AC 2009-369: COMPUTATIONAL DATA MINING FOR FEATURE EXTRACTION IN HEALTH INFORMATICS

Mahmoud Quweider, University of Texas, Brownsville

Dr. M K Quweider is an Associate Professor in the Computer & Information Sciences at the University of Texas at Brownsville/Texas Southmost College. He received his Ph.D. in Engineering Science and an M.S. in Applied Mathematics, M.S. in Engineering Science, and M.S. in Biomedical Engineering all from the University of Toledo, Ohio. After graduation, he worked at several places including Pixera, a digital multimedia processing company in Cupertino, CA, and 3COM, a networking and communication company in Schaumberg, IL. He joined the UTB/TSC in 2000. His areas of interest include Imaging, Visualization and Animation, Web Design and Graphics.

Adriana Perez, University of Texas, Brownsville

Gabriala Oropeza, University of Texas, Brownsville

Juan Iglesias, University of Texas, Brownsville

Dr. J R Iglesias is the Chair and Associate professor in the Computer & Information Sciences at University of Texas at Brownsville/Texas Southmost College. He received his Ph.D. in Computer Science from New Mexico State University (NMSU), New Mexico, USA, with specialization in Databases, and the B.Sc and M.S. in Computer Science from the National Autonomous University of Mexico. He has worked as an Associate Director for the Federal Electoral Institute (IFE), Mexico City, Mexico during the 1997 year. His areas of interest include Databases, Programming Languages, Data mining, Web Design, and e-Commerce Systems.

Computational Data Mining for Feature Extraction in Health Informatics

Abstract

This paper presents the methodologies and lessons learned from a cooperative effort within our institution involving the Health Science Center faculty, the Computer Science faculty, and senior/graduate students; the effort aimed at building a data mining module to input, process, and extract relevant information related to a pilot study on the effects of to Endocrine Disrupting Chemicals (EDCs) exposure on pregnancy, which was conducted by the Health Center and the School of Public Health.

Interdisciplinary in nature, the project brought together biostatisticians, medical doctors, and computer and information scientists (CIS). On the medical side, the team was trying to assess human health risks from exposures to Endocrine Disrupting Chemicals, measuring both the exposure level and its ramifications in pregnant women of the Rio Grande Valley. To aid in the process from a computational and engineering point of view, a professor and two computer science and engineering majors were put in charge of taking the requirements and specifications from the medical side and converting them into a robust and flexible software application with a friendly graphical user interface.

The study has progressed in a sequence of phases that included: obtaining approval of human subject research; preparing the necessary paperwork; recruiting subjects at local clinics, collecting blood and tissue samples, performing blood and tissue samples analysis, coding and entering data, constructing an integrated data base, performing statistical analysis, assessing human risk, and mining the database for trends, anomalies, and unusual cases.

The educational experience and the interaction between the students and the medical/health team were invaluable. The CIS students, and their professors, benefited immensely from not only coding the design and requirements but also from learning about concepts such as getting a certificate of training on research on human subjects, conducting and inscribing surveys, extracting and visualizing basic factors and trends from the collected data.

Our paper details the students' academic and professional experience in working with a real-life project with profound health and social impact on their local community. It also lays the foundation for continuous collaboration involving faculty and students between the involved departments.

Introduction

Computer Science is an applied science by its nature. Its applications are seen everywhere such as in the Internet, communications, e-commerce business to business and business to customer systems, electronics, and medical devices just to name few. This wide-spread range of applications brings a major challenge to computer science: the need to collaborate with other disciplines to bring about software that is of benefit to all stakeholders and users. This sentiment has been echoed by the leaders of the industry including Microsoft, the NSF, and the ACM society [1-5].

One of the areas that find computer science necessary for its advancement is health care services. Computer systems in this area have been successfully used to help clinicians gather and process data and then provide better patient care management. The University of Texas Health Science Center at Houston and the University of Texas School of Public Health, Brownsville Regional Campus, are actively engaged in education and research that addresses the need of the community especially in the Rio Grande Valley. Towards the goal of finding the effects of EDC on pregnancy, and in one of their research studies, the biostatistics division wanted to estimate the human health risks from additional exposures to Endocrine Disrupting Chemicals (EDCs), and see who if there are any symptoms connected to this exposure. To make more accurate estimates of human health risks from exposure to EDCs, the biostatistics division wanted to measure the exposure level and its ramifications in pregnant women of the Rio Grande Valley. The study also aimed at finding if race, age, or dietary habits are of relevance for diagnostic purposes.

Rather than running separate subsystems of software to collect, measure, and analyze the results, Dr. Perez, the PI from University of Texas School of Public Health, sought collaboration with the Computer Science department to conduct the study in a way that capitalizes on the power of computers as well as reducing human errors. Noting the recent convergence of Internet technologies and their deployment on different heterogeneous devices, Dr. Quweider, from the Computer Science, suggested building a web-based software platform for data-mining that caters to the needs of its medical-oriented users. The software would be web-based to allow access from any internet-ready computing device with a browsing capability. It would use MySQL and the platform-independent C# language. Details of transforming the project into an end product are given below.

Software Tools

The Pilot study application was developed using the spiral methodology for software development and two of the most popular technologies. The spiral methodology was used because it allows overlap among requirement analysis, design, analysis, coding, testing, and maintenance. .NET and MySQL. We have used .NET because of its platform independence and ease of use compared to other languages (such as C/C++) . Using .Net, the same language, C#, was used for programming, scripting, and designing the graphical user interface. MySQL is very easy to use and is a great choice for small size databases such as ours. The technologies used in the project are as follows:

DBMS: MySQL Server 5.0: MySQL Database System, version 5.0

SQLyog 5.25 : SQLyog provides you with powerful means to manage your MySQL databases.

