University of California, San Francisco

Larry L. Sautter Award Submission

Epilepsy Phenome Genome Project –

Data Warehouse, Data Dictionary, Data Visualization & Search Engine

Date: April 15th, 2010

CONFIDENTIAL

Table of Contents

1. Project Title

Epilepsy Phenome Genome Project – Data Warehouse, Data Dictionary, Data Visualization and Search Engine at the University of California, San Francisco

2. Submitter’s Details

Gerry Nesbitt, MBA PMP
Director of Bioinformatics
University of California, San Francisco
Department of Neurology
Epilepsy Phenome Genome Project
Telephone: (412) 889 3295
Email: gnesbitt *at* epgp.org

3. Names of Project Leaders and team Members

Bioinformatics Project Team
Mr. Gerry Nesbitt, MBA PMP
Director of Bioinformatics, UCSF

Mr. Kevin Miller
Data Manager and Senior Developer, UCSF

Mr. Alan Carpenter
Senior Developer, UCSF

Ms Vickie Mays
EEG/MRI Data Coordinator, UCSF

Mr. Harry LeBlanc
Senior Data Architect (Contract), UCSF

Mr. Chris Pragash
Senior Programmer (Contract), UCSF

Project Sponsors
Dr. Daniel Lowenstein, M.D.
Professor of Neurology, Department of Neurology at UCSF, Director of the UCSF Epilepsy Center

Ruben Kuzniecky, M.D.
Professor of Neurology, Comprehensive Epilepsy Center, NYU Medical Center

4. Project Significance

The Epilepsy Phenome Genome Project (EPGP) is a multi-center NIH-funded study to create a comprehensive phenomic (clinical) and genetic database in epilepsy. The EPGP project is collecting large amounts of phenotypic, imaging and genomic data on thousands of study participants. Here we present the projects that were undertaken to design and build a centralized phenotypic data warehouse and supporting data dictionary, and describe the data visualization tools implemented for exploring data and presenting results. We have successfully developed, implemented and integrated multiple technologies, and have accelerated the flow of critical information to key stakeholders as a result.

The EPGP data warehouse was designed based on a simple conceptualization consisting of interconnected architectural layers, and currently contains over 1.5 million data points. It is populated every 6 hours from disparate transactional databases that contain clinical, electrophysiological, and neuroimaging data using custom-developed ETL tools. The EPGP data dictionary contains definitions of more than 2670 clinical data elements and is used by various end-user applications to explore, visualize and analyze the huge amounts of data in EPGP's data warehouse.

We anticipate that the combined EPGP data warehouse will help researchers to identify the genetic contributions that cause specific epilepsy syndromes and predict the therapeutic efficacy of anti-epileptic drugs (AEDs). And finally, the EPGP data warehouse will establish a resource that will be available to other researchers who will apply new analytical methods in the future that are impractical or unimagined today.

5. Project Description

5.1. What is EPGP?

The Epilepsy Phenome Genome Project (EPGP) is a multi-center NIH-funded study to create a comprehensive phenomic (clinical) and genetic database in epilepsy. EPGP will conduct groundbreaking research that will characterize the clinical, electrophysiological, and neuroimaging phenotypes of 3,750 patients with discrete subtypes of idiopathic-generalized, focal, or severe early-onset pharmacoresistant epilepsy.

5.2. EPGP's Contribution Epilepsy Research

The EPGP project will provide an excellent patient population to address the significance of genetic contributions to certain types of epilepsy. The rigorous collection of phenotypic data, drug response data and a large sample of patients will provide significant power to detect clinically meaningful associations between genetic polymorphisms and the epilepsy phenotype. Furthermore, the proposed whole-genome analysis will allow us to consider novel drug response genes and is likely to significantly enhance our understanding of the biology of anti-epileptic drug response.

In addition, the EPGP will establish a national resource that will be available to other researchers who will apply new analytical methods in the future that are impractical or unimagined today. The Epilepsy Phenome/Genome Project engenders the prospect of major advances in epilepsy research that will ultimately be of direct benefit to patients.

5.3. EPGP's Data Warehouse

Definition:
Data Warehouse -- A data warehouse is a repository of an organization's electronically stored data.

The EPGP project is collecting a large amount of phenotypic, imaging and genomic data on thousands of study participants. To discover new knowledge from these data, they need to be centralized in a data warehouse and made accessible to various end-user data reporting, exploration and visualization applications.

Because it was difficult to develop reports or do data analysis against EPGP's transactional databases due to their complexity, the EPGP data warehouse was designed to facilitate easier reporting and analysis. It was necessary to integrate heterogeneous data sources into a composite data repository that would facilitate easy data exploration and data visualization. Custom tools were developed to build and maintain the data warehouse, and off-the-shelf tools were procured to provide data exploration and data visualization capabilities.

