

Avant-Garde HIV Research: Harmonizing and Visualizing Patient Data

Sandy Law

Advisor: Dr. Nadir Weibel
University of California, San Diego
Computer Science and Engineering Building
9500 Gilman Dr, La Jolla, CA 92093
s4law@eng.ucsd.edu

ABSTRACT

The human immunodeficiency virus (HIV), a virus that attacks and compromises the immune system, remains a global epidemic affecting the lives of millions of people. Tracking the transmission networks of HIV may be invaluable in identifying affected or at-risk populations in order to concentrate treatment and prevention efforts. However, the multitude of socio-demographic and phylogenetic factors that influence HIV transmission creates a complex problem to solve. This project looks at the challenges researchers face when handling these multidimensional data sets. In particular, researchers have a need to flexibly access data and perform meaningful analysis despite the complexity of the data set. However, data that has been collected over many years and across different studies and sites, might present issues in terms of harmonization. Furthermore, the multidimensional nature of the collected data creates challenges in terms of understanding the data across dimensions. We present a web-based interface to tackle the data management issues and look at applying Exploratory Data Analysis methods to the analysis of multidimensional HIV transmission factors.

Author Keywords

Avant Garde; HIV; visual; database; interface;

INTRODUCTION

The human immunodeficiency virus (HIV) is a world-wide epidemic that has affected the lives of millions of people. According to UNAIDS global report, approximately 35.3 million people world-wide were living with HIV in 2012 [10]. By identifying affected or at-risk populations, we can more accurately concentrate ongoing treatment and prevention efforts where it can help the most. Any improvement to how quickly and accurately we can identify these “hot spots” could potentially mean preventing transmission of the virus to any number of people. However, the complex socio-demographic and phylogenetic factors that influence HIV transmission can make it difficult to find these populations.

The National Institute of Drug Abuse (NIDA) awarded the Avant-Garde of HIV/AIDS Research Award in 2012 to Dr. David Smith for the purpose of building a novel system that uses various factors such as demographics, viral strains and geographic data to identify HIV transmission networks to find these hot spots [5]. This project aims to contribute in the realization of such a system. We integrate a variety of HIV

data sets from the San Diego (California) region and examine the difficulties that come with dealing with the complexity of this data. We explore the challenges researchers face when handling this multidimensional data, such as maintaining the currency of the data and harmonizing data coming from a variety of sources. After implementing ways to mitigate these challenges, we look to enhance existing methods of analyzing this data by creating a web-based visualization to give researchers ways to interact with HIV data and explore the factors relating to HIV transmission networks.

Scope

The UCSD AntiViral Research Center (AVRC) and multiple medical sites around the San Diego (California) and Tijuana (Mexico) area are conducting studies collecting thousands of data points for every patient participating. These data may be related to patient demographics, medical history, and other relevant data. In combination, this data can prove invaluable to the analysis of HIV transmission networks and, in turn, improve HIV prevention efforts.

For the scope of this project, the HIV patient data set is formed by the AVRC’s Acute and Early HIV (AEH) data and studies such as Amigo, Proyecto El Cuete, Mujer Mas Segura, Hombre Seguro, STAHR, STAHRII, and Parejas. Of all the questions patients answer and data collected, a set of approximately 50 data points have been carefully identified in collaboration with HIV researchers as a good coverage of the HIV factors typically used in analysis. These fields act as starting point to our Avant-Garde project and encompass patient information relating to demographics, medical history, lifestyle, and more. In total, our project draws these fields from 9 unique sources of patient data.

MOTIVATION

The primary motivation of this project is to aid in identifying HIV transmission networks. There are two major components in achieving this goal. Firstly, we wish to address what we believe is a sub-optimal workflow currently in place for HIV researchers at the UC San Diego AntiViral Research Center (AVRC). By creating a system with a clear architecture and easy to use workflow, we remove logistical obstacles such as maintaining current data and requiring extensive manual labor to assemble data from multiple sources. Secondly, we create a visualization tool that provides new ways of viewing and interacting with the multidimensional patient data. The

more intuitive and interactive the visualization, the more control researchers have over the analysis of their data set.