IDE: Microsoft Visual Studio.NET.

IIS: Microsoft .NET Framework Version.

ASP.NET

Programming Language: C#

Browser: Internet Explorer version 7.0.

Team Formation

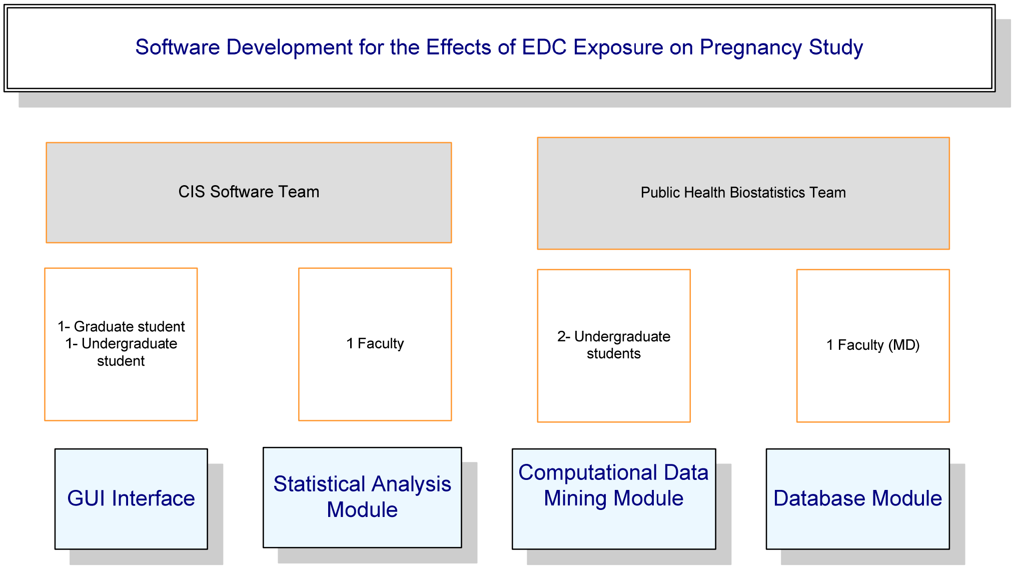

Software Development for the Effects of EDC Exposure on Pregnancy Study
CIS Software Team
Public Health Biostatistics Team
1- Graduate student
1- Undergraduate student
1 Faculty
2- Undergraduate students
1 Faculty (MD)
GUI Interface
Statistical Analysis Module
Computational Data Mining Module
Database Module

Figure 1. General View of Teams/Tasks

As an interdisciplinary project, a great amount of time was spent in formulating a set of specifications from which computer scientists were able to work and create program modules. The team consisted of one graduate student and one undergraduate student from the CIS department, the *Software Team*, and two students from the school of Public Health, the *Public Health* Team; the two teams were mentored and co-advised by Dr. Quweider and Dr. Perez. The Public Health team started by giving the software team a presentation about their study. This was followed by presenting the *Pregnancy Pilot Study* survey which is given in the appendix. The team went over all the fields in the survey and how are they measured and interpreted. They also detailed the current software tools available to them including the SAS statistical package and the FoxPro database package.

One of the findings revealed in this project was the importance of student-faculty and student-student interaction both electronically and face to face when needed. Arising issues were solved by assigning responsibilities to team members and motivating them to learn new material. The student-student interactions helped build teamwork, and communication skills and higher-level thinking skills.

During the whole software development process the two teams met on a regular basis to iron out any issues or detail and solve open questions. The Public Health team was kept abreast of every development from the specification to the design and implementation to the graphical user interface and operation and maintenance.

Software Development

After the two collaborating teams formed the specification for the project, the work on the design and implementation phases of the projects started. Use Case tools were used to define the role of users as shown in figure 2 and outlined in [6]. The figure shows in a compressed format the roles that an administrator and regular users will play. Security and authentication functionalities are given to the administrator in addition to the regular role. Users are uniquely identified by their login, and are also given certain privileges.

