Paper 181-28

Managing Clinical Trials Data with a SAS-Based Web Portal

J. Philip Miller, James Schauer, Jack D. Baty, Sara Littlewood, Kathryn Trinkaus,
Robert M. Lill, Reuben Richards
Division of Biostatistics, Washington University School of Medicine
and the Public Sector Group, SAS Institute

ABSTRACT

The proliferation of many small clinical trials at the A. J. Siteman Cancer Center challenges us to collect electronic data in a cost-effective way. The Siteman Clinical Information Portal (SCIP) was created by customizing a previously-developed suite of browser-based form design tools. SCIP allows for the creation and management of data collection instruments by nonprogrammers and supports web-based data entry and editing. The core of the tool set is a form-writing engine driven by a simple model of databases that contain both form and item metadata. The items, valid response options, online edit checks, and data handling properties, are all elements in the data model. SCIP provides an audit trail, patient and protocol management tools and client-side validation rules. The design of SCIP allows the creation of a library of frequently used items, reducing the forms implementation effort. Access to the forms is then controlled with a note tab-based portal model which provides access controls and customization of the interface for each user. The portal also supports access to other web-based resources, e.g., electronic versions of protocols. SCIP executes under SAS/IntrNet® on a Linux server.

THE PROBLEM

The Alvin J. Siteman Cancer Center (SCC) at Washington University School of Medicine and Barnes-Jewish Hospital was recently recognized as being one of the 61 National Cancer Institute designated cancer centers. As part of its research portfolio, a variety of clinical studies are conducted within the SCC. A priority of the SCC is to facilitate this research, particularly for institutional studies. The SCC is organized around a “matrix” model where each member’s primary appointment is in one of the academic departments. Faculty apply for membership in the SCC and their credentials are reviewed by a committee which evaluates their cancer related achievements. Since the members come from all of the clinical and basic science departments of the medical school, core resources were developed to be available to SCC members.

VOLUME OF STUDIES

At any given time, there are about 500 clinical studies active, half of which are clinical trials open to enrollment. Of these, about 75 are institutional studies. The remainder of the clinical trials are primarily those done under contract with the pharmaceutical industry or as part of cooperative groups. Institutional studies typically have recruitment goals of 20-50 participants over a period of several years. Many of the institutional studies do not receive support from external sources and must compete for scarce internal infrastructure resources.

CURRENT METHODS

Prior to the establishment of the SCC, each investigator was responsible for establishing their own data management procedures. While a clinical research associate (CRA) was normally responsible for abstracting the necessary information from clinical and research charts, the information was usually simply recorded on paper forms. Frequently these forms were not designed for electronic data entry and the information was subsequently entered into a spreadsheet for analysis by the CRA or the investigators themselves. Often there was not a controlled set of responses defined for each item, leaving a large data cleaning task for the statistician trying to analyze the data.

SOLUTIONS CONSIDERED

We have a long history of developing data entry/data management systems for individual studies (Achtenberg & Miller, 1975; Moore et al., 1976; Achtenberg et al., 1976; Miller et al, 1977; Achtenberg & Miller, 1978; Moore et al, 1978; Miller 1979; Krone et al., 1979; Krone et al., 1981; Mandel & Miller, 1988; Trinkaus et al., 2000; Thompson et al., 2002). In recent years, these have been primarily implemented with a web interface with SAS/IntrNet® as the back-end. We have developed a web-based system for the handling of Phase I studies conducted at the SCC. Phase I cancer studies are the first studies done in humans and are usually designed to determine the maximum dose of the new agent which can be tolerated without serious side effects. Often blood samples are obtained at frequent intervals to understand the pharmacokinetics of the new agent and its metabolites. At the SCC, most of these studies are conducted by a single research office with a template protocol and a standard set of data collection forms facilitating the development of a customized system which can be used for most of the Phase I studies.

As we contemplated what was needed to support the much more heterogeneous Phase II and III studies, we realized that the development cycle was too expensive to create a new data entry system for each study. Although we had prior experience with creating a generalized system (Achtenberg et al., 1976), the developmental costs for such a system were considered too great. We also have developed a flexible development environment to support web-based data entry for some of our multi-institutional studies where we function as the coordinating center (Thompson et al, 2002). This system requires a SAS programmer to customize each form. While the amount of effort to produce such data entry forms is much less than coding from scratch and is appropriate for large studies with many participants at different locations, the expense to set up each new clinical trial was too great.

We also considered a number of commercial systems designed for clinical trials for the pharmaceutical industry. These systems were both expensive and designed for a regulatory environment which far exceeded the needs of our studies. The system design was usually based on a client/server architecture with a proprietary client needed by the end user. The implementation tools frequently required an IT skill set. The computing environment at Washington University Medical Center (and thus the SCC) is diverse and there is no centralized desktop support for all users. This means that solutions which required the installation and support of proprietary clients would be an expensive ongoing solution.

The resources of the SCC Biostatistics Core and the SCC Clinical Trials Core are limited and are difficult to prioritize. If the investigative team could set up the data entry screens themselves, then it would be more efficient, particularly for the maintenance activities normally encountered over time. Our own experience with a thin client data entry system using a web

browser and SAS/IntrNet® as the server side engine has been positive.

The Public Sector Group at SAS Institute developed a system for the Department of Education for support of their responsibilities under Title II of the 1998 reauthorization of the Higher Education Act. The Secretary of Education is required to issue annual reports to Congress on the state of teacher quality nationwide. In order to implement this, a web-based system was established to allow institutions which train teachers and state reporting agencies to enter the required information for subsequent reporting. This application has been generalized and is currently known as the Data Acquisition Portal (DAP).

This solution was particularly attractive to us since the entire developer and user interface is accessed using a web browser. It also utilizes a system and computing environment with which we have considerable expertise. A contract was developed between SAS Institute and the SCC that outlined the changes in the system that would be necessary in order to support our needs for clinical trial data.

