of our application. The sole criticism we received on our application during testing was that for the older generation, and those less familiar with technology, the application may be difficult to navigate. Therefore, there is room for further development. However, with the simplistic nature of the application and backend system design, the team believes that any doctor can adjust to the slight learning curve and drastically decrease the amount of time spent digitizing their patient medical record database in its current state.

References

[1] M. R. Felizaire et al., “Perioperative Mortality Rates as a Health Metric for Acute Abdominal Surgery in Low- and Middle-Income Countries: A Systematic Review and Future Recommendations,” World Journal of Surgery, vol. 43, no. 8, pp. 1880-1889, Apr. 2019, doi: 10.1007/s00268-019-04993-1

[2] B. M. Biccard et al., “Perioperative patient outcomes in the African Surgical Outcomes Study: A 7-day prospective observational cohort study”, Lancet, vol. 391, no. 10130, pp. 1589–1598, Apr. 2018, doi: 10.1016/S0140-6736(18)30001-1

[3] J. L. Rickard, G. Ntakiyiruta and K. M. Chu, "Associations with Perioperative Mortality Rate at a Major Referral Hospital in Rwanda," *World Journal of Surgery*, vol. 40, no. 4, pp. 784-790, Apr. 2016, doi: 10.1007/s00268-015-3308-x

[4] “Electronic health records.” World Health Organization. https://www.who.int/gho/goe/electronic_health_records/en/ (accessed Apr. 8, 2021).

[5] V. Rho et al., "Digitization of Perioperative Surgical Flowsheets," 2020 Systems and Information Engineering Design Symposium (SIEDS), Charlottesville, VA, USA, 2020, pp. 1-6, doi: 10.1109/SIEDS49339.2020.9106679.

[6] M. O. Akanbi, A. N. Ocheke, P. A. Agaba, C. A. Daniyam, E. I. Agaba, E. N. Okeke, and C. O. Ukoli, “Use of Electronic Health Records in sub-Saharan Africa: Progress and challenges,” *Journal of Medicine in the Tropics*, vol. 14, no. 1, pp. 1-6, 2012

[7] M. Kumar and J. Mostafa, "Electronic health records for better health in the lower- and middle-income countries: A landscape study," *Library Hi Tech*, vol. 38, no. 4, pp. 751-767, March 2020, doi:10.1108/LHT-09-2019-0179

[8] F. F. Odekunle, R. O. Odekunle, and S. Shankar, “Why sub-Saharan Africa lags in electronic health record adoption and possible strategies to increase its adoption in this region,” *International Journal of Health Sciences*, vol. 11, no. 4 pp. 59-64, Sept.-Oct. 2017

[9] B. Sileshi, M. W. Newton, M. S. Shotwell, J. P. Wanderer, M. Mungai, J. Scherdin, P. A. Harris, S. H. Vermund, W. S. Sandberg, and M.D. McEvoy, “Monitoring Anesthesia Care Delivery and Perioperative Mortality in Kenya Utilizing a Provider-driven Novel Data Collection Tool,” *Anesthesiology*, vol. 127, no. 2, pp. 250-271, Aug. 2017, doi: 10.1097/ALN.0000000000001713

[10] D. K. Garg, D. Thakur, S. Sharma, and S. Bhardwaj, “ECG Paper Records Digitization through Image Processing Techniques,” International Journal of Computer Applications, vol. 48, no. 13, pp. 35-38, Jun. 2012

[11] R. Weeks, "The successful implementation of an Enterprise Content Management system within the South African Healthcare Services Sector," 2013 Proceedings of PICMET '13:

Technology Management in the IT-Driven Services (PICMET), San Jose, CA, USA, 2013, pp. 2590-2597.

[12] E. C. Ohuabunwa, J. Sun, K. J. Jubanyik, and L. A. Wallis, “Electronic Medical Records in low to middle income countries: The case of Khayelitsha Hospital, South Africa,” African Journal of Emergency Medicine, vol. 6, no. 1, pp. 38-43, Mar. 2016, doi: 10.1016/j.afjem.2015.06.003

[13] “Java AES Encryption and Decryption,” Baeldung, https://www.baeldung.com/java-aes-encryption-decryption, (accessed Apr. 8, 2021)

Big Data and Racial Inequity in Healthcare

A Research Paper submitted to the Department of Engineering and Society

Presented to the faculty of the School of Engineering and Applied Sciences
University of Virginia - Charlottesville, Virginia

In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science, School of Engineering

Sarah Rambo
Spring 2021

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments

Signature: _____ Date: _____
Sarah Rambo
Approved: _____ Date: _____
Sean Ferguson, Department of Engineering and Society