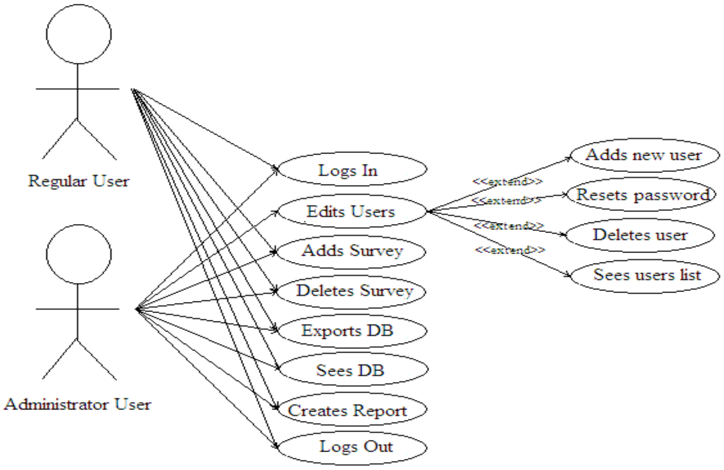


Figure 2. Use Case for Administrator/Users

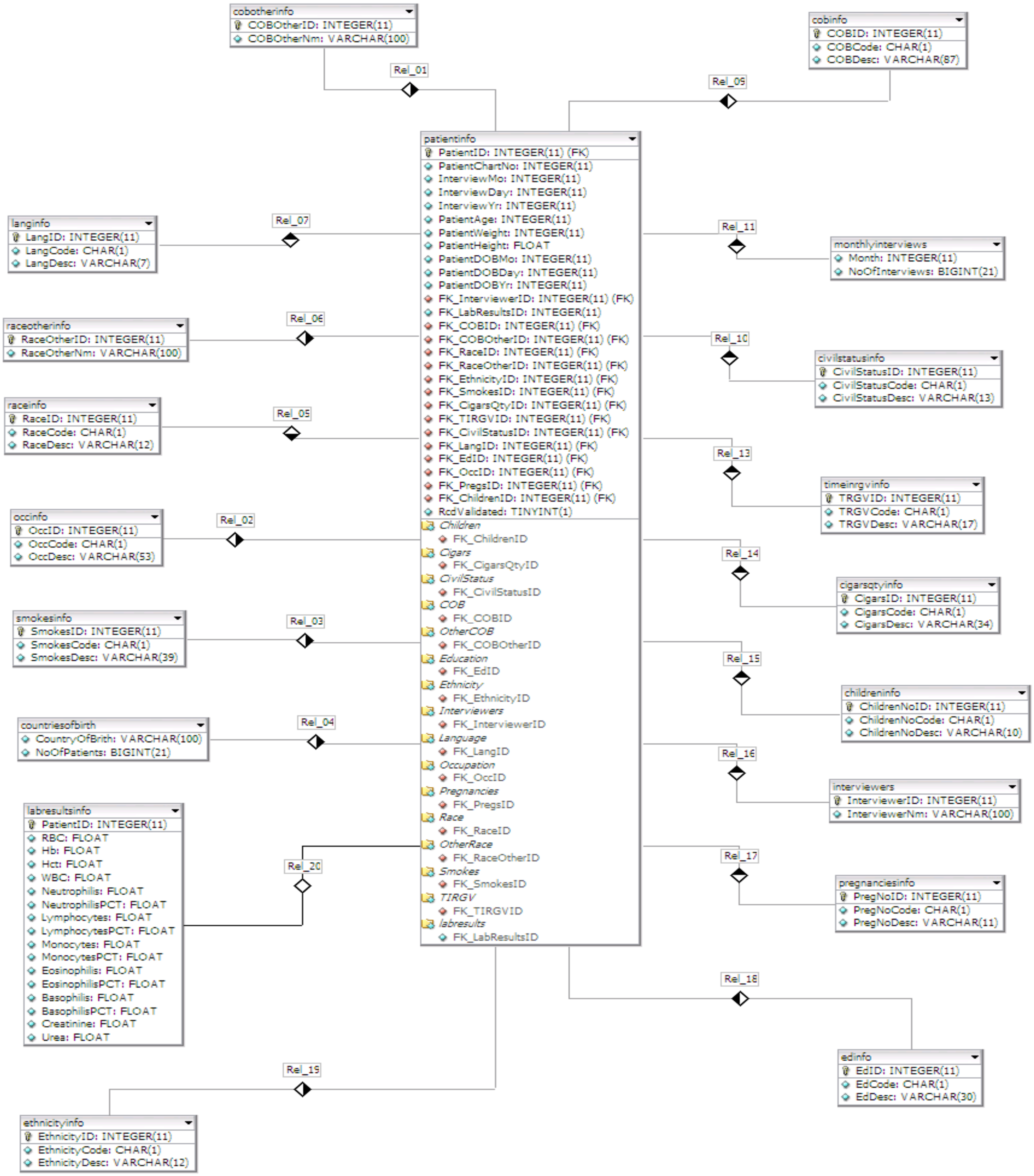
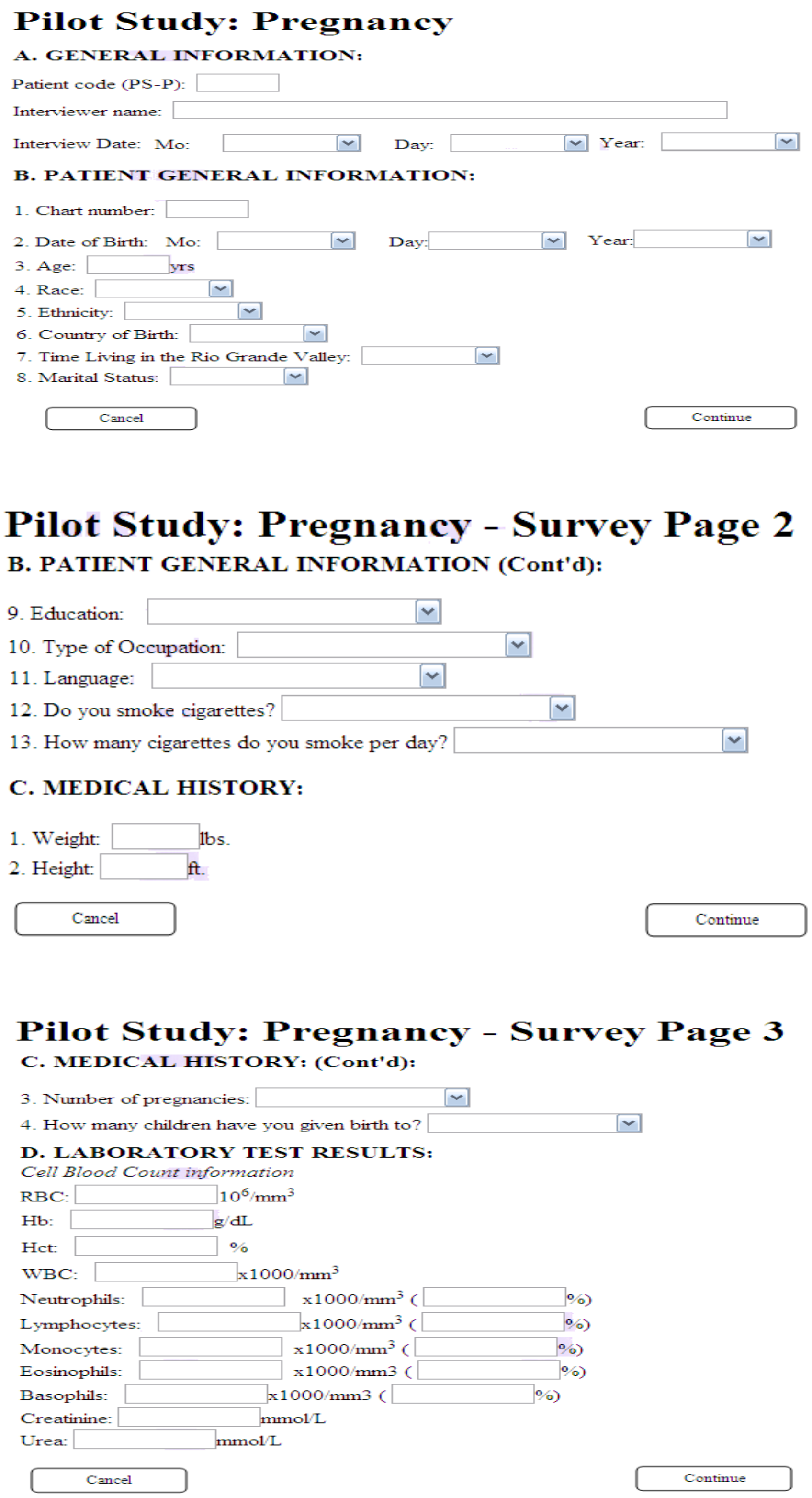