Data Structure and Workflow

As part of the project, we introduce an easy-to-use system to handle the logistical side of a variety of HIV analyses. That creates a clear separation between those collecting the HIV data and those who look to use it, allowing these two parties to operate independently. To reduce the need for manual labor, we want this system to automate any processing that must occur on the raw data, especially combining data from multiple sources. Another consideration is data storage: data should be stored in a manner that can accurately represent relationships the data may have, removing unnecessary repetition, and being extensible in the future.

“Avant-Garde” and Exploratory Data Analysis

Avant-Garde HIV research refers to research that features innovative and creative ideas that has the ability to open entirely new avenues for HIV prevention and treatment efforts.

In typical situations, a researcher forms a hypothesis and then does the necessary data gathering afterwards to examine the topic closer. When collecting very heterogeneous, multidimensional data, it becomes difficult to make sense of the whole picture and spot the patterns to form these hypotheses.

John W. Tukey, one of the primary supporters of Exploratory Data Analysis (EDA) techniques, believes there is value in using data to suggest hypotheses instead [9]. Typically, an EDA approach involves creating visual or statistical models to present raw data that can be used by researchers to form ideas. By providing interactive and exploratory tools, a researcher will be able to view the data in a different way, from fresh perspectives. In this project we embrace Tukey’s approach of EDA and we created a system to aid the discovery of “avant-garde” ideas and hypotheses that could help addressing the HIV pandemic.

DATA COLLECTION AND ANALYSIS WORKFLOW

Medical sites conduct HIV studies in which patients who choose to participate answer a multitude of questions and some clinical data is collected. This data is stored by each remote medical site in their own manner, using formats of their own choosing.

When a UCSD researcher wishes to get access to this data, the remote sites must export their data and send it via email. A researcher must then manually combine the data from the various sources and store it in a flat table structure (i.e. a large CSV file). Only then can analysis such as HIV sequencing begin.

This current workflow is tedious and must be repeated every time new data is sent from medical sites. The manual conversion necessary to keep the data consistent across the different studies is time-consuming. This issue is one of the primary issues we looked to resolve. In addition, this workflow is very linear; someone at the end of the pipeline must wait for all the previous steps to occur before he or she can begin.

OBSTACLES AND STRATEGIES TO SOLVE THEM

Data harmonization, combining data from a variety of sources while maintaining the consistency and integrity of the data, is a large concern of this project. In the current workflow, data from the different remote medical sites must be manually harmonized. In order to automatize or at least simply the harmonization process, a variety of challenges must be addressed with care in consultation with individuals who have experience in dealing with these specific HIV-related fields. This is a key requirement to ensure that the integrated data is valid for analysis being conducted with it. In this section, we detail several major obstacles that arise when harmonizing the HIV patient data and our proposed solution for addressing them.

Disaggregated Data

Although the medical sites ask similar questions, each study has its own goals and is managed by different people. As a result, their answer typically are slightly different. When harmonizing similar questions from different sources, the challenge comes from identifying where the differences are. This often takes the form of two studies asking the same question and receiving the same answer, but that answer is simply recorded in their data differently. For example, the Amigo study asks patients what gender they identify themselves as, and store a “0” to represent a female patient. The El Cuete study asks the same question, but chooses to use a “2” to denote a female patient. Nearly every field within our scope from every unique source of patient data shows such behavior.

To determine what these numerical values mean and to remove any uncertainty, we propose to create a universal “codebook” dictating all official data dictionary for these values. Then, a data mapping process must happen when processing all data points to ensure conformity to this standard.

Changing Data Formats

A number of relatively small transitions have occurred over the years that must be dealt with when harmonizing data. Most of these small transitions comes from one of two types: a change in answers or a change in where the answer is recorded. Sometimes, more options for answers are added to the same question being asked. For example, the AVRC’s AEH study ask patients to identify the gender of any sexual partner who is HIV positive. Prior to 2011, the potential answers made available to the patient differentiates “Transgender” as “Male to Female” and “Female to Male” as options. The later versions of the study do not make that differentiation. Other small transitions come from changes in the field name the data is stored under. For example, the results of a nucleic acid-based test (NAT) is stored in a “result” field prior to December 2013, but afterwards is found in a “nat_result” field.