OVERVIEW OF SCIP

We have implemented SCIP using the DAP application which was customized for our use. It is both a way of centralizing information for those working on clinical trials at the SCC and a way of facilitating the electronic capture of the necessary information for the conduct and analysis of clinical trials.

At the core of SCIP is a form writing engine which will dynamically generate a form that can be used by the CRA to enter the information about a particular participant in a clinical trial. This data is stored in a normal SAS dataset on the server, where it can later be used for analyses and reports. This engine is table driven with all tables stored as SAS datasets.

SCIP supports the development of a library of questions which comprise the data elements for the study database. The developer uses the web-based interface to populate the question library. With each question, not only is the text of the question stored, but also all necessary metadata. This metadata includes such attributes as the actual values to be stored in the actual study database for each choice, whether the item is to be required and keywords which can be used to search the question database. This architecture allows the development of standardized items which can then be used in many studies. We have worked to develop recommended data elements and sample forms to be used in SCC studies. When developing a new study, finding a data item which has already been developed is more efficient than inventing a new one.

These questions are then aggregated into a form using other web-based tools to build the metadata necessary for constructing the forms for data entry. The responses to each form will be stored in a standard SAS dataset. Efficiency will be gained by reusing forms from one study for another study. Even if the prior form is just used as an initial draft, the savings can be appreciable. The forms are then aggregated into a study along with a study definition to indicate which forms are to be completed at which visits.

Developers work with a series of pop-up windows where the necessary information is entered and submitted to SCIP. The window is then dismissed.

For the CRA (or other study staff) SCIP then provides a tool with a visual representation of the forms to be completed for each participant along with the status of data entry for that participant.

The CRA can also produce a variety of reports (delivered to the web browser) to track the progress of the study.

QUESTION LIBRARY

At the heart of SCIP is the creation of a question library. Developing data collection forms and the creation of appropriate data elements is a time consuming aspect of setting up a new trial. At the SCC, in collaboration among the Clinical Trials Office, the Biostatistics Core and individual investigators we are developing model forms for use in SCC clinical trials. One input into our efforts is the Common Data Element (CDE) project (http://cii-server5.nci.nih.gov:8080/pls/cde_public/cde_java.show). The CDE project is an NCI-sponsored initiative designed to standardize and simplify the collection and reporting of data for clinical trials, patient surveys and cancer patient care by collaboratively developing uniform and explicit data elements based on recognized standards. By identifying questions and the list of valid responses, this effort is intended to provide building blocks for trial development and promote data sharing and data mining through the use of common terminology.

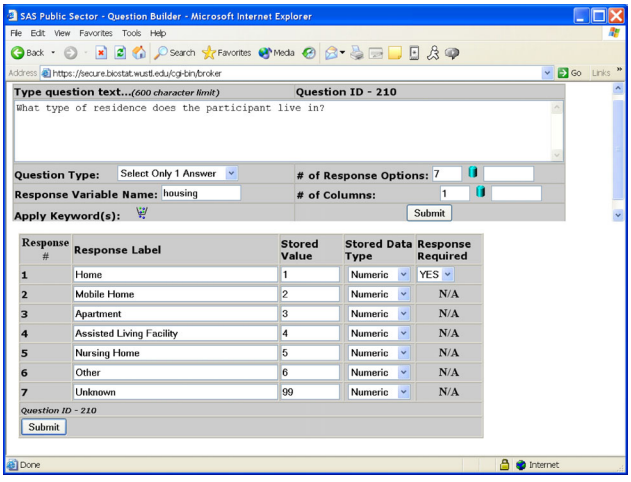
While the CDE project still covers only a portion of the areas needed to support the trials in the SCC, we will use those CDE projected developed questions supplemented by those identified as “best practice” from our own experiences. In this way, the developer can quickly create new forms.

A variety of question types are supported by the DAP code. While many of our customized systems support more question types (and greater flexibility in terms of window layout), this ensemble appears to cover most of the data items which are encountered in our clinical trials.

TYPES OF QUESTIONS

SELECT ONLY ONE ANSWER

The most common type of data item is where the allowable response is one of a set of mutually exclusive choices. After choosing to add a new item, a pop-up window allows for the specification of the text of the question, the number of responses, the value to be stored for each response, a response variable name which will be used in creating the SAS variable name for that item, the labels for the various responses and whether a response is required for this item. The example screenshot shows the creation of an item to capture the type of housing that the participant lives in.



When SCIP will present this item, the responses will appear as radio buttons, e.g.

Question 210

What type of residence does the participant live in?

- Home
- Mobile Home
- Apartment
- Assisted Living Facility
- Nursing Home
- Other
- Unknown

The green dot beside the first response indicates that this question must be completed. Instead of the radio button presentation, it is also possible to have the item appear with a drop down list. This can help to save on data entry screen real estate.

Question 220

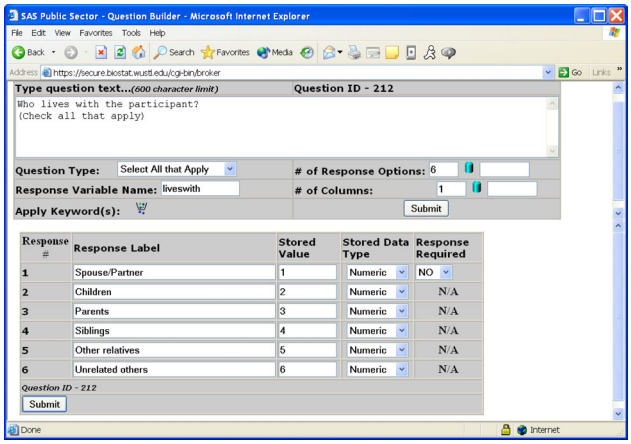
What type of residence does the participant live in?