Figure 3. UML Class Diagram

We used the Unified Modeling Language (UML) as shown in figure 3 for our design of the project. As a graphical object oriented tool for design, UML allows programmers to modularize functionality of the system as well as interact, test, and maintain the modules in an independent way. Additionally, UML presents the semantic elements in a way that is easily grasped and manipulated by a developer.



Pilot Study: Pregnancy

A. GENERAL INFORMATION:

Patient code (PS-P):

Interviewer name:

Interview Date: Mo: Day: Year:

B. PATIENT GENERAL INFORMATION:

1. Chart number:
2. Date of Birth: Mo: Day: Year:
3. Age: yrs
4. Race:
5. Ethnicity:
6. Country of Birth:
7. Time Living in the Rio Grande Valley:
8. Marital Status:

Pilot Study: Pregnancy - Survey Page 2

B. PATIENT GENERAL INFORMATION (Cont'd):

9. Education:
10. Type of Occupation:
11. Language:
12. Do you smoke cigarettes?
13. How many cigarettes do you smoke per day?

C. MEDICAL HISTORY:

1. Weight: lbs.
2. Height: ft.

Pilot Study: Pregnancy - Survey Page 3

C. MEDICAL HISTORY: (Cont'd):

3. Number of pregnancies:
4. How many children have you given birth to?

D. LABORATORY TEST RESULTS:

Cell Blood Count information

RBC: 10^6/mm^3
Hb: g/dL
Hct: %
WBC: x1000/mm^3
Neutrophils: x1000/mm^3 (%)
Lymphocytes: x1000/mm^3 (%)
Monocytes: x1000/mm^3 (%)
Eosinophils: x1000/mm3 (%)
Basophils: x1000/mm3 (%)
Creatinine: mmol/L
Urea: mmol/L

Figure 4. GUI Diagrams

The Graphical user interface elements (GUI), shown in figure 4, were designed using C#. By working with the public health team, GUI diagrams were created in a format that is familiar to the users. Actual values stored in the databases were different, mainly to allow for normalization and easier mining in the future, than those presented in an input or output. The data-mining engine running on the back-end was responsible for these conversions (this is of course in addition to all data extraction, manipulation, and visualization).