Introduction

“Big Data" and concerns over data privacy and ethics has been a popular point of controversy in recent years, yet many people do not understand what big data is, its applications, and how it may be affecting them and their health. In a 2019 review of the use of big data in healthcare, researchers found exponential growth in the amount of research and publications regarding the topic from 2010 to 2018 (Pastorino et al., 2019). A 2014 report found that the United States healthcare system alone had over 150 exabytes of data stored and was soon projected to approach the zettabyte and yottabyte scales (10^{21} and 10^{24} gigabytes respectively) (Raghupathi & Raghupathi, 2014). With such a vast amount of data available researchers have been able to leverage new technologies to advance medical knowledge and care at an increasing pace. With big data becoming such a force in the healthcare industry it is important to understand several things about the technology. How is big data used in medicine? Who does it help? Who does it hurt? And how can we decrease the consequences of using such tools? The use of big data in healthcare has a high potential for positive impacts, however, the use of the tools also has a significant potential to further perpetuate inequities and biases within the healthcare industry against those in underrepresented communities, namely Black and Indiginous People of Color (BIPOC). The potential for these consequences stems mainly from a lack of representation in data and poor model construction. In order for big data tools to be used in a manner that is safe and equitable for all, more oversight and regulation must be put into place.

What is "Big Data?"

In order to understand the nuances of the utilization of "big data" in healthcare, we first need to understand what the term big data refers to. Many working with big data today still rely on Gartner's definition which states that "big data is high-volume, high-velocity, and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation" (2001). Volume refers to the amount of data, which in the case of big data can be massive. Velocity refers to the speed at which the data is collected, which may be near real-time in some cases, and variety refers to the types of data which is collected, which could be in structured forms like traditional databases, or in unstructured forms such as audio recordings, video, etc. Working through this definition we can generalize that big data refers to exceptionally large and complex data sets that often require more sophisticated methods of analysis, typically complex algorithms to process. The techniques and tools utilized to analyze and work with these datasets also have a great variety. Techniques such as predictive analytics, machine learning, artificial intelligence (AI), and more are commonly employed to uncover deeper insights from the data. ("Data-driven healthcare", 2013). Some methods commonly employed in analyzing big data are neural networks and black box algorithms. In these models the internal processes between input and output are a "black box" and cannot be seen or explained, even by those building the model. These models can become incredibly complex and can be extremely powerful in identifying patterns in large amounts of data, but also may become increasingly unpredictable due to their lack of transparency (Bleicher, 2017).

How is Big Data Used in Healthcare?

Big data analytical methods are powerful tools that can have a profound impact on the healthcare industry and medical field. A 2018 European Union commission on big data in public health and healthcare identified four key areas where the utilization of big data tools could be especially impactful. First, that big data could increase "earlier diagnosis and the effectiveness and quality of treatments" through finding early signs and improved disease intervention. Next, big data could widen "widening possibilities for prevention of diseases" by identifying risk factors. Third, "improvement of pharmacovigilance and patient safety" by helping to improve the ability to make informed medical decisions. Finally, the commission identified that big data could be instrumental in prediction of outcomes of diseases (Pastorino et al., 2019). While these areas of benefit are nonexhaustive of the potential positives, they help to illustrate areas where the use of big data could be pivotal in improving treatment and patient care.

There are numerous examples of big data tools improving medical knowledge. For instance, researchers and doctors at Columbia University were able to use streams of physiological data from patients with brain injuries from strokes and aneurysms to detect severe complications over 48 hours earlier than with traditional methods. In another case the Hospital for Sick Kids in Toronto, Canada used sophisticated analytics on vital-sign data collected from bedside devices over 1,000 times per second and were able to detect signs of a potential infection up to 24 hours earlier ("Data-Driven Healthcare", 2017).

To reap these benefits sophisticated models have been developed and utilized, including in many cases black box models. These black box models are valuable in providing insights, but also are controversial in medicine due to their lack of transparency (Price, 2018).

Beyond improving medical knowledge and care, the use of big data in healthcare could also bring about large financial and economic benefits. McKinsey estimated that the usage of big data tools in healthcare could save an estimated $300-450 million dollars over one year in the United States (Kayyali, Knott, & Kuiken, 2020).

Potential for Bias and Inequity

Although all people regardless of any characteristics or personal attributes should receive fair and equitable medical care, it has been well documented that implicit and explicit biases have significant impacts on care for many minority groups. A review of bias in healthcare found widespread bias affecting care based on race, sexual identity, socioeconomic status, and more (FitzGerald & Hurst). In a study of implicit bias in healthcare, Blair et al. found that implicit biases may affect clinical judgements and that treatment outcomes may also be affected by differences in judgement (2011). While biases and inequity in healthcare are nothing new, with the advent of big data analytics a new method for perpetuating bias has been created. In cases where black box models are utilized it becomes much harder to ensure that conclusions and predictions are not being made in ways that will perpetuate bias, these models lack the transparency to ensure equity and fairness. Similarly, many models are built in ways that fail to consider the interconnectedness of health and other factors.