EPGP’s transactional databases were highly normalized to ensure the data was free of data anomalies, thereby providing better data integrity. This made the underlying data architecture of the transactional databases very complex and would not be feasible for end-users to consider running data reporting or data visualization applications against them. Therefore, we needed to build a read-only data warehouse that was denormalized to simplify the data architecture, and would empower end-users to conduct their own business intelligence activities.

EPGP’s transactional databases contain all the phenotypic data collected on all study participants, and was gathered using web-based data collection instruments. These data are used to populate the data warehouse. We knew we had to decouple these transactional databases data from the data warehouse and decouple the data warehouse from the end-user business intelligence applications, so the architecture of the data warehouse was designed on a simple conceptualization consisting of the following interconnected layers:

Operational Database Layer
This layer is the source data for the data warehouse, and includes data stored in various transactional databases and file systems.

Data Access Layer
This layer is the interface between the operational and informational access layers. It consists of custom developed ETL tools to extract the data from the source databases and load it into the EPGP data warehouse.

Informational Access Layer
This layer consists of the data stored in the data warehouse and is accessed by the tools and applications that facilitate data reporting, data visualization, data mining, and analysis.

Metadata Layer
This layer is comprised of the EPGP data dictionary that describes all the data elements in the EPGP data warehouse, and is used by the end-user business intelligence applications, like the data reporting and data visualization applications.