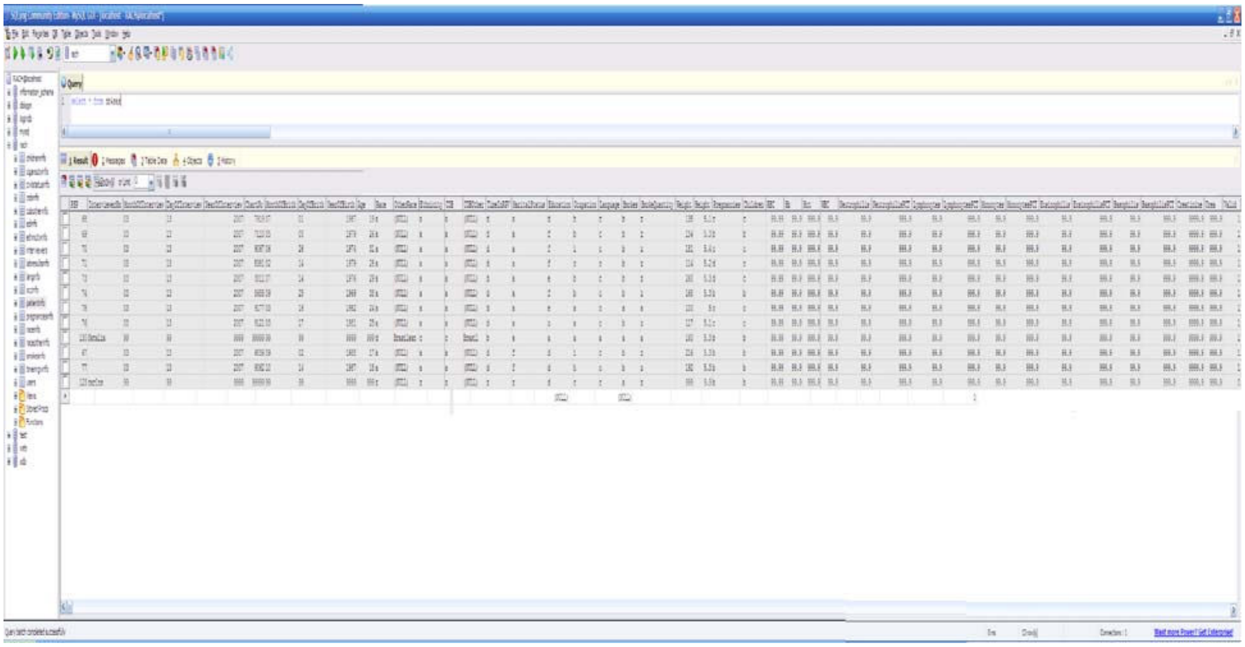


Figure 5: DB4SAS MySQL View

MYSQL was used to create the database along with a set of queries that were inspired by interaction with the public health team. While the tables were stored to conform to good relational database design, several queries were designed to join or merge or present the data in a way familiar to the public health workers. This entailed the creation of several interface layers before actual output display.

The software team also implemented many data mining functionality in which the Public Health team was interested in. These functionalities range from basic statistical analysis to finding peculiar patterns, to clustering data based on specific symptoms, to detecting anomalies.

Operation & Maintenance:

The two teams worked together on testing parts of the software where confusion could occur. Collected data was compared with converted and stored data. All modules went through unit tests then system test.

In addition to testing the application with invented data, the application was tested with 10 real sample surveys. The appendix shows a sample of one of such surveys. As the software was put to use, the staff from the school of public health started giving feedback on any discovered bugs in the system. The GUI was adjusted more than once to reflect new requirements. Major changes and new recommendation were left for upcoming releases and future collaborations.

Assessment and Institutionalization

The project was engaging to the CIS and Public Health students in many ways. To capitalize on the success of this experiment and benefit future students, the leaders of the project, Dr. Quweider and Dr. Perez, sought steps to reproduce similar collaborative outcomes in the future in a methodical way. Since the CIS students are required to finish a capstone project during their senior year, it was thought very useful to allow interested student to pursue a similar experience while earning credit for graduation. While taking the senior project, students who pursue this path will be asked to follow the steps.

Project Selection: Student(s) will be given a range of projects to choose from. The selection will be coordinated by a joint faculty from the CIS and the Public Health departments.

Project Presentation: Student(s) will be required to present the project upon its completion to the rest of the class; they will detail their experience and point out any improvements or modifications when they fill the course evaluation form.

Project Dissemination: Both an Oral Report as well as a Written Report will be required. Written reports will be made available to future classes as samples.

It is important to note that such interdisciplinary projects feed directly into some of the Department's Objectives of graduating students who: demonstrate a good understanding of mathematics, show proficiency in the use of analytical and problem-solving skills, able to apply their design skills, show proficiency in written and oral communication, able to work in a multi-disciplinary team environment, and appreciate the need for lifelong learning.

Conclusion

One of the major outcomes of this experience was the increase student's awareness of the role other disciplines play when writing software. In addition to getting familiar with other disciplines on an intimate level, the students realized that transforming real-life problems into a software system takes constant interaction between technical and non-technical people and between experts in seemingly different fields. No one doubts the role of computers in our present and future and through collaborative inter-disciplinary projects the role for computer scientists is more defined and easier to grasp. By opening the door of collaboration for future student, they now have a chance to engage their computer skill to solve real life problems in medicine, public health, and patient management in addition to the more traditional technology oriented applications. We also hope that the department's concerted effort and repeated assessment and modification will generate more collaborative programs across many fields.

Future Work

Enhancements for the systems are under way. As non-technical users (i.e. the health professionals) are using the system, they are reporting not only bugs but also features they would like to have. The next release of the software will be updated to reflect these features. Computer science majors can choose to work on the project to work on as part of their undergraduate work. More importantly, the involved departments, The University of Texas Health Science Center at Houston and the University of Texas School of Public Health, Brownsville Regional Campus, and the Computer Science department, now allow CIS students to choose projects of mutual interest to be used in their senior project, which is a required course for graduation.