For instance, in one case evidence was found that a commercial algorithm used by many United States health systems had strong racial biases. The algorithm, which assigned levels of risk to patients, was often used in determining if patients would be referred to more specialist care. However, this model used money spent on health costs as a predictor for patient health, and on average less money is spent on Black patients with the same level of need. Therefore, the

algorithm incorrectly identified healthier white patients as higher risk than their sicker Black counterparts, and therefore the number of Black patients identified for extra care by less than half. This results in sicker patients, specifically Black patients receiving care they need less often (Obermeyer et al., 2019).

While cases such as this one are almost certainly not intentionally biased, what they do fail to do is understand the more complex relationships between cultural, societal, and economic factors that may play a role in someone's health. This inadequacy in considering biases and differences creates a model that is not effective and may consequently cause harm toward groups of people.

Diversity in Data

One of the most prominent contributors to perpetuating biases and inequity with big data in healthcare is the overall lack of diversity in healthcare data.

Recognizing a need for diversity in healthcare data, the National Institutes of Health (NIH) Revitalization Act was passed in 1993. This act worked to require the NIH to ensure that all federally funded research include women and minorities and that research participant characteristics be disclosed as well (Chen et al., 2014).

While the NIH Revitalization Act was an important step in recognizing and addressing the need for diversity in medical data, nearly thirty years later it's intentions may not have been fully realized. A 2015 study from researchers at UC San Francisco found that less than 2% of over 10,000 cancer studies and less than 5% of respiratory studies have enough minority participants to be statistically significant (Bole, 2021). Similarly a review of genome-wide association studies (GWAC), which have been a popular tool for uncovering genetic factors in

common diseases, found that 80% of participants overall were of European descent, with most non-European participants coming from populations of Asian descent (Popejoy & Fullerton, 2016).

Given the legislative action requiring the inclusion of minority groups in clinical research, why is it then that little progress has been made over the last three decades? To begin, some of the bias toward a lack of diversity may come from logistical and historical factors. For instance, in the case of GWAS, researchers may prefer to use existing and established datasets, such as the Framingham Heart Study, which has followed multiple generations of participants in Framingham, Massachusetts since 1948 to study cardiovascular disease. While these established and often large datasets might provide simplicity for researchers, they often are not diverse or representative of populations as whole (Popejoy & Fullerton, 2016).

Furthermore, a large reason for lack of diversity may be lack of geographic diversity in study participants. In 2020 researchers at Stanford found that most medical data used in publications about AI and machine learning from the previous five years came from only three states, with 71% of data being from California, Massachusetts, and New York. They also found that thirty-four states were not represented at all in any publications reviewed (Lynch, 2020). With lack of geographic diversity data findings and insights gleaned from these models may reflect ethnic, cultural, educational, and social features that are not representative of other populations, and therefore may come to poor conclusions (Kashual et al., 2020).

Moreover, a 2015 study explored potential reasons why minority groups may often be left out of clinical research and found several key reasons for the lack of diversity. First they found that minority participants are often more likely to face barriers that prevent them from being offered the opportunity to participate in clinical studies. These barriers may range from lack of

transportation to lack of insurance, to lack of access to specialists who may serve as referral sources for the studies. Furthermore they found, many minority groups are more likely to harbor higher levels of skeptiscsm and distrust in clinical research, largely stemming from past abuses such as the Tuskegee Syphilis Experiment, or potentially personal incidents of discrimination while recieving care. Researchers found that Black Americans and Native Americans were most likely to feel distrust due to historical mistreatment (Durant et al., 2014).

Finally, in addition the researchers perhaps most likely to consider and reach minority groups, those who are a part of those communities, are often awarded research grants at much lower rates. A 2012 study of race and NIH research grants found that Black researchers are over 10% less likely to be awarded NIH funding than white researchers (Ginther et al., 2011). Increasing equity on the researcher side may help potentially increase or bring better attention to increasing equity and representation in participants.

While it is clear that there is an overwhelming lack of diversity and representation in healthcare data, why does this actually matter, and how does this lack of representation actually hurt underrepresented and minority populations? Machine learning and AI systems are built and "trained" using preexisting data sets. These methods use this data to learn and identify hidden patterns in the data to make predictions or conclusions about the data. If the data used is not representative, this could mean that predictions or conclusions from the data do not take into account potential differences ("What is Machine Learning", n.d.). Char et al. discuss how issues such as lack of diversity in data, and "overfitting" a model may provide skewed results and estimates (2020). For example, in some cases, specific racial or ethnic groups' genetics may play a role in their risk or reaction to certain diseases or treatments. For example, genetic variations found in those with ancestry specific to certain regions in Africa have a much higher incidence

rate of and are at a higher risk for kidney disease than those of European or Asian descent (Genovese et al, 2010). Similarly, potentially up to 75% of those with Pacific Islander ancestry are unable to convert clopidogrel, an antiplatelet drug, into its active form and therefore are at a much higher risk following angioplasty (Wu et al. 2015), a procedure that helps to restore the normal flow of blood through an artery ("Angioplasty and Stent placement for the heart", n.d.).