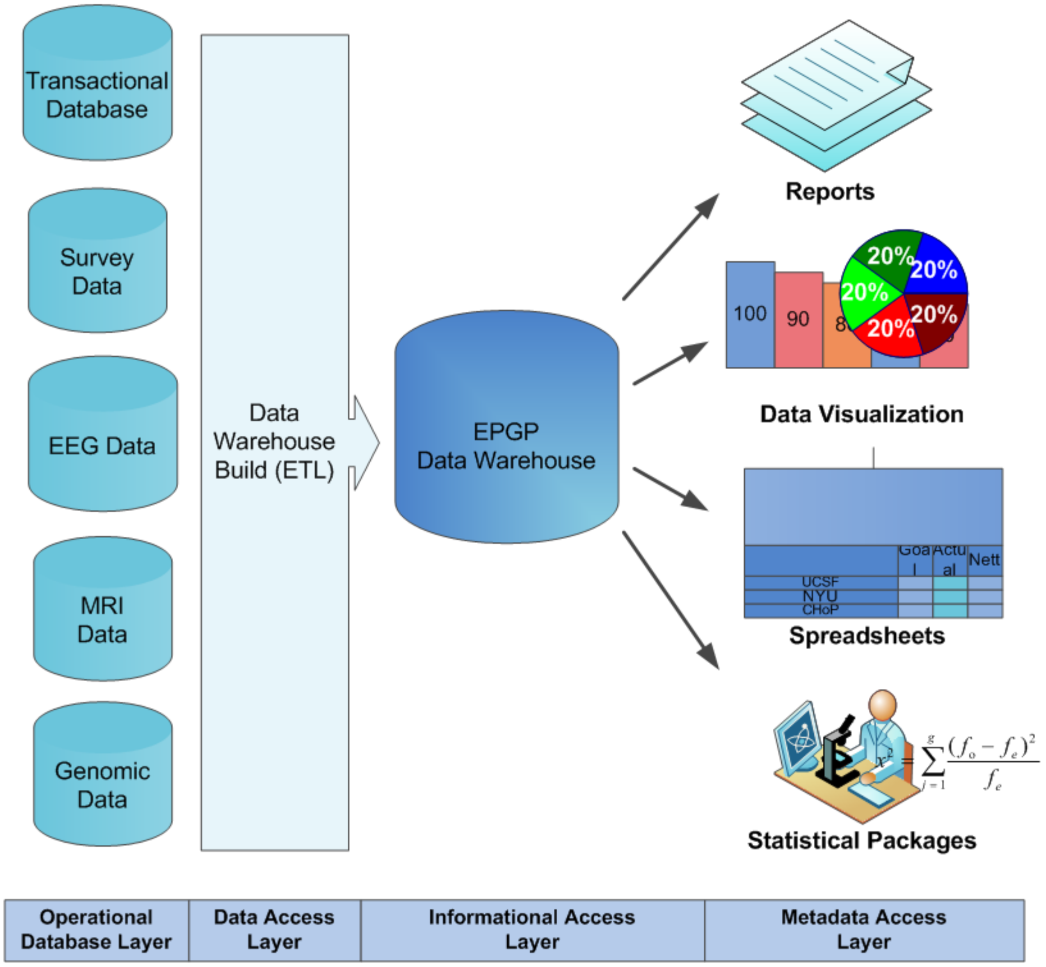


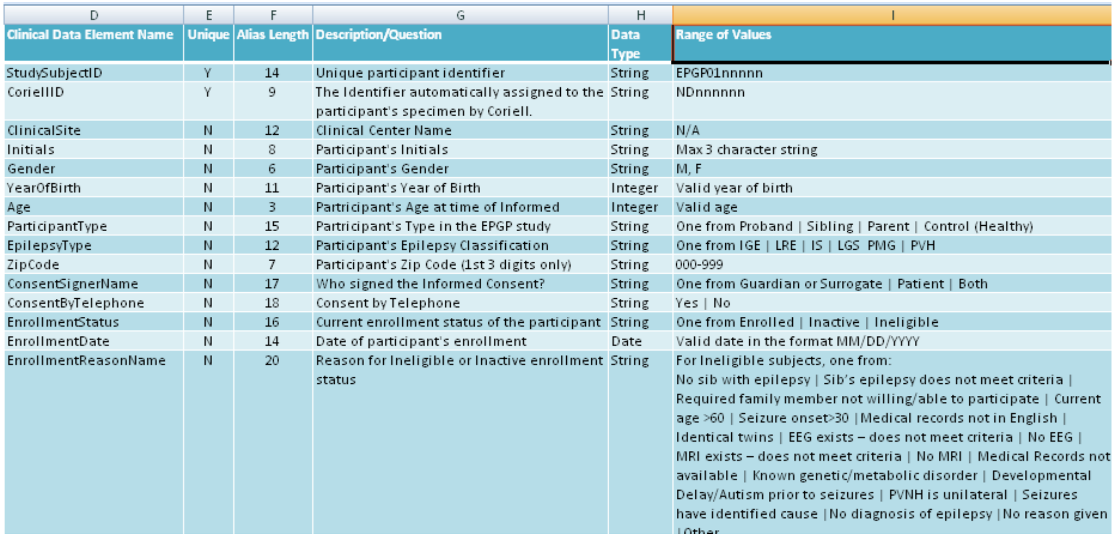
Figure 1 - High-level Data Warehouse Architecture

EPGP has designed and implemented a data warehouse that contains all of EPGP's phenotypic data. We have automated the ETL process to import clinical data from disparate transactional databases ever 6 hours so that the data stays relatively current. Over 20 reports, developed using MS SQL Server Reporting Server 2008®, are now using the data warehouse and execute a lot faster than if programmed against the transactional databases.

5.4. EPGP Data Dictionary

The EPGP data dictionary is a catalogue of all EPGP's clinical data elements. It consists of metadata that describes the underlying data in the data warehouse, and is a critical component due to the volume and diversity of data in the data warehouse, the many end-users who will access the data and the multiple data sources used to populate it. The data dictionary contains information about EPGP's data items, such as data point name, data type, length, description, origin, usage, format and encoding.

The EPGP data dictionary contains definitions of more than 2670 clinical data elements. We used a MS Excel spreadsheet to store the metadata on each clinical data element. The reason for this was because it was much simpler approach than say implementing the somewhat complex ISO/IEC 11179 Metadata Registry (data dictionary) standard, and the clinical data elements were still in a state of flux so the data dictionary needed to be extremely adaptable to constant and rapid change.