Acknowledgment

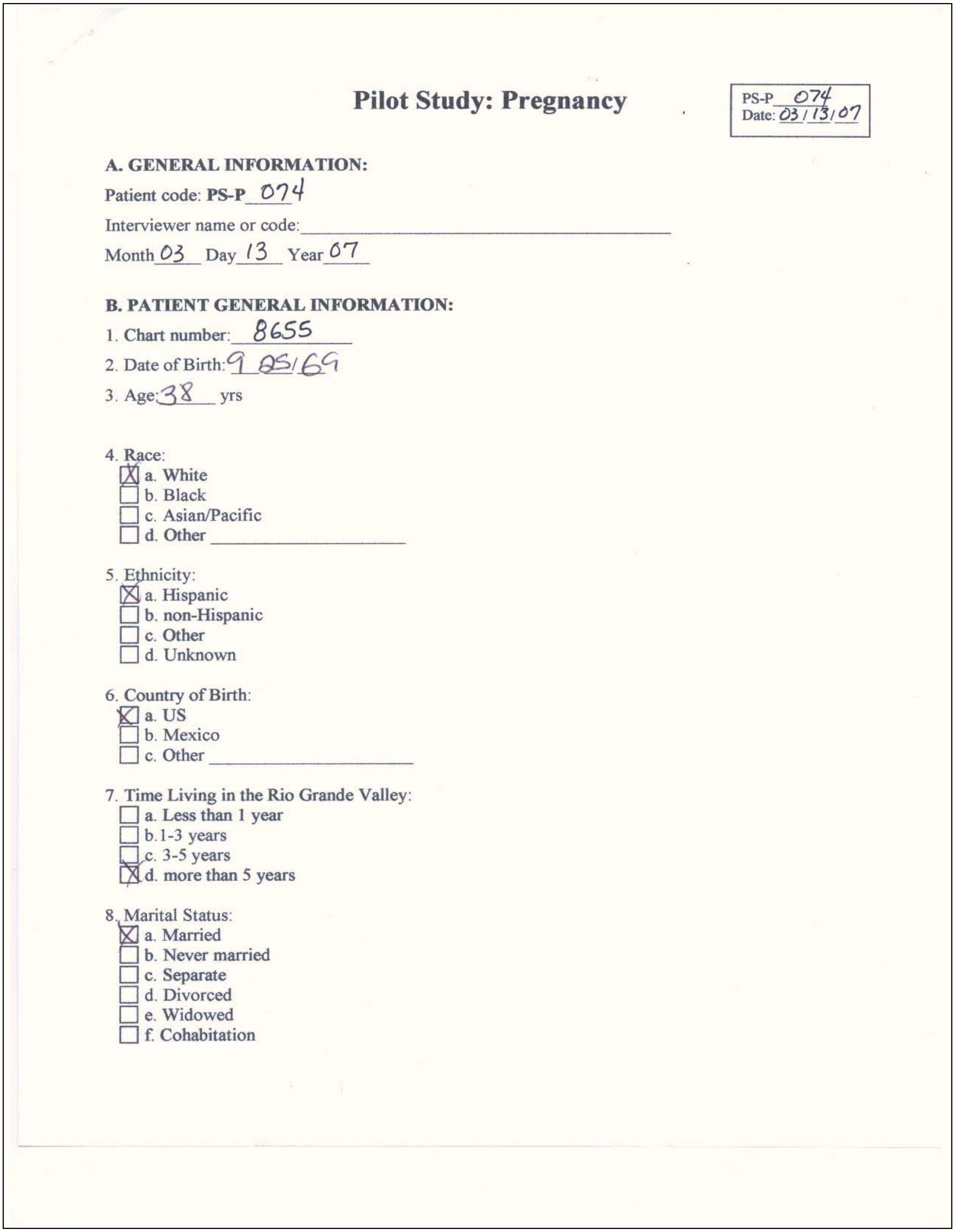
Special thanks to Dr. Adriana Perez from the School of Public Health for providing the subject of this study and for facilitating access to the surveyed data and software tools used for analysis. Thanks to Gabriela Oropeza for conducting the initial software development as part of her internship.

References

1. Committee on Science, Engineering, and Public Policy. Facilitating Interdisciplinary Research. National Academies Press, Washington DC, 2004
2. Kurland and Rawicz, Involving students in undergraduate research and development: two perspectives, ASEE/IEEE Frontiers in Education Conference, 1995.
3. Madler, L., Genesis of an undergraduate research experience, ASEE/IEEE Frontiers in Education Conference, 1998.
4. Anwar, S. and P. Ford. Use of a Case Study Approach to Teach Engineering Technology Students. International Journal of Electrical Engineering Education, 38 (1), 2001.
5. http://research.microsoft.com/towards2020science/background_overview.htm.
6. Jacobson, Iror, Griss, Martin and Jansson, Patrick, Software Reuse Architecture, Process and Organization for Business Success. Addison Wesley, 1997.

Appendix A

Data entry forms with filled data



Pilot Study: Pregnancy

PS-P 074
Date: 03/13/07

A. GENERAL INFORMATION:

Patient code: **PS-P** 074

Interviewer name or code:____________________

Month 03 Day 13 Year 07

B. PATIENT GENERAL INFORMATION:

1. Chart number: 8655
2. Date of Birth: 9 25/69
3. Age: 38 yrs

4. Race:
☒ a. White
□ b. Black
□ c. Asian/Pacific
□ d. Other ________________

5. Ethnicity:
☒ a. Hispanic
□ b. non-Hispanic
□ c. Other
□ d. Unknown

6. Country of Birth:
☒ a. US
□ b. Mexico
□ c. Other ________________

7. Time Living in the Rio Grande Valley:
□ a. Less than 1 year
□ b. 1-3 years
□ c. 3-5 years
☒ d. more than 5 years

8. Marital Status:
☒ a. Married
□ b. Never married
□ c. Separate
□ d. Divorced
□ e. Widowed
□ f. Cohabitation

PS-P______
Date: ___/___/___

9. Education:
- □ a. None
- □ b. Elementary
- □ c. Middle school
- □ d. Some High School
- □ e. Graduated High School/ GED
- ☒ f. Graduated College/ University

10. Type of Occupation:
- □ a. Housewife
- ☒ b. Works in Office/Business/shopping mall
- ☒ c. Works in Education Teacher/Student/Administrator
- □ d. Works with agricultural produce: Food store/warehousing
- □ e. Works on a farm in agriculture
- □ f. Other outdoor occupation (truck driver, etc.)
- □ g. Disabled
- □ h. Unemployed
- □ i. Other
- □ j. Unknown