With such clear and drastic consequences for oversights and lack of prioritization in diversity, there is a real need to improve the accessibility of participation for a more representative population. As long as the data used to build models is not representative of all relevant populations, it will be difficult to ensure any type of equity.

Improving Representation and Decreasing Bias

As noted many of the issues causing an increase in inequity in healthcare stem from logistical, historical and systemic factors that will not be easily fixed or changed. However, there are steps that can be taken to decrease the potential consequences. In order to apply big data tools to the healthcare industry without worsening inequality in health, users must make a serious and concerted effort to thoroughly evaluate and address potential sources of bias. Several studies and researchers have laid out guidelines or frameworks on evaluating appropriate usage of big data in healthcare. As a specific example, Ibrahim, Charleson, and Neill (2020) lay out a five point approach to evaluation. They assert that in order for big data tools to be used effectively in a manner that promotes healthcare equality, algorithms and tools must (1) explicitly measure relevant equity criteria, (2) explicitly include a "fairness criteria" as part of the model design, (3) emphasize variation in treatment effects within the patient population, (4) prioritize use of transparent algorithms or avoid "black box" algorithms, and (5) conduct algorithm audits to

combat areas with lack of transparency. Frameworks such as this are essential in ensuring safe and equitable healthcare as the use of big data increases. However, frameworks such as this are not enough. While taking the diversity included into a model is an important and positive step, more regulation and prioritization of collecting and using more diverse and representative data from the beginning will also help to combat biases in models. While the NIH Revitalization Act should have increased diversity in clinical research, in practice there have been subpar improvements. In a review of diversity in clinical research Oh et al. (2015) explain that the NIH's approach to prioritizing and monitoring sex and gender inclusion through their Office of Research on Women's Health should be applied to ensuring higher rates of racial and ethnic representation. Additionally, they assert that research applications from minority-serving institutions should be judged more based on their capacity to do research, rather than their past track record of engaging in research. Institutions with stronger ties to minority communities often have a higher capacity to recruit and retain participants from underrepresented groups.

Finally, while creating more oversight and inclusion are necessary for decreasing inequity and bias caused by big data in healthcare, simply bringing more attention and education to the issue could be beneficial. While attention and education alone will not fix the problem, it is a clear step in working towards a solution. Many of the people building these models and conducting research are not people of color or are not from underrepresented groups and may not be heavily considering and understanding the potential for bias their work has. Ideally with more education on the issue, researchers would feel more compelled to more seriously consider diversity and its lack thereof in their work.

Even beyond moral and ethical benefits of decreasing the lack of diversity in health data, there is potential for significant long-term economic benefits as well. A 2011 study (LaVeist et

al., 2011) found that over $230 billion in medical spending and over $1 trillion in indirect costs related to illness and death could have been saved between 2003 and 2006 by eliminating health disparities for minorities.

Overall, there is no easy fix to solve these issues, big data is used in so many different ways, even just within the healthcare field, but with these recommendations, steps can be taken to reduce the potential consequences felt by underrepresented groups and can help to increase equity in healthcare.

Conclusion

Big data analytics has the incredible potential to improve healthcare practices, save money, and improve patient's outcomes and lives. However, as noted these benefits are not without potential for harm, especially toward minority communities, specifically communities of people of color. In order to ensure equity and safety for all people in healthcare and in medical treatment work needs to be done to improve the usage of big data in healthcare in terms of representation and consideration in model building. With a high emphasis on racial justice, equity, and equality over the past year, ignoring clear issues revolving around minority communities in healthcare, would be an injustice. An overall fix, or to make the healthcare industry perfectly equitable and unbiased is a complex and multifaceted problem, but ensuring that the use of powerful and impactful tools and big data do not further perpetuate biases is a positive step toward decreasing disparities.

References

Angioplasty and Stent placement for the heart. (n.d.). Retrieved from https://www.hopkinsmedicine.org/health/treatment-tests-and-therapies/angioplasty-and-stent-placement-for-the-heart

Bleicher, A. (2017, August 09). Demystifying the black box that is ai. Retrieved from https://www.scientificamerican.com/article/demystifying-the-black-box-that-is-ai/

Bole, K. (2021, April 01). Diversity in medical research is a long way off. Retrieved from https://www.ucsf.edu/news/2015/12/401156/diversity-medical-research-long-way-study-shows

Char, D. S., Abràmoff, M. D., & Feudtner, C. (2020). Identifying ethical considerations for machine learning healthcare applications. The American Journal of Bioethics, 20(11), 7-17. doi:10.1080/15265161.2020.1819469

Chen Jr, M. S., Lara, P. N., Dang, J. H., Paterniti, D. A., & Kelly, K. (2014). Twenty years post‐NIH Revitalization Act: enhancing minority participation in clinical trials (EMPaCT): laying the groundwork for improving minority clinical trial accrual: renewing the case for enhancing minority participation in cancer clinical trials. Cancer, 120, 1091-1096.