Own home
Mobile home
Apartment
Assisted living facility
Nursing home
Other
Unknown

When using the drop down list question type, it is also possible to use a predefined SAS Format to generate the response labels (and corresponding stored codes). Using this methodology assures an alignment between the format for the use of the dataset and the input options. It is also useful for questions with a large number of possible responses, e.g. states or drugs. Easy SAS programs can be written which take external valid value definitions and create appropriate SAS Formats.

SELECT ALL THAT APPLY

When the alternatives are not mutually exclusive, a very similar layout is used.



When this item is presented, it uses check boxes rather than radio buttons.

Question 212

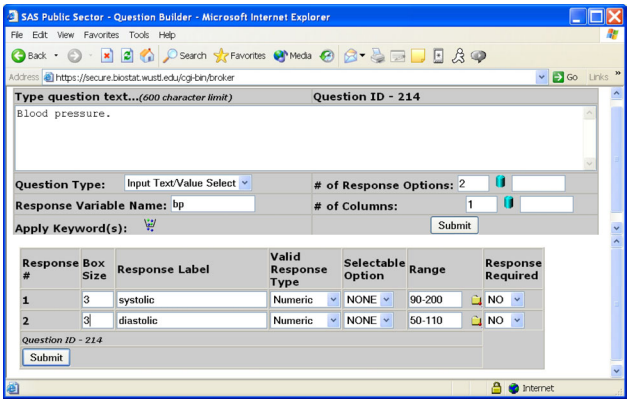
Who lives with the participant? (Check all that apply)

- Spouse/Partner
- Children
- Parents
- Siblings
- Other relatives
- Unrelated others

All of these methods use a controlled response set providing a well structured dataset. Since the data entry forms are created dynamically, if additional responses are discovered during the course of the study, the developer can easily edit the question and redeploy the study and data entry then proceed with the revised question.

INPUT NUMBER/TEXT

Many data items require the entering of a specific value. SCIP allows for numeric or character response types. For numeric values, an allowable range can also be specified. The range checks are implemented with the use of SAS Formats. This provides considerable flexibility. For example, mixed ranges (1.4-3.4, 5.2-6.8) are valid as well as low to high definitions (5.5 to high). If there are several values for the data element, items can be created with multiple boxes for input.



This data item is then presented with the responses labeled as specified.

Question 214

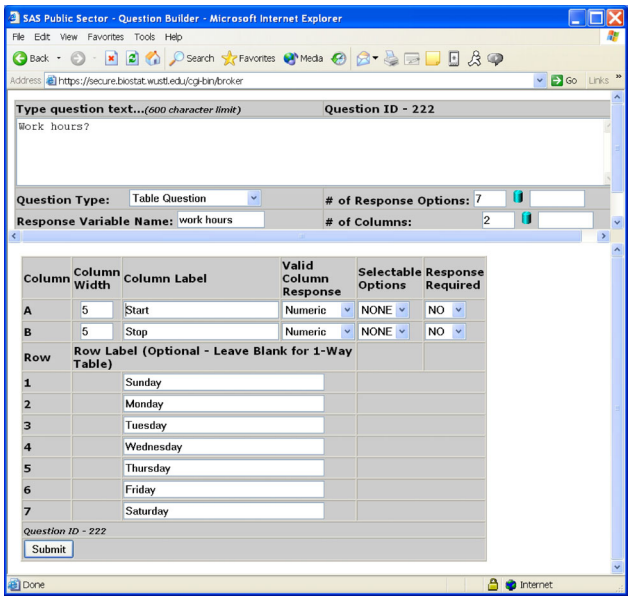
Blood pressure.

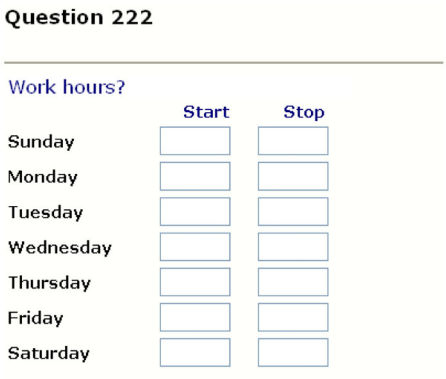
systolic

diastolic

TABLE ORIENTED

The boxes can also be arranged into a table for data items which are more logically organized in that fashion.

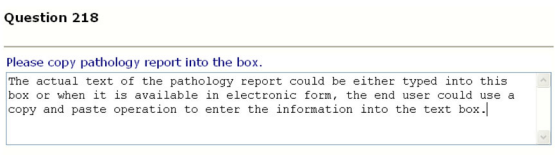




These tables can be made more complex with different response types as different columns. It is also possible to nest a particular question as a column in a table. This allows, for example the creation of a Likert scale question which is then applied repeatedly within the table. The advantage provided is the option for more complex combinations of question types and the re-use of a single question definition over and over again.

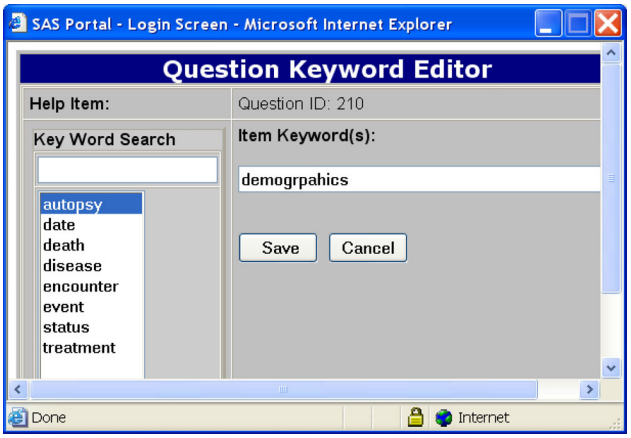
LARGE TEXT BOX

Actual text can be entered into large text boxes. The information can either be typed into the box or the user could use the cut and paste operation of the windowing system on their desktop to place the information into the box.