11. Language:
- □ a. English
- □ b. Spanish
- ☒ c. Both
- □ d. Other

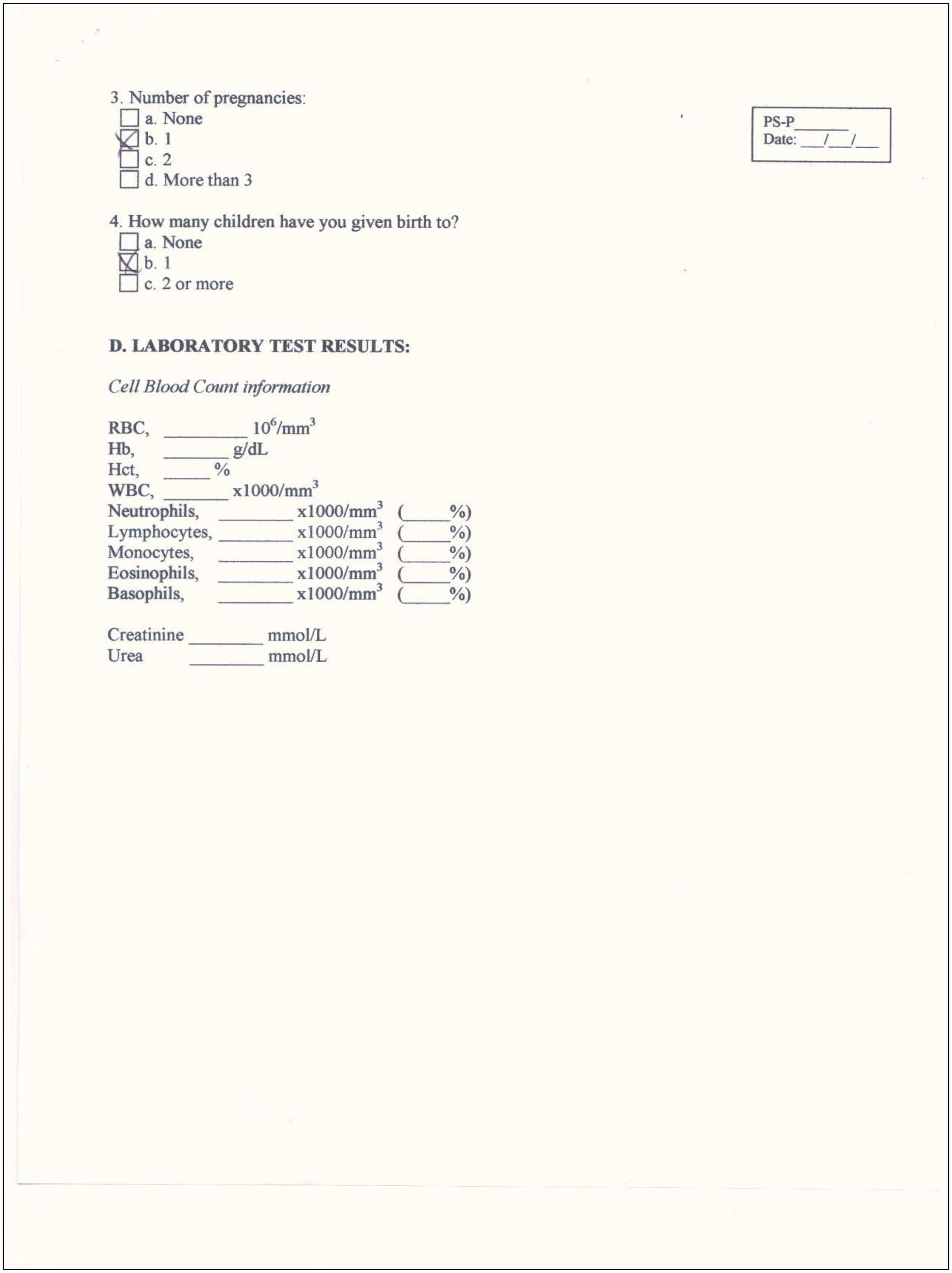
12. Do you smoke cigarettes?
- □ a. Yes
- ☒ b. No, I never smoke
- □ c. I used to smoke cigarettes, but I quit.

13. How many cigarettes do you smoke per day?
- □ a. 14 or fewer cigarettes a day
- □ b. Between 15 and 25 cigarettes a day
- □ c. More than 25 cigarettes per day

C. MEDICAL HISTORY:

1. Weight 168

2. Height: 5'3



PS-P______
Date: ___/___/___

3. Number of pregnancies:
☐ a. None
☒ b. 1
☐ c. 2
☐ d. More than 3

4. How many children have you given birth to?
☐ a. None
☒ b. 1
☐ c. 2 or more

D. LABORATORY TEST RESULTS:

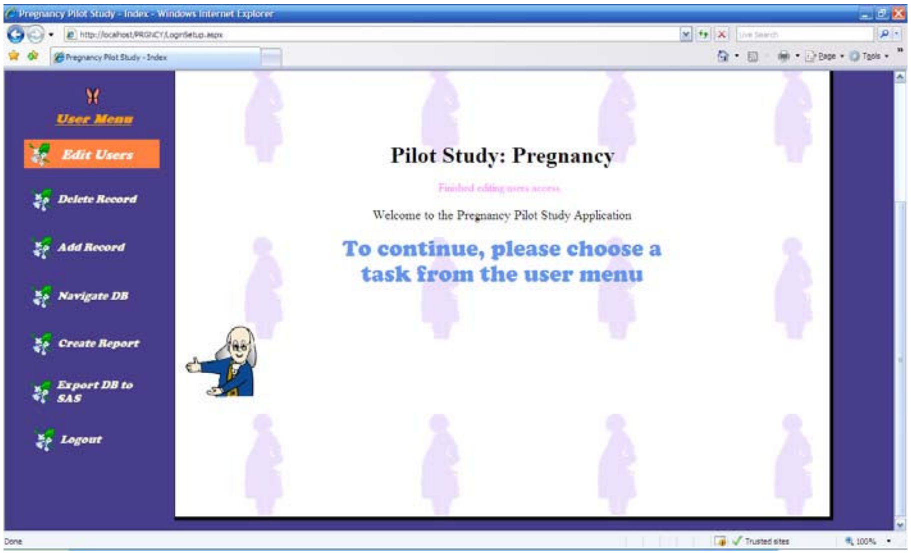
Cell Blood Count information

RBC, _________ $10^6/mm^3$
Hb, _______ g/dL
Hct, _____ %
WBC, _______ $x1000/mm^3$
Neutrophils, ________ $x1000/mm^3$ (_____%)
Lymphocytes, ________ $x1000/mm^3$ (_____%)
Monocytes, ________ $x1000/mm^3$ (_____%)
Eosinophils, ________ $x1000/mm^3$ (_____%)
Basophils, ________ $x1000/mm^3$ (_____%)