Clayton, J. A., & Collins, F. S. (2014). Policy: NIH to balance sex in cell and animal studies. Nature, 509(7500), 282–283. https://doi.org/10.1038/509282a

Data-driven healthcare organizations use big data analytics for big gains. (2013). Retrieved from https://www.ibmbigdatahub.com/whitepaper/data-driven-healthcare-organizations-use-big-data-analytics-big-gains

Definition of Big Data - Gartner Information Technology Glossary. (n.d.). Retrieved November 01, 2020, from https://www.gartner.com/en/information-technology/glossary/big-data

Durant, R. W., Wenzel, J. A., Scarinci, I. C., Paterniti, D. A., Fouad, M. N., Hurd, T. C., & Martin, M. Y. (2014). Perspectives on barriers and facilitators to minority recruitment for clinical trials among cancer center leaders, investigators, research staff, and referring clinicians: enhancing minority participation in clinical trials (EMPaCT). Cancer, 120 Suppl 7(0 7), 1097–1105. https://doi.org/10.1002/cncr.28574

FitzGerald, C., & Hurst, S. (2017). Implicit bias in healthcare professionals: A systematic review. BMC Medical Ethics, 18(1). doi:10.1186/s12910-017-0179-8

Genovese, G., Friedman, D. J., Ross, M. D., Lecordier, L., Uzureau, P., Freedman, B. I., Bowden, D. W., Langefeld, C. D., Oleksyk, T. K., Uscinski Knob, A. L., Bernhardy, A. J., Hicks, P. J., Nelson, G. W., Vanhollebeke, B., Winkler, C. A., Kopp, J. B., Pays, E., & Pollak, M. R. (2010). Association of trypanolytic ApoL1 variants with kidney disease in African Americans. Science (New York, N.Y.), 329(5993), 841–845. https://doi.org/10.1126/science.1193032

Ginther, D. K., Schaffer, W. T., Schnell, J., Masimore, B., Liu, F., Haak, L. L., & Kington, R. (2011). Race, ethnicity, and NIH research awards. Science (New York, N.Y.), 333(6045), 1015–1019. https://doi.org/10.1126/science.1196783

Ibrahim, S. A., Charlson, M. E., & Neill, D. B. (2020). Big Data Analytics and the Struggle for Equity in Health Care: The Promise and Perils. Health equity, 4(1), 99–101. https://doi.org/10.1089/heq.2019.0112

Kaushal A., Altman R., & Langlotz C. (2020). Geographic Distribution of US Cohorts Used to Train Deep Learning Algorithms. JAMA. 324(12):1212–1213. doi:10.1001/jama.2020.12067

Kayyali, B., Knott, D., & Kuiken, S. (2020, March 01). The big-data revolution in US health care: Accelerating value and innovation. Retrieved October 18, 2020, from https://www.mckinsey.com/industries/healthcare-systems-and-services/our-insights/the-big-data-revolution-in-us-health-care

LaVeist, T. A., Gaskin, D., & Richard, P. (2011). Estimating the economic burden of racial health inequalities in the United States. International journal of health services: planning, administration, evaluation, 41(2), 231–238. https://doi.org/10.2190/HS.41.2.c

Lynch, S. (2020, September 14). The geographic bias in Medical AI Tools. Retrieved from https://hai.stanford.edu/news/geographic-bias-medical-ai-tools

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019, October 25). Dissecting racial bias in an algorithm used to manage the health of populations. Retrieved from https://science.sciencemag.org/content/366/6464/447.full

Oh, S. S., Galanter, J., Thakur, N., Pino-Yanes, M., Barcelo, N. E., White, M. J., de Bruin, D. M., Greenblatt, R. M., Bibbins-Domingo, K., Wu, A. H., Borrell, L. N., Gunter, C., Powe, N. R., & Burchard, E. G. (2015). Diversity in Clinical and Biomedical Research: A Promise Yet to Be Fulfilled. PLoS medicine, 12(12), e1001918. https://doi.org/10.1371/journal.pmed.1001918

Pastorino, R., De Vito, C., Migliara, G., Glocker, K., Binenbaum, I., Ricciardi, W., & Boccia, S. (2019). Benefits and challenges of Big Data in healthcare: an overview of the European

initiatives. European journal of public health, 29(Supplement_3), 23–27. https://doi.org/10.1093/eurpub/ckz168

Popejoy, A. B., & Fullerton, S. M. (2016, October 12). Genomics is failing on diversity. Retrieved from https://www.nature.com/news/genomics-is-failing-on-diversity-1.20759#/b1

Raghupathi, W., & Raghupathi, V. (2014). Big data analytics in healthcare: promise and potential. Health information science and systems, 2, 3. https://doi.org/10.1186/2047-2501-2-3

What is machine learning?: How it works, techniques & applications. (n.d.). Retrieved from https://www.mathworks.com/discovery/machine-learning.html