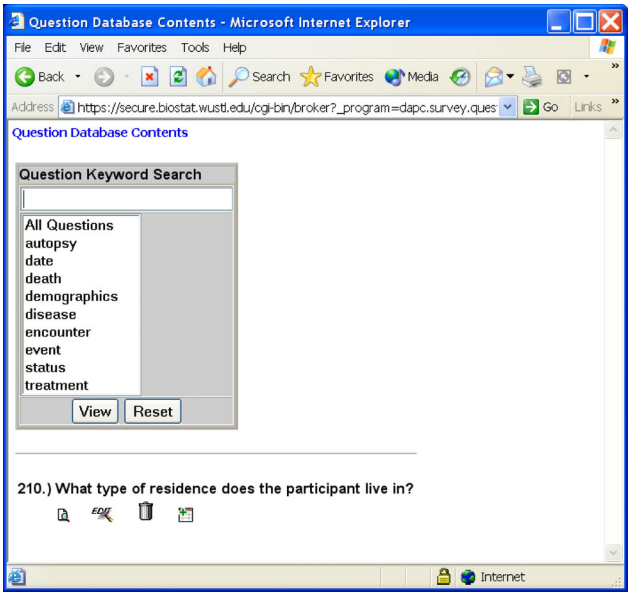


MANAGING THE QUESTION DATABASE

In order to better navigate the question data base, a question keyword editor has been provided where key words can be attached to each question. The panel on the left shows those key words which have already been assigned to assist in the selection of the keyword(s) for the question.



The questions are then searched using a simple selection which displays either all questions or those which have the specified key word attached. The select list for the key words to be used is built dynamically with the question keyword editor.

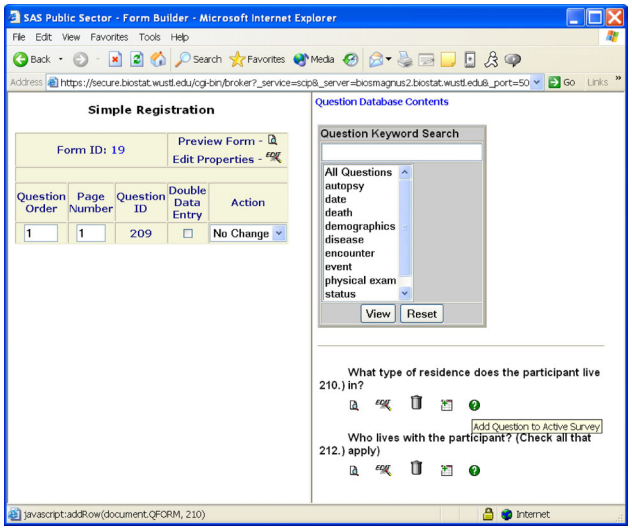


Along with the text of each question are icons which allow the question to be previewed (displayed as if it were part of a form), edited, deleted or copied to form the basis of another question.

FORM CREATION

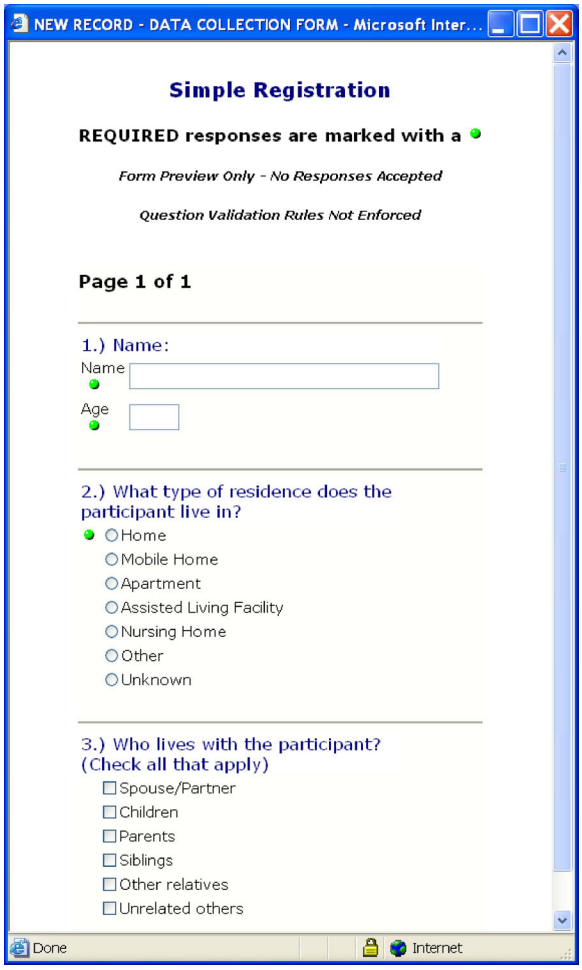
BUILD A NEW FORM

After all of the questions for a form have been either verified to exist in the database or be newly created, a form can be created by adding questions with a tool which has the same search engine as is used for building the question database in the right panel. A list of the questions currently on the form is displayed in the left panel. With each question found in the database, the same icons for editing and displaying the question appear along with an icon for adding the question to the form.