When harmonizing fields that have either of these behaviors, we must use other fields, such as the date the interview occurred or when the specimens are collected, to determine which version of the question was used or where the results can be found.

Sometimes, changes in the format of how data from studies is recorded can be drastically altered. This is particularly true for the data coming from the AVRC. Approximately five years ago, “BBL” data was stored in a database structure that was effectively abandoned. A new schema was put in use in which all variable names were changed, many questions rephrased, removed, or added, and their answers altered. For example, when recording whether a patient is symptomatic of pharyngitis, the data was stored under “EHXRF2” in BBL format and simply “pharyngitis” in the new format. The old BBL data was only partially imported to the new schema, and there is currently no reconciliation between the two formats.

We felt that the old BBL format was different enough from the new schema that treating them as separate sources was a solution to handling this change. The BBL format would undergo the data harmonizing process as if it was simply another study from a different medical site.

Data Integrity

Perhaps the most important consideration when combining data from multiple sources is maintaining the integrity of the data collected. Each of the medical sites within our scope ask similar questions and receive similar answers. However, there are subtle differences that are important to consider. In particular, when the questions being asked in each study do not map to the other studies in a 1 to 1 fashion.

Many times, differences between a particular question across multiple sites comes down to syntax. When recording a patient’s sexual orientation, there were three ways of asking the question:

1. “What is your sexual orientation?”
2. “Whom do you have sex with?”
3. “Do you think of yourself as heterosexual, homosexual, or bisexual?”

Is the intention of the questions the same? A patient who identifies as heterosexual may still engage in homosexual intercourse.

For another example, we have to determine whether the following three fields gave answers that could be considered equivalent.

1. Site A asks patients for the total number of sexual partners.
2. Site B asks for the number of male sexual partners and the number of female sexual partners.
3. Site C asks for the number of casual sexual partners and whether a regular partner exists.

The question is whether questions 2 and 3 are the sum of their parts. Could we be potentially excluding certain possibilities if we simply add the numbers together to obtain a total number of sexual partners?

A particularly complicated issue deals with questions whose answers are tied directly to a period of time. Each of the data sources within our scope ask questions of the form: “Have

you be diagnosed with chlamydia in the past x months?” for a variety of time-sensitive topics such as various sexually transmitted diseases or frequency of drug usage. There are a few ways to handle this harmonization.

1. Take lowest value of x: “Have you been diagnosed with chlamydia in the past month?”
2. Take highest value of x: “Have you been diagnosed with chlamydia in the past 6 months?”
3. “Have you EVER been diagnosed with chlamydia?”

Each of these options illustrate situations that may violate data integrity. If we take the smallest value of x, a patient who answered “no” to being diagnosed in the last three months may not have answered “no” if asked within the last 6 months or 4 months. If we take the largest value of x, a patient who answered “yes” to having chlamydia in the past 6 months may not answer “yes” if asked within the past month or last three months. If we correlate the data based on whether the patient has ever been diagnosed for chlamydia at all, a patient could say “no” to the past 1-6 months, but we cannot say for certain that he or she has never been diagnosed with the disease ever before.

In our opinion, in order to have meaningful queries for the HIV patient data, some compromise must be made in order to combine these similar yet different questions to a universal field for more fruitful analysis. This may be equating similar questions that perhaps have different intentions (such as the sexual orientation example) or consolidating several questions in one study to better match questions from other sources (the number of sexual partners example). While they cannot be called completely equivalent fields, we believe an approximation can still be useful. Perhaps, an aggregated field can simply be an approximation of an answer instead of us simply being unable to combine the data at all.

Sometimes, there is just no easy solution to combining fields. Situations where an answer is tied to the period of time chosen by each of the studies (such as the chlamydia example), simply cannot be combined into a single data point. Because maintaining data integrity is paramount yet we still want to do some meaningful analysis on these types of questions, we propose treating the data with its time frame as a couplet and showing the data as a pair of two values. By tying the answers directly to the time duration of the question, manipulating the data is more difficult but there is still some potential for querying.