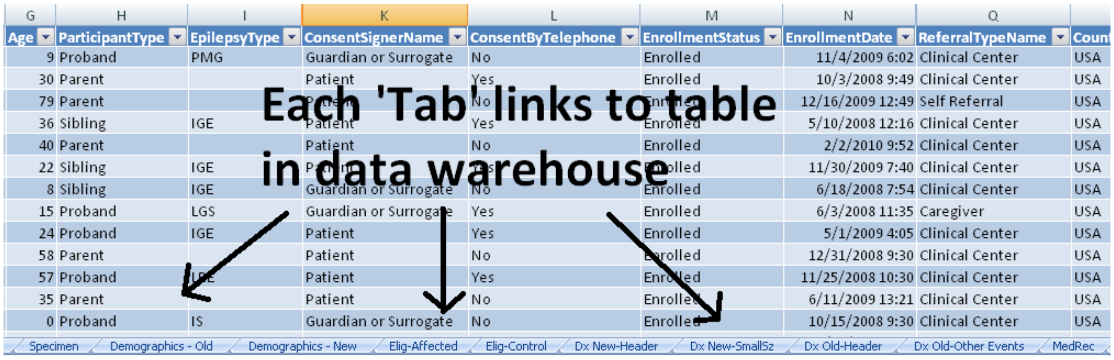
D	E	F	G	H	I
Clinical Data Element Name	Unique	Alias Length	Description/Question	Data Type	Range of Values
StudySubjectID	Y	14	Unique participant identifier	String	EPGP01nnnnn
CorielIID	Y	9	The Identifier automatically assigned to the participant's specimen by Coriell.	String	NDnnnnnn
ClinicalSite	N	12	Clinical Center Name	String	N/A
Initials	N	8	Participant's Initials	String	Max 3 character string
Gender	N	6	Participant's Gender	String	M, F
YearOfBirth	N	11	Participant's Year of Birth	Integer	Valid year of birth
Age	N	3	Partricipant's Age at time of Informed	Integer	Valid age
ParticipantType	N	15	Partricipant's Type in the EPGP study	String	One from Proband | Sibling | Parent | Control (Healthy)
EpilepsyType	N	12	Participant's Epilepsy Classification	String	One from IGE | LRE | IS | LGS PMG | PVH
ZipCode	N	7	Participant's Zip Code (1st 3 digits only)	String	000-999
ConsentSignerName	N	17	Who signed the Informed Consent?	String	One from Guardian or Surrogate | Patient | Both
ConsentByTelephone	N	18	Consent by Telephone	String	Yes | No
EnrollmentStatus	N	16	Current enrollment status of the participant	String	One from Enrolled | Inactive | Ineligible
EnrollmentDate	N	14	Date of participant's enrollment	Date	Valid date in the format MM/DD/YYYY
EnrollmentReasonName	N	20	Reason for Ineligible or Inactive enrollment status	String	For Ineligible subjects, one from: No sib with epilepsy | Sib's epilepsy does not meet criteria | Required family member not willing/able to participate | Current age >60 | Seizure onset>30 |Medical records not in English | Identical twins | EEG exists – does not meet criteria | No EEG | MRI exists – does not meet criteria | No MRI | Medical Records not available | Known genetic/metabolic disorder | Developmental Delay/Autism prior to seizures | PVNH is unilateral | Seizures have identified cause |No diagnosis of epilepsy |No reason given |Other

Figure 2 - Data Dictionary

To date, EPGP has collected 1,551,408 phenotypic data points on over 1700 study participants, encompassing clinical, electrophysiological, and neuroimaging data.

5.5. EPGP Data Visualization

The EPGP study is gathering a vast amount of research data, which could provide novel insights into epilepsy syndromes when explored. Data visualization provides an excellent approach for exploring data and presenting results using meaningful charts, and therefore plays a crucial role in trying to understanding data. The goal of data visualization is to communicate information clearly and effectively through graphical means, and facilitate data reporting, data mining and data analysis. It enables data stored in EPGP's large phenotypic datasets to be condensed into meaningful visual representations and facilitates visual comparisons of data. Researchers wishing to explore or visualize the data are not restricted to a specific package, i.e. they can use commonly available applications like MS Excel®, MS Access® and SAS® to explore the data, and other 3rd party software tools to visualize the data, such as Omniscope™.



G	H	I	K	L	M	N	Q	
Age	ParticipantType	EpilepsyType	ConsentSignerName	ConsentByTelephone	EnrollmentStatus	EnrollmentDate	ReferralTypeName	Count
9	Proband	PMG	Guardian or Surrogate	No	Enrolled	11/4/2009 6:02	Clinical Center	USA
30	Parent		Patient	Yes	Enrolled	10/3/2008 9:49	Clinical Center	USA
79	Parent		Patient	No	Enrolled	12/16/2009 12:49	Self Referral	USA
36	Sibling	IGE	Patient	Yes	Enrolled	5/10/2008 12:16	Clinical Center	USA
40	Parent		Patient	No	Enrolled	2/2/2010 9:52	Clinical Center	USA
22	Sibling	IGE	Patient	Yes	Enrolled	11/30/2009 7:40	Clinical Center	USA
8	Sibling	IGE	Guardian or Surrogate	No	Enrolled	6/18/2008 7:54	Clinical Center	USA
15	Proband	LGS	Guardian or Surrogate	Yes	Enrolled	6/3/2008 11:35	Caregiver	USA
24	Proband	IGE	Patient	Yes	Enrolled	5/1/2009 4:05	Clinical Center	USA
58	Parent		Patient	No	Enrolled	12/31/2008 9:30	Clinical Center	USA
57	Proband	IGE	Patient	Yes	Enrolled	11/25/2008 10:30	Clinical Center	USA
35	Parent		Patient	No	Enrolled	6/11/2009 13:21	Clinical Center	USA
0	Proband	IS	Guardian or Surrogate	No	Enrolled	10/15/2008 9:30	Clinical Center	USA

Each 'Tab' links to table in data warehouse

Figure 3 - Viewing Data in the Data Warehouse
End-users can use a MS Excel spreadsheet to view the data in the data warehouse, using each 'tab' to point to a specific database table.