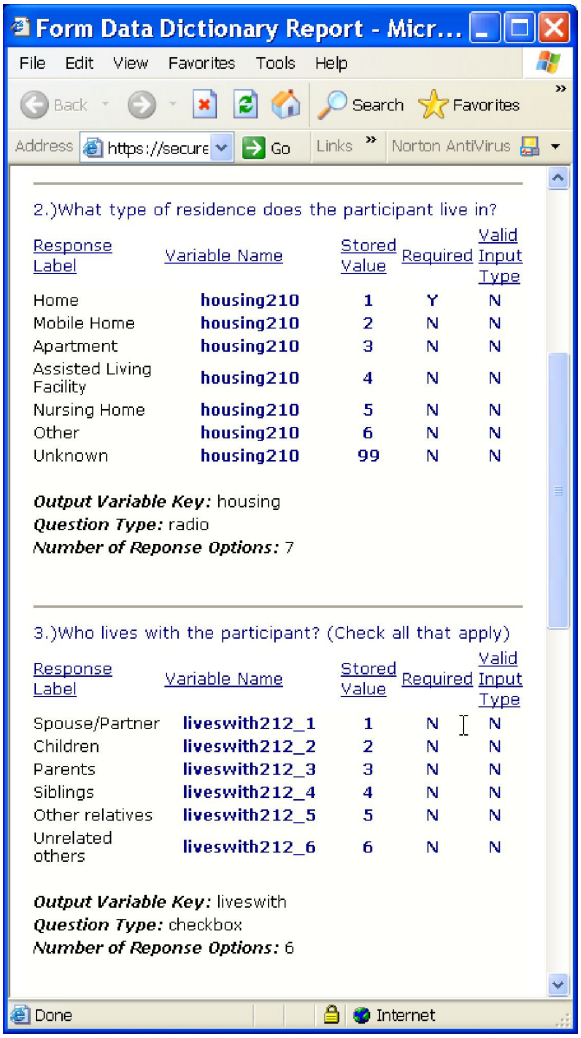
For each study, a registration form is required which identifies a new participant in the study. The first question of that form is critical as it is used as the index to identify the specific subject. By using a multipart question, it is possible to use multiple fields to identify the individual, e.g. study ID and name.

It is possible to view the form at any time during the creation process to verify its content. While the radio buttons, check boxes, etc. are functional, actual data cannot be added while in the form builder tool.



The layout of the form is simple, with the questions listed in order along with the appropriate widget to allow for data entry. This simple linear layout was one of the compromises which needed to be made in order to have the layout dynamically created. For our customized data entry applications we tend to use more layout control in order to make the browser window resemble the paper form and to conserve space on the paper form. Current users consider this an acceptable trade-off.

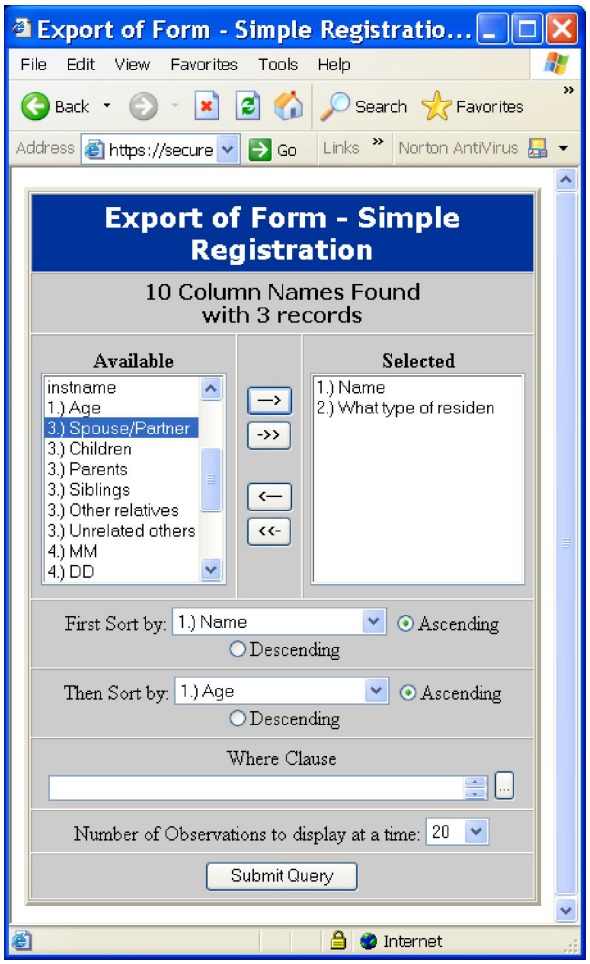
The forms design tool also allows the creation of a data dictionary for the SAS dataset which is created to contain the data entered for that form, e.g.



Since the metadata for the forms are all stored as SAS datasets, more complex data dictionary reports can then easily be created by simple SAS programs.

As would be expected, when the question type allows only a single response, a single SAS variable is created using the name supplied when creating the question (e.g. “housing”) along with the internal question number (210). This naming convention then helps to avoid name collisions. When multiple responses are allowed, a suffix is added to the variable name stem.

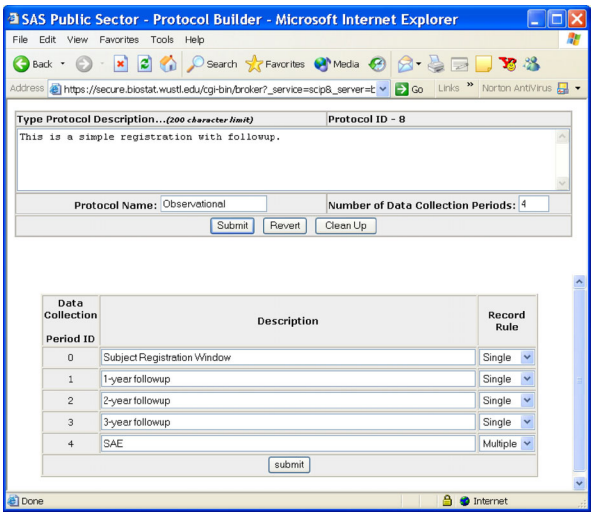
The forms design tool also has an export facility which can be used to generate a simple report. The tool has a familiar format with two boxes to select which variables in the data set are to be used. The export facility can sort the observations by any two variables in the dataset, use a SAS WHERE clause to subset the observations, and control the page size.