Wu, A. H., White, M. J., Oh, S., & Burchard, E. (2015). The Hawaii CLOPIDOGREL lawsuit: The possible effect on clinical laboratory testing. Personalized Medicine, 12(3), 179-181. doi:10.2217/pme.15.4

Prospectus

A Data Infrastructure for Global Perioperative Outcomes: Digitization of Perioperative Surgical Flowsheets

(Technical research project in Systems Engineering)

An Examination of the Use of Big Data in Healthcare Systems

(STS research project)

By Sarah Rambo
Fall 2020

Technical Project Team Members:
Mary Blankemeier
John Raddossich
Charles Thompson

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments

Signature: _____ Date: _____
Sarah Rambo

Approved: _____ Date: _____
Dr. Donald E. Brown, Department Engineering Systems and the Environment

Approved: _____ Date: _____
Dr. Sean Ferguson, Department of Engineering and Society

Introduction

Throughout the past 10-20 years the use of electronic medical data has grown exponentially. In most high income countries it is commonplace for most medical and health records to be recorded digitally (Zeng, 2016). With this ever expanding amount of healthcare data comes many potential benefits and consequences. Big data in healthcare has the capacity to improve medical knowledge, inform best practices, and decrease medical spending. However, these analytical methods also have the potential to perpetuate biases and possibly worsen inequalities in healthcare (Gianfrancesco, Tamang, Yazdany, & Schmajuk, 2018).

Meanwhile, the use of electronic medical records (EMRs) has not been implemented as widely in most low to middle income countries (LMIC), potentially widening the gap in quality of care. Low and middle income countries often face barriers in adopting new technologies that are not as pervasive in higher income countries. To help lessen these barriers, my capstone team is working on a project to develop and improve a data infrastructure for the University Teaching Hospital of Kigali (CHUK). Located in the capital city of Rwanda, a small landlocked country of under a million people in central Africa. Although Rwanda is still a relatively poor nation, they have made great strides in recovering from the devastating 1994 genocide. In rebuilding the country and economy, Rwanda has prioritized healthcare accessibility and equity to very positive results (Binagwaho et al., 2014). With Rwanda's major investment in health, providing an infrastructure for medical data could help providers more effectively measure health metrics and provide important insights into healthcare practices.

This paper aims to examine the potential benefits and challenges associated with the application of big data analytics in a healthcare setting, and to explore implementation of EMR digitization in Rwanda.

Technical Project

Research suggests that between two and five billion people globally do not have access to safe surgical or emergency healthcare. With this lack of access many of these people are vulnerable to unnecessary death, injury or complications. In 2010, it was estimated that 32.9% of deaths worldwide were associated with conditions that required surgical care (Meara et al., 2015). Among leading surgical and anesthetic societies perioperative mortality rate (POMR) has emerged as a leading indicator of access to safe surgical care (Watters et al., 2014). In order to compute metrics such as POMR, or gain insight into other trends related to care and best practices, reliable access to medical data is imperative, however, in many LMIC, most healthcare records are still recorded on paper, decreasing the flexibility and usability of the data. Due to various barriers such as structural, legal, or financial constraints only about 35% of lower-middle incomes countries, and about 15% of low income countries have adopted the use of EMRs (Akhlaq, Mckinstry, Muhammad, & Sheikh, 2016; WHO, 2019). Over the course of this year my team and I are working to combat these barriers to improve a system to digitize medical records at CHUK.

The technical advisor for this project is Dr. Don Brown of the Department of Engineering Systems and the Environment. Additionally, Dr. Marcel Durieux of the University of Virginia School of Medicine provides close support and guidance to the team. My team is continuing work done over the course of the 2019-2020 school year by a previous Systems Engineering Capstone group.

The procedure in which physical medical records are converted to electronic medical records involves three key steps: transmission, image processing and storage. In the current system the first phase of this process involves scanning the physical medical record using a

scanning app on a mobile device along with the Scanning Apparatus for Remote Access, or SARA, which is a wooden box specially designed by the previous capstone team to ensure consistent lighting and cropping in scans. These scans are then typically transferred by the user to a laptop or desktop computer and uploaded through a web app to be encrypted and sent to UVA. Once uploaded and decrypted, the second phase, image processing, begins. In this phase algorithms crop the sheet into several sections, then read and extract checkbox, graph and handwriting data from the image. This year a team of Master's of Data Science students will be primarily responsible for improving the image processing phase of the system. Finally, the storage phase of the system involves cleaning and storing the extracted data and returning data to CHUK. In the current system there is no real process in place to handle this final step.

Our goal as a team is to improve upon the system created by the previous capstone group in order to streamline and optimize the upload process for users in Rwanda, and to handle issues and inefficiencies that have arisen within the current system. Through our research and conversations with users and stakeholders we have identified several key issues. First, is a high level of inefficiency and required equipment to upload sheets. Currently the chart upload process requires several steps with several devices with additional equipment. In order to upload a sheet the user must have a device with the scanner app downloaded and access to the SARA, they then typically must transfer images from a mobile device to a laptop or desktop computer in order to upload on the web app. In our project we propose to handle this issue by developing a single app that integrates the scanner functionality, streamlining the process so that users may easily upload records using only a mobile device. In decreasing these inefficiencies we may not only be able to decrease the time and effort needed to upload a record, but then in turn may be able to increase adoption rates of the technology and increase the amount of data collected. Furthermore,

decreasing the required equipment and devices may help to increase the flexibility and scalability of the system, therefore increasing the potential for expanding the system to additional locations.