IMPLEMENTATION

The nature of this project call for an agile-like approach to the development process. Periodic meetings and ongoing conversation with medical researchers (particularly Dr. Sanjay Mehta and Christy M. Anderson) were required to ensure we were capturing an appropriate collection of fields from the study, a detailed understanding of the workflow, and an accurate representation of the data collected. Their expertise and familiarity when dealing with HIV medical data is paramount to creating a system that maintains data integrity and also meets our project goals.

Architecture

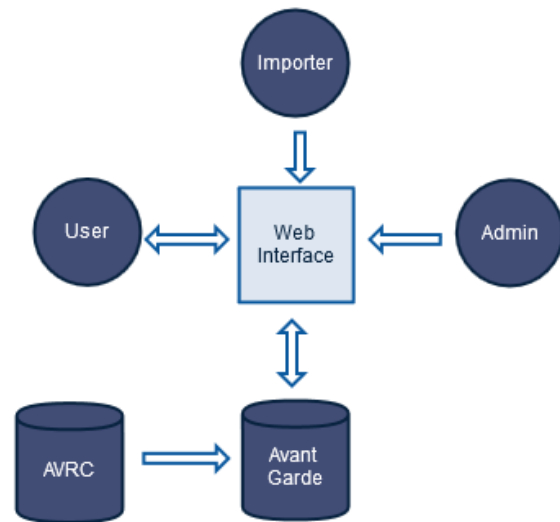


Figure 1. Diagram of the software architecture

The architecture of our system (Figure 1) involves a web interface that displays HIV patient data from a database. Administrators monitor the data and make changes to fix issues as needed. Medical sites import data from studies through the web interface to be stored into our database. HIV researchers view, manipulate, and export data from the interface. An AVRC database periodically updates AEH data in our database.

Data Harmonizing and Mapping

With multiple sources of data in multiple formats, a clear documentation of what the data means is required. In order to satisfy this requirement, we created a codebook to map values from each data source to their counterpart in our database. For our universal codebook, we can then say with certainty that a “1” value in the gender data point will always mean “Male”.

Questions and answers must be combined in a logical manner from the different sources. For example, when asking about a patient’s heroin usage, some studies ask “Injected” and “Non-Injected” heroin separately, while another study does not. In our version, we decided to combine these questions and that decision is then documented. However, this may not work for some questions who cannot combine as easily. Where this is the case, a future extension of this project would be to allow the person querying the data to make these combining decisions “on-the-fly”.

Relational Database

An essential component to the improvement of the analysis workflow is the creation of a database that improves upon the flat table format currently being utilized when consolidating

data across the different sources. Even with a smaller collection of fields, handling thousands of data points is made easier by logical separations to normalize the database data. A major benefit we receive from moving away from a flat table format is being able to assign unique keys to reference specific surveys and subjects. Each source of HIV patient data has its own ID format. However, assigning an universal ID that is uniform across all data points helps us manipulate and maintain the data in an easier way.

Figure 2 shows an overview of the relational structure introduced to handle the HIV fields in question. This database structure provides us several benefits that the previous representation did not. Firstly, the separation of patients from their surveys allows for the situations where a subject participates in multiple surveys. Previously, this information would have to be represented as multiple rows and would create repetitive data. Fields that are unlikely to change can then be stored specific to the subject in question. These include fields recording a patient’s country of birth, education level, gender, sexual orientation, and ethnicity. Meanwhile, fields that were time-sensitive would be tied to the surveys. This change allows us to observe survey values changing as time passes for the same patient, something harder to see in the previous flat structure.