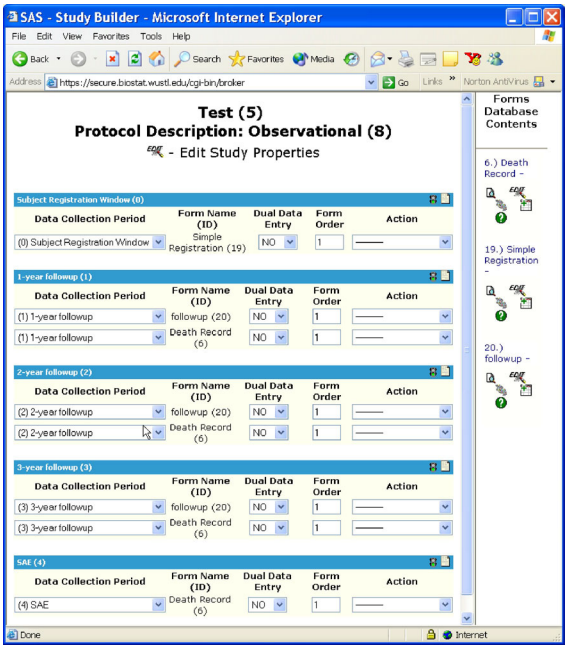
For more complex forms, the questions can be arranged into pages. This reduces the download time during data entry making the system appear more responsive and makes the process less cumbersome. In addition, since the responses are stored only on the user's computer until the submit button is clicked, the use of pages allows for the saving of responses after each page is completed. Skip logic can also be specified so that particular pages can be skipped depending on responses to earlier questions.

STUDY DEFINITION

A study is then defined by specifying a data collection protocol. The template for the protocol defines a set of data collection time intervals (windows). The first of these must contain the registration form which adds a new participant to the study. Other windows are then defined along with an option specifying a single instance of a form for that window or multiple instances. This allows buckets to be defined with a variable number of instances of the form. Examples of multiple forms within a window are for serious adverse events (SAEs), hospitalizations or other recurring event reports.

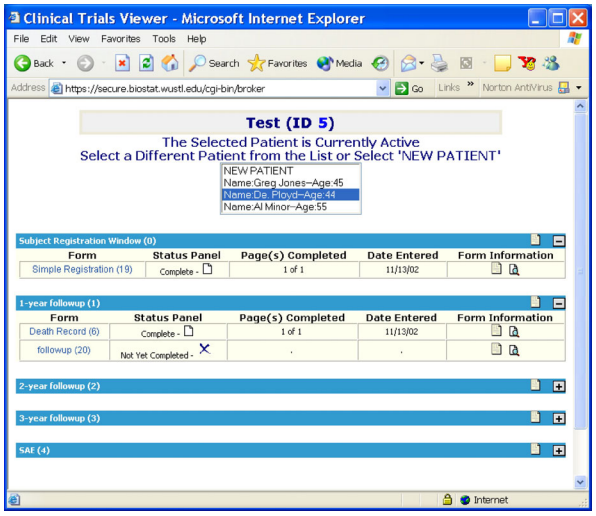


The study development tool then allows specific forms to be associated with each study window.



FORM COMPLETION

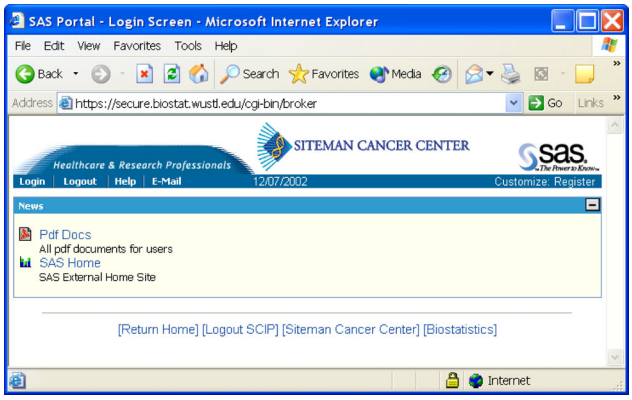
To enter data, the user chooses the appropriate study on the initial portal page. A selection screen appears which allows a particular participant to be selected and displays the windows defined for that study. Each window can be expanded to show each form that was defined for that window and its status.



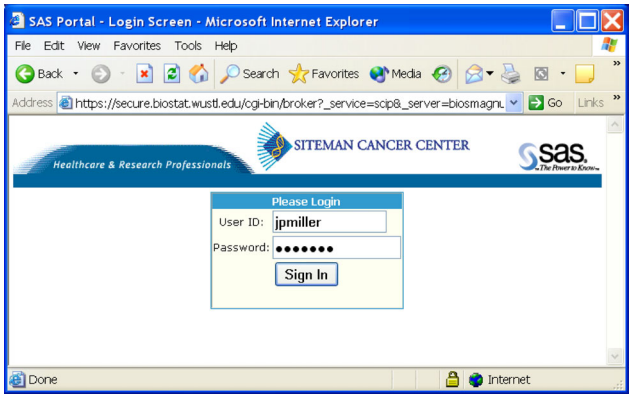
By clicking on the form within a study window, that particular form can be entered or previous responses can be edited.