There are seven stages to visualizing data, which are

1. Acquire
2. Parse
3. Filter
4. Mine
5. Represent
6. Refine
7. Interact.

EPGP procured a data visualization and charting tool, called Omniscope™, because it allowed the end-user to control each of these steps in the data visualization process, and perform data visualization in a progressive and iterative manner. This tool also suited the intuitive design of EPGP's data warehouse, and was easy to implement and use from an end-users perspective.

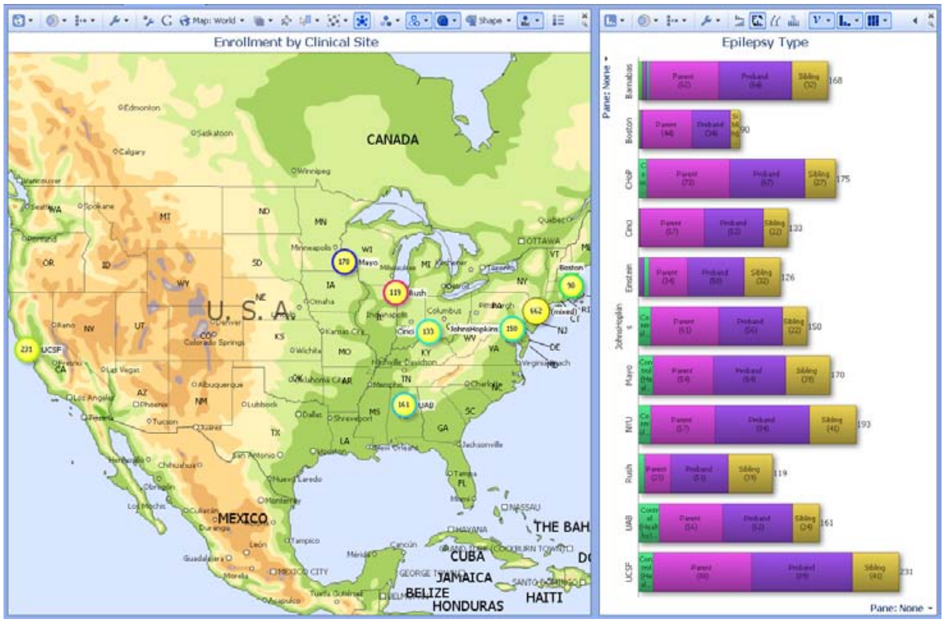
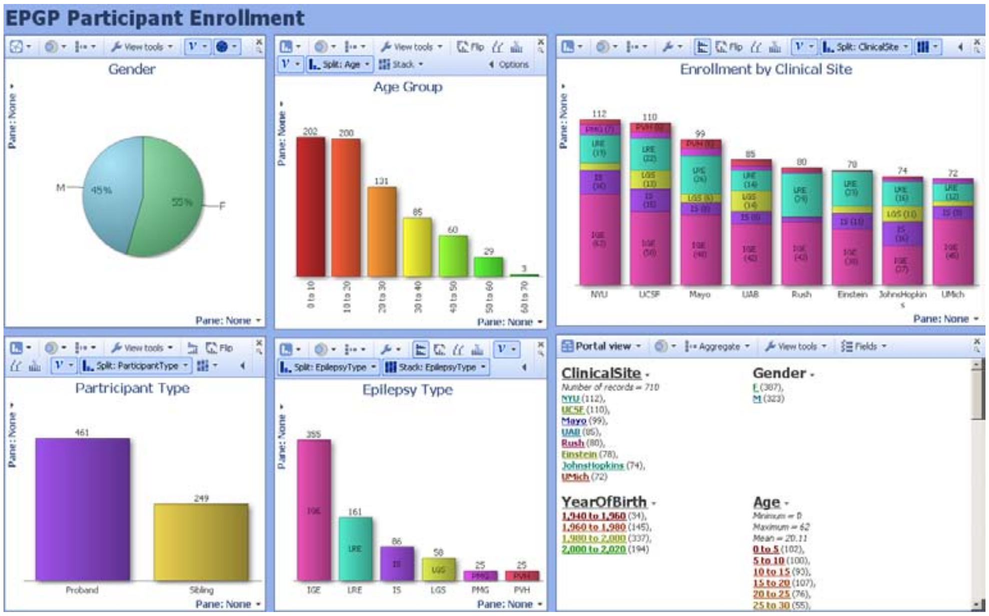


Figure 4 - Data Presented by Geographical Location



ClinicalSite
Number of records = 710
NYU (112),
UCSF (110),
Mayo (99),
UAB (85),
Rush (80),
Einstein (78),
JohnsHopkins (74),
UMich (72)

Gender
F (387),
M (323)