Another issue that we have identified as a priority is the current lack of user feedback presented in the system. We plan to provide immediate feedback within our app to users to let them know what medical records have been successfully uploaded and read, and to provide them with basic statistics and data in an easily accessible manner.

In all with our proposed app the entire system will be streamlined so that doctors will be able to scan and upload images all within one app on one device, thereby increasing upload efficiency and availability of EMRs.

STS Topic

In order to understand the nuances of the utilization of "big data" in healthcare, we first need to understand what the term big data refers to. Many working with big data today still rely on Gartner's definition which states that "big data is high-volume, high-velocity, and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation" (Gartner, 2001). Working through this definition we can generalize that big data refers to exceptionally large and complex data sets that often require more sophisticated methods of analysis, typically complex algorithms to process. These methods of analysis can act as a catalyst for incredible growth in medical knowledge and practice, but also have the capacity to cause great harm and increase inequities in healthcare.

I will examine both the potential benefits and consequences of the usage of big data in healthcare, and then evaluate a framework for complex data analytics methods to be implemented in an equitable and safe manner.

As the usage of big data increases, potential benefits of the practice are becoming increasingly evident in the healthcare sector. One major benefit of the usage of big data in medicine is the potential medical advancement that could be made, which could improve patient care and best practices (Raghupathi, 2014). For example, in one study at Columbia University, researchers analyzed "complex correlations" within physiological data from patients with brain injuries. They found that with these advanced analytics they were able to predict and diagnose serious complications up to 48 hours sooner than with standard practices (IBM, 2013).

Besides improvement in patient care and scientific advancement in medicine, the use of big data in healthcare can also create positive changes in the costs and effectiveness of healthcare systems. McKinsey recently reported that the usage of big data tools in healthcare could save an estimated $300-450 million dollars over the coming year (Kayyali, Knott, & Kuiken, 2020).

With so much potential for advancement it is clear to see why big data is becoming a fixture within healthcare. However, looking only at the benefits of big data usage could be incredibly dangerous. Just as biases and lack of transparency in big data presents large issues in other applications such as law enforcement and resource allocation, these issues are also primary concerns in healthcare applications (Crawford, 2013). In many cases those creating analytical tools neglect to consider and address lack of diversity in data sets, along with other oversights in data collection such as issues with sample size and and misrepresentation of groups or characteristics in the data creates large opportunities for resulting algorithms and analysis to reflect only a narrow set of perspectives and backgrounds (Gianfrancesco, Tamang, Yazdany, &

Schmajuk, 2018). This lack of consideration increases biases that may be harmful to underrepresented populations, oftentimes racial minorities and/or those who are socioeconomically disadvantaged (Ibrahim, Charlson, & Neill, 2020).

In order to apply big data tools to the healthcare industry without worsening inequality in health, users must make a serious and concerted effort to thoroughly evaluate and address potential sources of bias. Several studies and researchers have laid out guidelines or frameworks on evaluating appropriate usage of big data in healthcare. As a specific example, Ibrahim, Charleson, and Neill lay out a five point approach to evaluation. They assert that in order for big data tools to be used effectively in a manner that promotes healthcare equality, algorithms and tools must (1) explicitly measure relevant equity criteria, (2) explicitly include a "fairness criteria" as part of the model design, (3) emphasize variation in treatment effects within the patient population, (4) prioritize use of transparent algorithms or avoid "black box" algorithms, and (5) conduct algorithm audits to combat areas with lack of transparency (Ibrahim, Charlson, & Neill, 2020). Frameworks such as this are essential in ensuring safe and equitable healthcare as the use of big data increases.

Conclusion and Next Steps

In both my technical and STS research projects I am working with data in healthcare. For my technical project my team is focusing on providing the infrastructure to transform handwritten data into an electronic record that can be analyzed to help develop better insights into best practices and other trends that could potentially improve care and medical outcomes for patients. Meanwhile, in my STS research I will be examining the usage of medical data on a larger scale and how implementation of complex algorithms in the healthcare setting can provide

both helpful insights and information to improve care or system efficiency, but also introduce harmful biases or other negative consequences.

Moving forward with my research, by the end of the fall 2020 semester my technical project team hopes to have built the foundation of our app, and have completed necessary research into the components that will be implemented in the application such as a scanner function, necessary security features, and relevant upload feedback. Furthermore, into the spring semester we plan to have a working version of our app deployed so that we may test its functionality and further refine the components and user experience so that we may implement the app in The University Teaching Hospital of Kigali. We hope that our app and system development may provide the tools to improve patient care in the future.