PORTAL INTERFACE

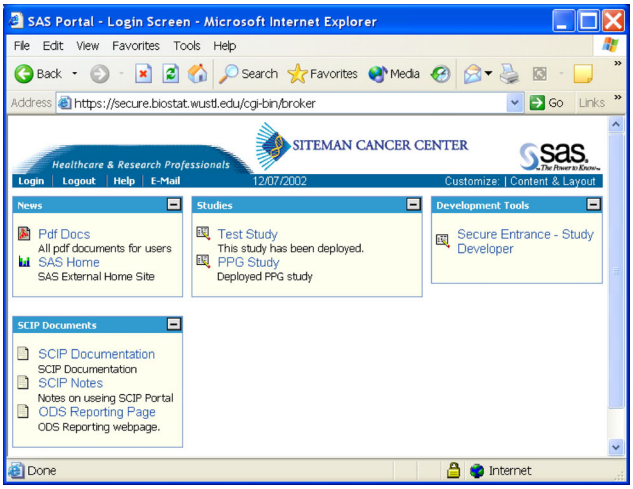
The interface to SCIP is based in a portal interface which can be customized by each user. On initially going to the site, a public interface to the portal is presented. One of the tabs presented in this view is to login. The system supports an automatic registration module if you want to allow new users to register to receive a userid with limited privileges.



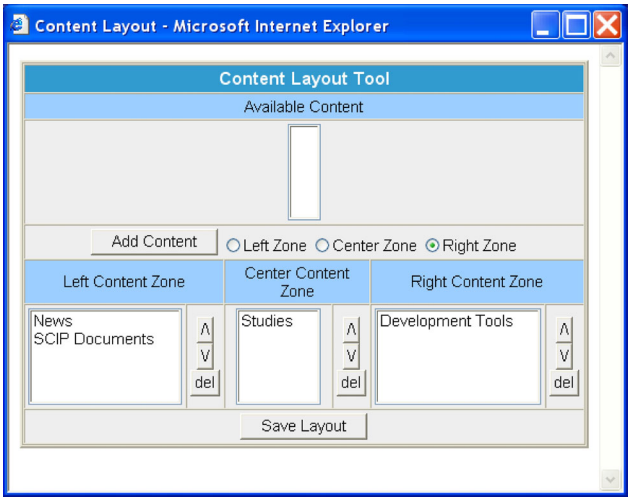
The login is presented in a new window. The system uses a strong security model (see below).



Upon logging in, the view that the user is presented with is based on those components appropriate for the particular user and is customizable by each user using a link on the top tab-bar. The layout for a developer might appear as:



Following this link brings up a popup window which allows individual content modules to be arranged into the left, center or right hand column of the window.



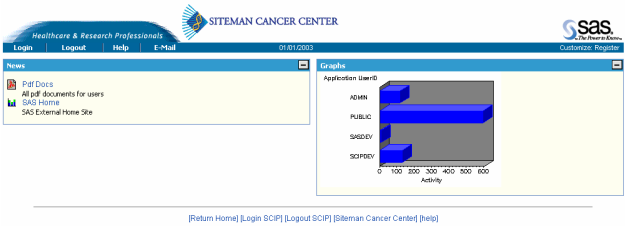
OTHER RESOURCES AVAILABLE

The DAP allows the administrator to add additional content material that the user can add to their portal interface. For SCIP we have provided a link to an electronic protocol facility where electronic versions of the consent forms, protocols, study order sheets, etc. can be accessed for many of the SCC clinical trials. Another content provided is a link to a web based tool that allows the user to access the tumor registry for patients treated at the SCC during the last 16 years. The tool (implemented with SAS/IntrNet®) allows for counts and graphs of patients based on tumor location, stage of the cancer and demographics. This tool is particularly useful during the planning phase of new studies to determine the historical patient flow. Access to this tool is restricted based on the same userid/password database as is used for accessing SCIP.

It is also possible for the site administrator to create “protected” content items. These are pre-defined for a specific user group and therefore can not be customized by the end user (added or subtracted, moved from one zone [left, right middle] to another).

Another example of the use of protected content items are for the public user (who does not need to login to authenticate themselves). Since the public user does not even have a link to customize content – all of their content has been pre-defined for them.

The portal also has "web-services" capabilities. A web-service is a component request from one application to another application, which in turn, makes the particular functionality available to the requesting application. Instead of having to call the entire application that contains the component of interest, you may simply call the component of interest that is required to fulfill the request. This decouples the functionality from the application. For example we could add to the portal a simple horizontal bar chart of user activity.



This web service executes an SCL method which returns a horizontal bar chart showing user activity. Instead of having to click on a link to execute a report, the web service is run upon login and the html output from the web service is grabbed and placed within the portal content zone. This automatically provides users with the most up-to-date information.

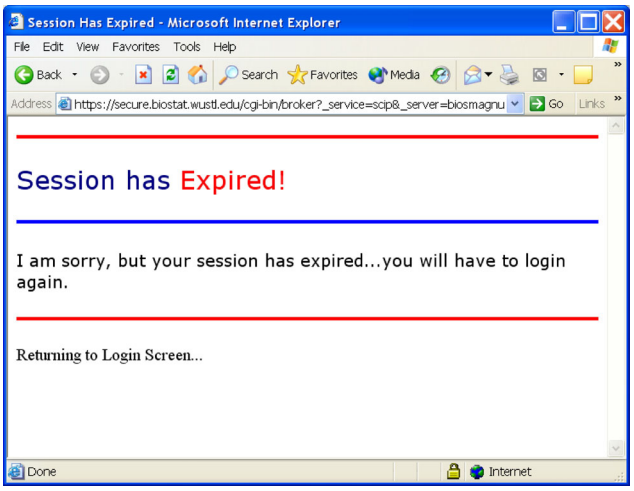
SECURITY MODEL

All research which collects data about the health of individuals needs to be particularly concerned about the security of the information in order to protect the subject's privacy. The medical community is currently involved in examining its security procedures in order to conform to new HIPAA regulations. SCIP is implemented using an Apache server running with the secure sockets layer (SSL) transport protocol for a web server (https). This ensures that all dialogs between the server and the user's browser are encrypted.