YearOfBirth
1,940 to 1,960 (34),
1,960 to 1,980 (145),
1,980 to 2,000 (337),
2,000 to 2,020 (194)

Age
Minimum = 0
Maximum = 62
Mean = 20.11
0 to 5 (102),
5 to 10 (100),
10 to 15 (93),
15 to 20 (107),
20 to 25 (76),
25 to 30 (55),

Figure 5 - Various Data Visualization Charts

A number of measures are in place to control access to the applications and the data, ensuring compliance with standards like HIPAA and 21CFR11. No patient-identifying information is stored in the databases. All end-users are assigned a unique username and password, and access to the applications depends on the permissions assigned to them. The databases have role-based security implemented and all web-based applications are run over HTTPS using SSL 128-bit encryption, ensuring that all communications between the browser and web application are encrypted.

End-users can exploit a variety of business intelligence tools to explore and view the data in the data warehouse. EPGP implemented the Omniscope™ tool to empower the end-users with data visualization capabilities. It is used by study management personnel and researchers to report on and analyze the data, and requires minimal effort to implement the software and train end-users. These tools have enabled study management personnel and researchers to explore the ever-expanding volume of phenotypic data that are stored in easy to interpret formats.

5.6. *EPGP ‘s ‘google-like’ Search Engine*

In an effort to make EPGP’s phenotypic data available to as wide an audience as possible, we implemented a ‘google-like’ search engine called Tabula DX, which enables searching across large collections of PDFs. We developed a program that creates PDFs for all the phenotypic data (surveys, EEG data and MRI data) in the data warehouse. This program is run automatically each night to create PDFs for the data collected since the last run. Once these PDFs are created, the search engine indexes the contents of each new PDF and makes the document available via the search website.

To run a search, the end-user enters the search string in the field and clicks [Search]. The search results returned include a thumbnail of the PDF document, the document title, a section of the test in the PDF that contains the search string, and a link to open the PDF file with ‘Find’.

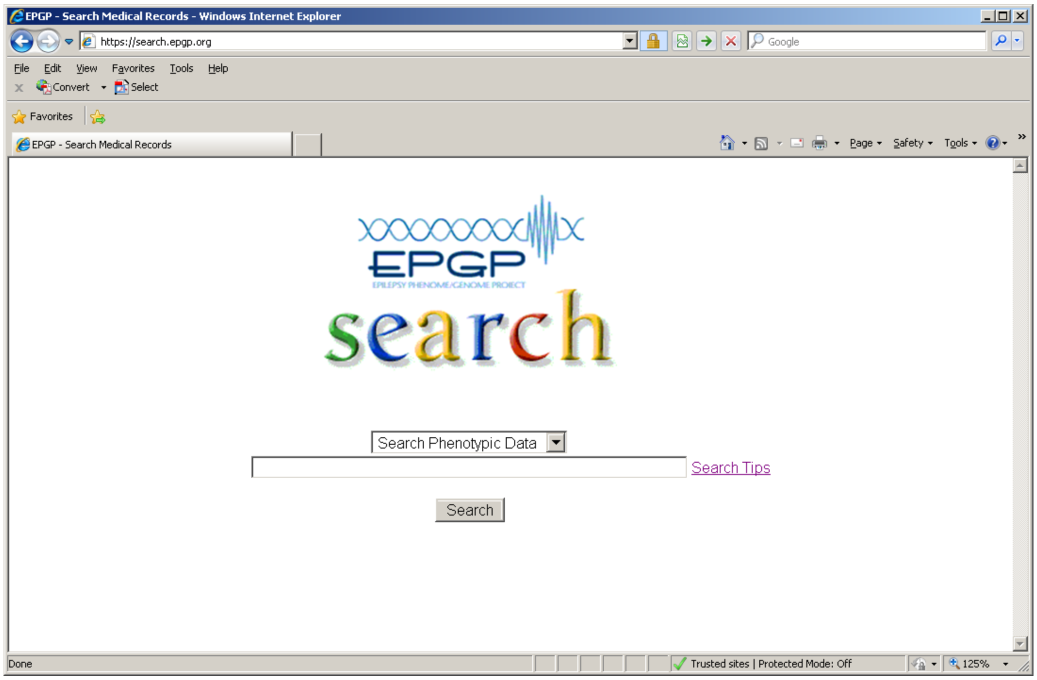
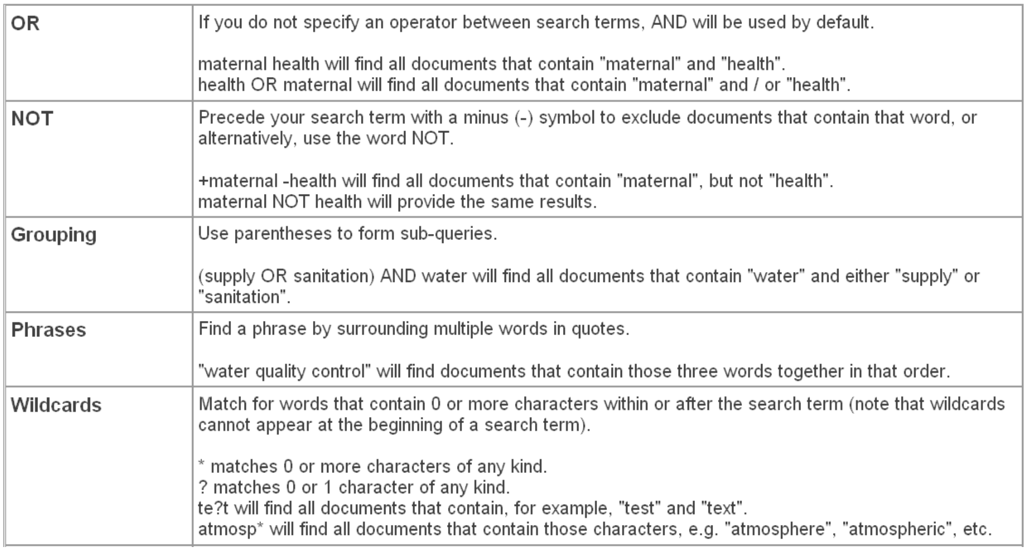