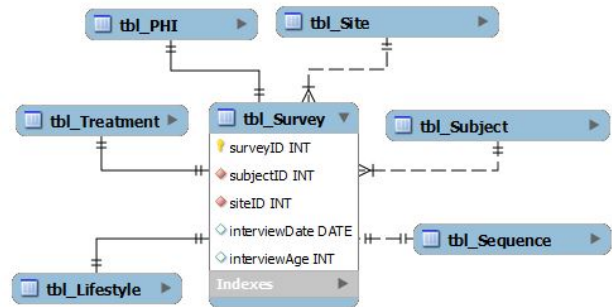


Figure 2. Relational Database Schema

In addition to that separation, we sort similar fields into categories such as lifestyle or treatment to improve visibility. Previously, viewing demographic data, for example, comes with many non-demographic fields that contain data the researcher is not interested in. The Lifestyle table contains data from questions related to recreational drugs usage, sharing needles, participating in transactional sex, and sexually transmitted disease diagnoses. On the other hand, fields in the Treatment table primarily deal with more clinical measurements and HIV specific treatments such as patient’s viral load, CD4 levels, and whether they are currently receiving antiretroviral therapy (ART).

Protected Health Information (PHI) data is another subcategory similar to the lifestyle and treatment division. PHI refers to private patient data that must be stored and managed to higher standards of security and confidentiality to protect the privacy of the patients participating in the study. This data is beyond the scope of this project; we primarily deal with fields where this information has been stripped out. Our main

infrastructure including our database already account for this data and can be easily expanded on in the future.

Another component of the avant-garde database is a place to record the sequences and HIV markers generated by medical researchers after the data has been processed. This sequencing data plays a large role in identifying the different strains of HIV and can aid in tracing each to geographic locations. This table does not record the entire sequence, but provides a mapping of the generated sequence id to the survey the strain was taken from.

In order to allow users to see the data in its original format and allow us to trace anomalies in the data, each source of data also has their own table in our server, separate from the database structure created to be the aggregation of the data. Before the HIV patient data is mapped and stored, we store its original form in the site-specific table.

Scheduled Script

One of the components to improving the workflow of HIV researchers at UCSD is giving access to more current data. The AVRC AEH data is exported into CSV (comma-separated values) files on their server. To keep our database current, a Python script runs nightly joins the CSV files containing the fields we are interested in using the Pandas library [6]. The aggregation of these CSV files forms one file that is archived on our server. By running the Linux “diff” command on the new file and the archived file from the night before, we can identify the new or modified entries that can then be added to the database.

Web Interface

After enhancing the data structure and creating a database to store and manage the data, we wanted to ensure easy access to our users. One of the primary components of this project was to reduce the logistical aspects of data analysis at the AVRC. The aspects we are concerned with tackling include reducing the need for researchers to manipulate raw data, remove the need for emailing to exchange data, and changing the linearity of the workflow. To create an application that would meet those needs, we developed a web interface using Xataface [3], an open source PHP framework. Primarily, we focus on three types of users who will utilize this system: administrators, medical site importers, and researchers.

Administrators of the interface have the ability to monitor the data as well as control user access with fine granularity. Administrators define roles that can control what actions a user can take as well as which tables he or she can see or manipulate. For example, the Amigo study will only have agency over their own data (importing, adding, deleting data), but will be unable to see or manipulate other medical sites. Additionally, administrators can monitor the incoming and outgoing patient data to deal with any potential issues. We record the history of any changes for data in the database. The history shows which user made the changes, when it was made, and which fields are changed. These changes can be easily reverted to a previous version. This feature can help administrators discover and handle any issues with data. These fea-

tures are based on existing features in Xataface that have been adapted and extended where necessary.

Medical site importers can now make use of an interface for uploading their latest data. Because the medical sites use IBM Statistical Package for the Social Sciences (SPSS) software [8] to record their survey data, their data is exported in a .sav file format. Importers can export their data as a .sav file and store it on their local machine. They can then directly import their file into our web interface, preview the data in order to visually confirm it’s correctly being imported, and store it. Data is then automatically imported into their site-specific table in our database.

We developed custom import filters to handle different file formats and this is extensible. We can continue to add new filters, handling each file format individually, as need arises. For .sav files, we define a filter that uses PSPP (an open source replacement of SPSS) [7], to convert the format to CSV. After inserting to the site-specific table, we map the values to conform with our codebook and insert it into our overall database.