Creatinine ________ mmol/L
Urea ________ mmol/L

Appendix B

Sample Output Screens



Pregnancy Pilot Study - Index - Windows Internet Explorer

http://localhost/PRGNCY/LoginSetup.aspx

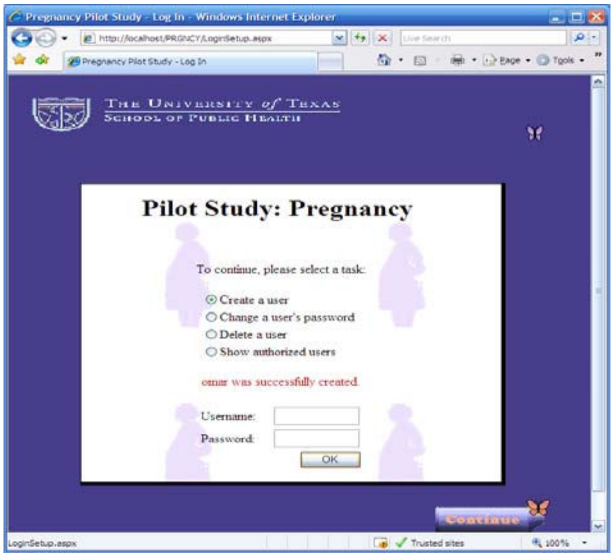
Pilot Study: Pregnancy

Finished editing users access.

Welcome to the Pregnancy Pilot Study Application

To continue, please choose a task from the user menu

Initial Screen Display



Pregnancy Pilot Study - Log In - Windows Internet Explorer

http://localhost/PRGNCY/LoginSetup.aspx

The University of Texas School of Public Health

Pilot Study: Pregnancy

To continue, please select a task:

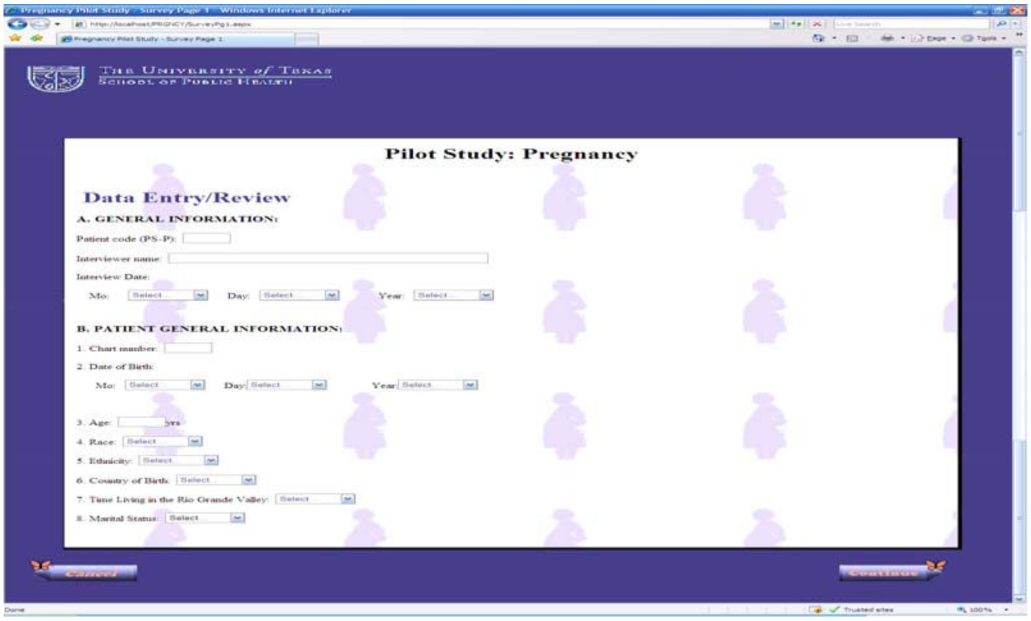
● Create a user
○ Change a user's password
○ Delete a user
○ Show authorized users

omar was successfully created.

Username:

Password:

Administrator log-in.


Pregnancy Pilot Study - Survey Page 1 - Windows Internet Explorer
http://localhost/PRGNCY/SurveyPg1.aspx
Pregnancy Pilot Study - Survey Page 1
Live Search
Page
Tools
The University of Texas
School of Public Health
Cancel
Continue
Done
Trusted sites
100%

Pilot Study: Pregnancy

Data Entry/Review

A. GENERAL INFORMATION:

Patient code (PS-P):

Interviewer name:

Interview Date:

Mo: Select Day: Select Year: Select

B. PATIENT GENERAL INFORMATION:

1. Chart number:

2. Date of Birth:

Mo: Select Day: Select Year: Select

3. Age: yrs

4. Race: Select

5. Ethnicity: Select

6. Country of Birth: Select

7. Time Living in the Rio Grande Valley: Select

8. Marital Status: Select

Data Entry Form