EPGP - Search Medical Records - Windows Internet Explorer
https://search.epgp.org
Google
File Edit View Favorites Tools Help
Convert Select
Favorites
EPGP - Search Medical Records
Page Safety Tools
EPGP
EPILEPSY PHENOME/GENOME PROJECT
search
Search Phenotypic Data
Search Tips
Search
Done
Trusted sites | Protected Mode: Off
125%

Figure 6 – EPGP’s Search Engine Website

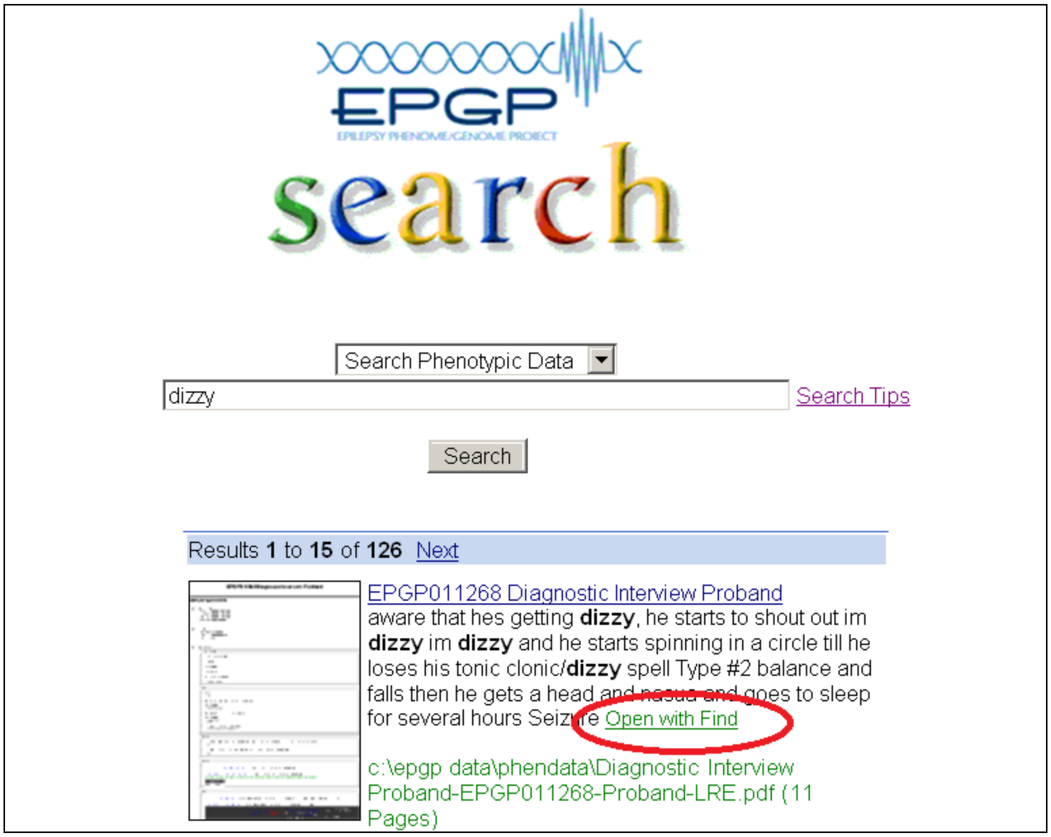
The search website also includes options for filtering the data. Examples of some search strings include:

Search String	**Explanation**
EPGP011100	Retrieves survey responses, EEG data and MRI data for the subject EPGP011100.
EPGP011100 AED	Retrieves the AED Data Sheet for EPGP011100.
EPGP011100 Demographics	Retrieves the Subject Demographics & Ethnicity survey response for EPGP011100.
Baseball	Retrieves all survey responses that contain the word “Baseball”
Baseball Laser	Retrieves all survey responses that contain the words “Baseball” AND “Laser”
Baseball OR Laser	Retrieves all survey responses that contain the word “Baseball” OR “Laser”
Baseball NOT Laser	Retrieves all survey responses that contain the word “Baseball” but not “Laser”
"*Baseball bat*"	Retrieves all survey responses that contain the text string “Baseball bat”
keywords:IGE	Retrieves all survey responses that contain the keyword “IGE”
title:Final	Retrieves all survey responses that contain the word “Final” in the title



OR	If you do not specify an operator between search terms, AND will be used by default. maternal health will find all documents that contain "maternal" and "health". health OR maternal will find all documents that contain "maternal" and / or "health".
NOT	Precede your search term with a minus (-) symbol to exclude documents that contain that word, or alternatively, use the word NOT. +maternal -health will find all documents that contain "maternal", but not "health". maternal NOT health will provide the same results.
Grouping	Use parentheses to form sub-queries. (supply OR sanitation) AND water will find all documents that contain "water" and either "supply" or "sanitation".
Phrases	Find a phrase by surrounding multiple words in quotes. "water quality control" will find documents that contain those three words together in that order.
Wildcards	Match for words that contain 0 or more characters within or after the search term (note that wildcards cannot appear at the beginning of a search term). * matches 0 or more characters of any kind. ? matches 0 or 1 character of any kind. te?t will find all documents that contain, for example, "test" and "text". atmosp* will find all documents that contain those characters, e.g. "atmosphere", "atmospheric", etc.

Figure 7 – Search Filters



dizzy

Results **1** to **15** of **126** Next

EPGP011268 Diagnostic Interview Proband
aware that hes getting **dizzy**, he starts to shout out im **dizzy** im **dizzy** and he starts spinning in a circle till he loses his tonic clonic/**dizzy** spell Type #2 balance and falls then he gets a head and nausea and goes to sleep for several hours Seizure Open with Find

c:\epgp data\phendata\Diagnostic Interview Proband-EPGP011268-Proband-LRE.pdf (11 Pages)

Figure 8 – Open Document with Find

5.7. Conclusion

EPGP's data warehouse and suite of data visualization tools have been extremely successful, not because of its architecture or its structure, but rather on its ability to generate ideas that help to build, maintain and enhance the use the data warehouse. The EPGP data warehouse is now accessible to all EPGP researchers using a broad range of reporting and data visualization tools, and the number of end-users continues to grow. The ETL informatics tools were custom developed for EPGP but the architecture provides a feasible model for widespread use on other clinical studies throughout UCSF.

We anticipate that the combined EPGP data warehouse will help researchers to identify the genetic contributions that cause specific epilepsy syndromes and predict the therapeutic efficacy of AEDs. And finally, the EPGP data warehouse will establish a resource that will be available to other researchers who will apply new analytical methods in the future that are impractical or unimagined today.

5.8. Technical Architecture

O/S: Windows Server 2003 R2
Database: MS SQL Server 2008
Reporting: MS SQL Server Reporting Server 2008
Development: MS Visual Studio, Visual C#, SQL Stored Procedures
Data Visualization: Visokio - Omniscope
Search Engine: AquaForest - Tabula DX

5.9. Abbreviations Used

EPGP:	Epilepsy Phenome Genome Project
ETL:	Extract, Transform and Load
AED:	Anti-epileptic Drug
HIPAA:	Health Insurance Portability and Accountability Act (HIPAA) of 1996
ISO/IEC 11179:	ISO/International Electrotechnical Commission (IEC) 11179 Metadata Registry (data dictionary) Standard

Feedback from Stakeholders

Dr. Daniel Lowenstein, M.D.
Professor of Neurology, Department of Neurology at UCSF, Director of the UCSF Epilepsy Center

"The EPGP data warehouse is a superb resource that will be mined and explored by researchers trying to discover the genetic contributions that cause specific epilepsy syndromes and to identify better epilepsy drug therapies. The EPGP bioinformatics team at UCSF has done a phenomenal job implementing the data warehouse, data dictionary, data visualization tools and search engine for such a large-scale research study in less than six months. There is no question that the work of the bioinformatics team has been a critical factor in the success of the project to date."