Researchers using our infrastructure experience a variety of improvements. For example, data dictionaries can be pulled up to quickly view what the values in each table represents. Previously, we researchers needed to navigate to a medical site’s codebook file to find this information. Another benefit is available through the Sequence table. A MySQL view shows the HIV sequence IDs linked to their corresponding surveys and researchers can see quickly which surveys have no sequencing data yet and export these data points to a CSV file. When the ID is filled in, researchers can import that back into our database via the same process used by medical site importers.

A major component of our web interface for researchers use (Figure 3) is a view of the of the data in our database. Xataface provides an advanced search form that we customize to match values in our database. For example, a query is executed to populate a drop-down box with all the unique values of spoken languages found in our data. Researchers can search and narrow the data sets to what he or she wishes to look at. The advanced search form allows researchers to define values or ranges for every field represented in our database. For example, a researcher may query for patients from the El Cuete study that are males between the ages of 30 to 40 years old. Selecting dates come with a small calendar widget for choosing ranges based on time frame as well. The search form allows users to only see the data they are interested in, and these data sets can be exported to a CSV file or used in the data visualization.

Visualizing Data

While in the last section we looked at handling the data, we shift the focus to examine the second major component of our project: creating tools to visualize the HIV data set to facilitate an Exploratory Data Analysis approach. For the purpose of this project, we implemented a parallel coordinates graph for exploring HIV data. Parallel coordinates are a type of graph that displays multiple variables in parallel to visually represent connections between these dimensions. Alfred

Aggregate | Sequence | Amigo | ElCuete | HombreSeguro | MMS | Parejas | STAHR | STAHR II | BBL | AEH



Search Aggregate [Advanced Search](#)

New

Import

Data Dictionary

List Records

Export CSV | Visualize Selected | Visualize All

<input type="checkbox"/>	Survey ID	Site Subject ID	Site Name	Date of Interview	Baseline Age (in years)	Question Duration (in months)
<input type="checkbox"/>	1	A025	Amigo	2011-07-11 00:00:00	60	4
<input type="checkbox"/>	2	A053	Amigo	10/21/2011	36	4
<input type="checkbox"/>	3	A188	Amigo	08/02/2012	27	4
<input type="checkbox"/>	4	EC029	ElCuete	04/06/2011	34	6
<input type="checkbox"/>	5	EC043	ElCuete	04/08/2011	35	6
<input type="checkbox"/>	6	EC060	ElCuete	04/11/2011	33	6
<input type="checkbox"/>	7	EC092	ElCuete	04/13/2011	30	6
<input type="checkbox"/>	8	EC135	ElCuete	04/19/2011	39	6

Figure 3. Web interface built with Xataface, displaying a table of HIV patient data.

Inselberg, from Tel Aviv University, believes parallel coordinates “transforms multivariate relations into 2-D patterns” [4]. These patterns can be more apparent when displayed as a visual problem.

Once researchers select which data points they wish to look at, through using the advanced search form or selecting rows via check-boxes, the system populates a parallel coordinates graph to view this data. Figure 4 shows a graph displaying a collection of fields from our database. This is created using D3.js (Data-Driven Documents) [1], a JavaScript library that provides a framework for creating a variety of graphs to represent data.

Interactivity is necessary for maximizing the effectiveness of Exploratory Data Analysis methods. The parallel coordinates visualization provides several ways to interact with the HIV data set. Fields can be added or removed from the graph to only show researchers the dimensions of the data they wish to view at the moment. These dimensions can also be rearranged by dragging left or right to see connections that the default ordering of axis may have hidden. A table displaying the data being shown in the visualization allows researchers to easily trace the lines displayed to individual data points within our database. An additional benefit of the parallel coordinates graph is the utilization of a technique called “brushing” which Stephen G. Eick and Graham J. Willis describes as a user moving a rectangle across data such that any points in the path are “drawn in a manner which separates them from the rest of the data” [2]. Researchers can select the areas of each dimension they wish to view and show only the data points that are in the brushed areas. By combining these tools, researchers gain powerful control over which data points and fields they wish to view in the visualization. By displaying

the data in a new, visual way and allowing the user to interact with the data, we hope to facilitate EDA methods.