SCIP uses its own password management system. Users are added by the SCIP administrator only after a confidentiality agreement is signed by the user and approved by their supervisor. Password creation rules can be customized and can be forced to expire after a specified number of days. Unsuccessful login attempts are logged and after a specified number of unsuccessful attempts the account is locked (temporarily or permanently depending on the number of unsuccessful attempts). Each user can be associated with particular access level (view reports only, edit data, study developer, site administrator) along with a profile which defines their role in the study.

Passwords can be authenticated against either a local SAS encrypted dataset or against an LDAP server. Using an LDAP server allows for a single userid/password to be used for other applications. In SCIP we use a single LDAP database to also authenticate usage for related web based services such as access to the tumor registry statistics and electronic access to protocols, consent documents and necessary standardized orders. It could also be used to control access to protected web pages and even Linux logins.

The system also detects inactivity and forces a logout for that user.



The DAP provides a number of customizable reports for security administration. Since the logs are also SAS datasets, customized reports are easily created. These reports can be launched by adding additional content areas to the portal for systems administrators or by cron jobs.

REPORT GENERATION

For most of our coordinating center activities we have moved to the electronic publishing of study reports. With the SAS ODS facility it is easy to generate the reports in either HTML or PDF formats. Study specific reports are defined in SCIP with the reports being dynamically generated by users initiating SAS programs. Links to generate these reports are then included as content items which can be configured on the user's customized portal interface. Regularly scheduled reports are generated as PDF documents which are then e-mailed by cron jobs on the Linux system.

For many of the studies at the SCC there is considerable scientific interest in genetic analyses of samples obtained from subjects in ongoing SCC studies. For bioinformatics data mining activities, it is appropriate to deal with anonymized datasets. For these studies we have written SAS programs which access the clinical information entered into SCIP and creates a fully anonymized dataset by removing all study IDs except those that provide the link to the biological samples and deleting and granularizing the data elements to reduce the identifiably of the data, e.g. recoding the age to the age decade where the original data element is computed by subtracting the date of birth from the date of diagnosis and dividing by 365.25.

CONCLUSIONS

We have implemented a system which demonstrates the power of SAS for implementing a web-based system for the collection of data from clinical trials at the Siteman Cancer Center. By creating a library of questions we are able to bring more standardization across trials and increase efficiency in developing new data entry forms. Because all of the tasks involved in implementing new forms can be done with familiar web tools, the end users can be directly involved in creating and maintaining the forms and those who are directly involved in the collection of the data can also enter it. The system is implemented through a customized portal for each user and it uses a strong security model.

REFERENCES

Achtenberg J, Miller JP. Diabetes Center Patient Registry. **Proc 1975 MUMPS Users Group**, J Zimmerman (ed), MUMPS Users' Group, St. Louis, 1-7, 1975.

Achtenberg JA, Miller JP, Cryer P, Santiago J. Data 3 - A Forms Management System. **Proc 1976 MUMPS Users Group**, J Rothmeier, (ed), 1-8, MUMPS Users' Group, St, Louis, 1976.

Achtenberg J, Miller JP. Interfacing a MUMPS-Based Data Entry System to SAS. **Proceedings of the 3rd Annual Conference of SUGI**, RH Strand (ed), SAS Institute, Raleigh, NC, 161-7, 1978.

Krone RJ, Igielnik S, Miller JP. A Computerized Cardiology Report Generating/Data Management System. **Proceedings of the 1979 Computers in Cardiology**, IEEE 79-CH1462-1, 469-72, 1979.

Krone JR, Igielnik S, Miller JP. A Computerized Report Generating/Data Management System for a Cardiology Division. **Proceedings of the 4th Annual WAMI Meeting**, FB Begon (ed), Strasbourg, France, 277-82, 1981.

Mandel SA, Miller JP. A Data Base Interrogation System with the SAS System for Personal Computers. **Proceedings of the 13th Annual SUGI Conference**, SAS Institute, Cary NC, 681-5, 1988.

Miller JP, Moore P, Achtenberg J, Thomas LJ, Jr. Computerized Information Handling for Long-Term Clinical Studies. **Proceedings of the 1977 Computers in Cardiology**, IEEE Catalog 77-CH-1243-2C, 151-9, 1977.

Miller JP. The Role of SAS in the Support of Clinical Studies in Medicine. **Proceedings of the 4th Annual Conference of SUGI**, SAS Inst. Raleigh, NC, 300-5, 1979.

Moore P, Miller JP, Moran M, Clark KW, Oliver GC. A Management Information System Designed to Oversee a Clinical Study, **Proceedings of 1976 MUMPS Users Group**, J Rothmier (ed), 85-96, MUMPS Users Group, St. Louis, 1976.

Moore P, Miller JP, Oliver GC, Thomas LJ, Jr. Task-Oriented Management Information Systems for Clinical Studies. **Proceedings of 1978 Computers in Cardiology**, IEEE 78-CH-1391-2C, 149-54, 1978.

Thompson PA, Littlewood S, Adelman AJ, Miller JP. The Web Data Entry System: Methods for Web Development and SAS Data Management, **Proceedings of the SAS Users Group International**, SAS Institute, 2002.

Trinkaus KM, Thompson PA, Miller JP. Using SAS/IntrNet Software to Support Distributed Data Entry, in **Proceedings of the Twenty-Fith Annual SAS Users Group International Conference**, SAS Institute, 1212-1217, 2000.

ACKNOWLEDGMENTS

Supported in part by NIH grant P30 CA91842.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Corresponding Author Name: **J. Philip Miller**
Company: **Washington University School of Medicine**
Address: **Campus Box 8067, 660 S. Euclid**
City state ZIP: **St. Louis, MO 63110**
Work Phone: **314-362-3617**
Fax: **314-362-2693**
Email: **phil@wubios.wustl.edu**
Web: **www.biostat.wustl.edu/~phil**