Technical Project Timeline 2020-2021:
Oct.-Nov. 2020: Initial research and development of app and components
Dec. 2020-Feb. 2021: Early phase deployment
Feb.-Mar. 2021: App refinement, revisions and bug fixes
Mar.-Apr. 2021: Initial deployment of app to client for testing
May 2021: Final deployment of app to hospital

STS Thesis Timeline 2020-2021:
Nov. 5, 2020: Submit Prospectus
Nov. 2020-Jan. 2021: Additional research on healthcare data and techno-political issues
Feb. 2021: First Draft of Thesis and subsequent revisions
Mar. 2021: Second Draft of Thesis and subsequent revisions
Apr. 2021: Potential third draft of Thesis and final revisions
May 2021: Final Portfolio submission

References

Akhlaq, A., Mckinstry, B., Muhammad, K. B., & Sheikh, A. (2016). Barriers and facilitators to health information exchange in low- and middle-income country settings: A systematic review. *Health Policy and Planning, 31*(9), 1310-1325. doi:10.1093/heapol/czw056

Binagwaho, A., Farmer, P. E., Nsanzimana, S., Karema, C., Gasana, M., Ngirabega, J. D., . . . Drobac, P. C. (2014). Rwanda 20 years on: Investing in life. *The Lancet, 384*(9940), 371-375. doi:10.1016/s0140-6736(14)60574-2

Crawford, K. (2013, April 1). The Hidden Biases in Big Data. Retrieved from https://static1.squarespace.com/static/5b5df2f5fcf7fd7290ff04a4/t/5b8d8261562fa736b2002818/1536000609606/05+The+Hidden+Biases+in+Big+Data+%28Crawford%29.pdf

Dornan, L., Pinyopornpanish, K., Jiraporncharoen, W., Hashmi, A., Dejkriengkraikul, N., & Angkurawaranon, C. (2019). Utilisation of Electronic Health Records for Public Health in Asia: A Review of Success Factors and Potential Challenges. *BioMed Research International, 2019*, 1-9. doi:10.1155/2019/7341841

Gartner. (n.d.). Definition of Big Data - Gartner Information Technology Glossary. Retrieved November 01, 2020, from https://www.gartner.com/en/information-technology/glossary/big-data

Gianfrancesco, M. A., Tamang, S., Yazdany, J., & Schmajuk, G. (2018). Potential Biases in Machine Learning Algorithms Using Electronic Health Record Data. *JAMA Internal Medicine, 178*(11), 1544. doi:10.1001/jamainternmed.2018.3763

IBM. (2013). Data-driven healthcare organizations use big data analytics for big gains. Retrieved from

https://www.ibmbigdatahub.com/whitepaper/data-driven-healthcare-organizations-use-big-data-analytics-big-gains

Ibrahim, S. A., Charlson, M. E., & Neill, D. B. (2020). Big Data Analytics and the Struggle for Equity in Health Care: The Promise and Perils. *Health Equity, 4*(1), 99-101. doi:10.1089/heq.2019.0112

Kayyali, B., Knott, D., & Kuiken, S. (2020, March 01). The big-data revolution in US health care: Accelerating value and innovation. Retrieved October 18, 2020, from https://www.mckinsey.com/industries/healthcare-systems-and-services/our-insights/the-big-data-revolution-in-us-health-care

Meara, J. G., Leather, A. J., Hagander, L., Alkire, B. C., Alonso, N., Ameh, E. A., . . . Yip, W. (2015). Global Surgery 2030: Evidence and solutions for achieving health, welfare, and economic development. *The Lancet, 386*(9993), 569-624. doi:10.1016/s0140-6736(15)60160-x

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

Reid, M. J. (2017, April 05). Black-box machine learning: Implications for healthcare. Retrieved November 02, 2020, from https://www.polygeia.com/post/black-box-machine-learning-implications-for-healthcare

Rickard, J. L., Ntakiyiruta, G., & Chu, K. M. (2015). Associations with Perioperative Mortality Rate at a Major Referral Hospital in Rwanda. *World Journal of Surgery, 40*(4), 784-790. doi:10.1007/s00268-015-3308-x

Watters, D. A., Hollands, M. J., Gruen, R. L., Maoate, K., Perndt, H., Mcdougall, R. J., . . . Mcqueen, K. A. (2014). Perioperative Mortality Rate (POMR): A Global Indicator of

Access to Safe Surgery and Anaesthesia. *World Journal of Surgery, 39*(4), 856-864. doi:10.1007/s00268-014-2638-4

WHO. (2019, November 08). Electronic health records. Retrieved November 05, 2020, from https://www.who.int/gho/goe/electronic_health_records/en/

Zeng, X. (2016). The Impacts of Electronic Health Record Implementation on the Health Care Workforce. *North Carolina Medical Journal, 77*(2), 112-114. doi:10.18043/ncm.77.2.112