RESULTS

Database

The introduction of a relational database structure to replace the previous formatting offers numerous benefits. Dividing the fields into categories allows for easier reading and more complicated queries to be executed. We can also utilize MySQL views to tailor displays of data to a researcher’s needs. From an administrative standpoint, adding fields or tables will have minimal impact on the overall database.

Web Interface

A web interface to tie together the database, visualization, administrators, researchers, and medical sites is beneficial to all parties involved. We have more control in regulating access to patient data, we can allow medical sites to importing survey data at their discretion, and we have a system to automatically harmonize data across multiple sources.

Parallel Coordinates

A parallel coordinates graph is just one of a variety of visualizations we can implement to show HIV patient data in a new light. This format, in particular, allows researchers to view a variety of complex fields in parallel and interact with these fields by reordering or “brushing” the dimensions. Choosing specific data sets through the website and adding or removing axes allows researchers to tailor exactly what data they wish to look at easily.

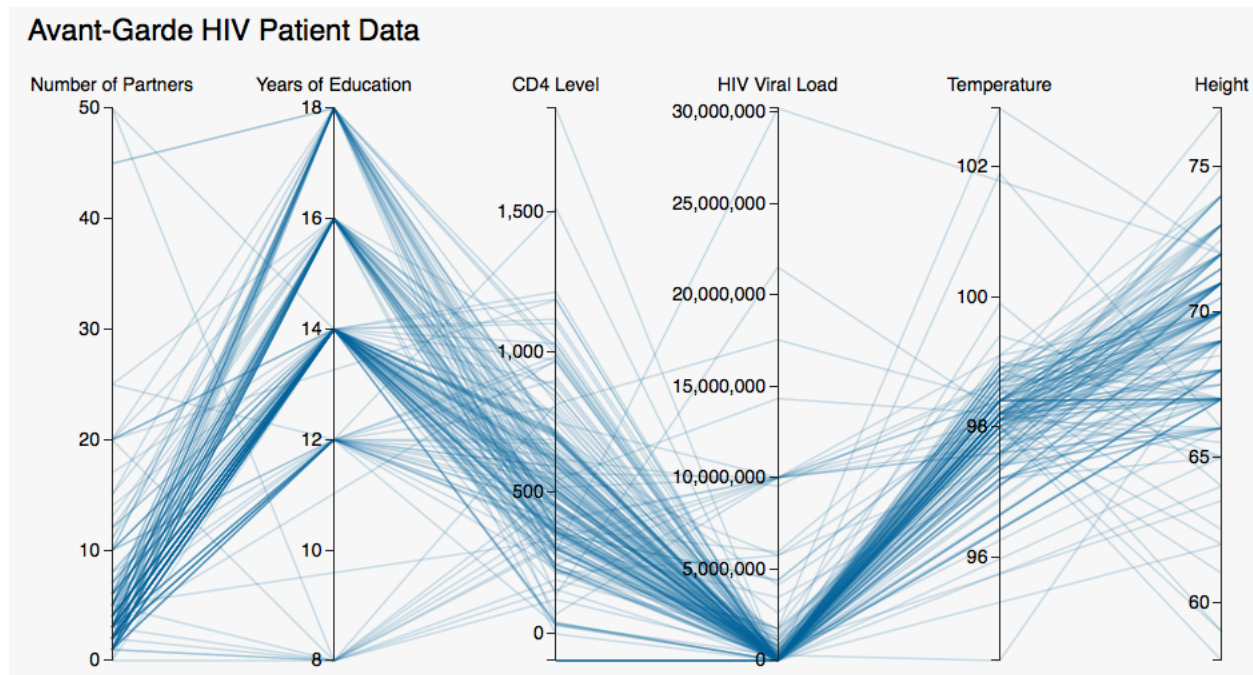


Figure 4. A parallel coordinates visualization displaying HIV patient data across multiple dimensions.

Workflow Comparison

The workflow concerns we hoped to address for this project were based on three elements: (a) handling of raw data, (b) linearity of the process, and (c) manual labor in harmonizing the data. Through our web interface, researchers, medical site importers, and administrators can now operate independently and no longer must be done in a serial manner. We have removed the need to send data via email by providing medical sites a place to import their data and automatically fetching AVRC AEH data from their servers. Data harmonizing is now a process done by our server, so researchers do not need to manipulate the data in its unconverted form the majority of the time. There remain certain situations where some manual decisions must be made to resolve harmonization processes we have decided upon.

Feedback

The web interface and parallel coordinates visualization was discussed with several HIV researchers and individuals who spend a large portion of their day dealing with fetching HIV data for researchers. These researchers are the ones we consider the primary users for the system, as well as the victims of the issues that we hoped to address in this project. The feedback for the web interface was overall positive and many additional improvements were suggested to be added. These include the ability to export data from the visualization directly, adding an area to view all recent changes to the HIV data, and a place in the database for notes on specific data.

Unfortunately, we were unable to demonstrate this project to the individuals working at the remote medical sites. We will require further feedback from this category of users to analyze the benefits of the system we have built and what unique challenges importers may face.

CONCLUSION

There were two primary motivations for the work in this project: improvement of the current workflow in which logistical issues hinder HIV research and an exploration of potential benefits Exploratory Data Analysis methods could provide. We have successfully created an interactive web-based tool that allows users to view, import into, manipulate, and visualize complex HIV data. We were able to aggregate data from multiple medical sites while maintaining the integrity of the data in question. The web interface, database, and visualization are extensible and will lead to further expansion of this project.

Future Work

The next step will be to put this project into production providing the medical sites and researchers regular access. More feedback as researchers use the interface will provide new directions for expansion, and more data being added will reveal new data considerations and obstacles that must be addressed.

Future work would primarily address expanding the number of data points that we considered within the scope of this project. The more fields we can represent in our database, the more likely previously unknown relationships between HIV transmission factors can be discovered. Primarily, data related to protected health information would be an interesting avenue of expansion. Access to the more sensitive data requires more care, but can be invaluable for analysis as well.

Another major direction for this project to expand on is in adding more visualizations. There are hundreds of additional ways to represent this complex data. The more visualizations made available to researchers, the more perspectives can be realized. In the same vein, expanding on the interactivity and

intuitiveness of the visualizations can give researchers more agency.

ACKNOWLEDGMENTS

I'd like to thank Dr. Sanjay Mehta for helping us narrow the scope of our project, choosing relevant data fields, and explaining in detail the current workflow process for HIV researchers. I'd like to thank Christy M. Anderson for providing explanations and access to the codebooks and paper forms for AVRC data. I'd especially like to thank my advisor, Professor Nadir Weibel for the continuous aid, advice, and feedback.

REFERENCES

1. Data-Driven Documents. <http://d3js.org>.
2. Eick, Stephen G., Wills, Graham J. High Interaction Graphics. *European Journal of Operations Research*, 81(3):445-459, March 1995.
3. Hannah, S. Xataface. <http://xataface.com/>.
4. Inselberg, A. Multidimensional Detective. <http://web.cs.ucdavis.edu/~ma/ECS289H/papers/Inselberg1997.pdf>.
5. NIDA Avant-Garde Program for HIV/AIDS and Drug Use Research. <http://www.drugabuse.gov/about-nida/organization/offices/office-nida-director-od/aids-research-program-arp/avant-garde-award-hivaids-research>.
6. Python Pandas Library. <http://pandas.pydata.org/>.
7. PSPP. <http://www.gnu.org/software/pspp/>.
8. IBM Statistical Package for the Social Sciences (SPSS). <http://www-01.ibm.com/software/analytics/spss/>.
9. Tukey, J. W. *Exploratory Data Analysis*. Addison-Wesley Series in Behavioral Science: Quantitative Methods, Reading, Mass 1, 1977.
10. UNAIDS Global Report 2013. <http://www.unaids.org/en/resources/campaigns/globalreport2013/factsheet/>.