FROM: Zweben A, Barrett D, Carty K, McRee B, Morse P, Rice C. Strategies for Facilitating Protocol Compliance in Alcoholism Treatment Research. Project MATCH Monograph v. 7 (NIH Pub. No. 98-4144). Bethesda, MD: National Institute on Alcohol Abuse and Alcoholism (NIAAA), 1998.

Setting Up and Maintaining an Electronic Data Base

Christopher Rice, Ph.D., Lisa Freda, and Constance Lawson
School of Social Work, State University of New York at Buffalo

Conventional strategies for monitoring participant compliance place the burden of adhering to the research protocol on the participant. There may be no theoretical formulation for this, nor does the current empirical evidence indicate that participant-centered compliance is the most effective strategy. The convention appears to stem from the notion that an interactive strategy that anticipates rather than simply reacts to protocol participation barriers is too costly. To interactively monitor each participant's progress through the research protocol requires staff time over and above the essential tasks of assessment and data entry. Hence, efforts to enhance participant compliance usually involve adding staff. However, the accompanying cost increase makes this solution unattractive.

A computerized, participant-tracking data base easily fits into the working technology of the staff of a research project. It allows the staff to shift from a reactive perspective that puts the burden of compliance solely on the participant to a proactive position that anticipates barriers, standardizes participant contact, and limits intensive efforts at reengaging participants to those few who do not react favorably to the routine.

Yet this is only the tip of the iceberg. Many other opportunities remain to move clinical trials toward full automation of administrative and logistical procedures. Along that path lies the ability to reduce staff tedium and focus their attention on the human contact that it takes to enhance the participation of research clients.

Computerized Tracking

It might seem too obvious in this age of microcomputers to promote the use of a computerized data base program for information storage during a clinical trial because it allows for more efficient compilation, organization, tracking, and retrieval of data.

However, many research projects still use file cabinet systems for managing participant compliance. The disadvantages of these low-tech systems are not always appreciated, particularly since the bulk of the trial's resources and attention is devoted to establishing sophisticated systems for the collection of data for the planned research analyses.

Project MATCH took a different approach. We reduced the contextual barriers to developing a participation enhancement strategy by developing specialized client data bases that allowed staff to—

- Minimize the amount of time it took to track client status.
- Easily retrieve client-specific information.
- Modify client contact tactics to achieve maximum response with minimal effort.
- Allow standardization and automation of frequent client reminder mailings.
- Continually update client information.

This data base tool allowed the research staff time to develop a routine approach to all participants as well as devote extra time to the small percentage of participants who required additional attention.

The Low-Tech Alternative

The basic difficulty with file cabinet systems is that they must be physically organized. Since the primary purpose of collecting data is to address the research questions, the file cabinet is usually organized to provide the most efficient access to the data for that purpose. However, organizing the file cabinet in a way that is efficient for data analysis might not be suitable for other important project tasks.

For example, it is often necessary to maintain copies of portions of the assessment batteries as separate information files for each ancillary purpose. This typically hampers the timely retrieval of the bits and pieces of information that are required when project staff are attempting to reengage a recalcitrant participant. They are less likely to use routine methods for enhancing participant compliance if they need to rummage through the physical records in order to retrieve information to compile assessment due dates, draw up mailing lists, or send personalized form letters. Continually updating physical files is time-consuming, confusing, and prone to error. In addition, when there are several distinct second-

ary files, it is more likely that some information will not be updated at all.

There are two typical methods of physically organizing participant data, by case and by assessment instrument. The case organization method uses participants' folders as the repository of all data collected on each client recruited into the research project. Most often the case file is marked by a participant identification (PID) number.

One major drawback of this method relates to data entry. Computerized files are generally constructed for each specific instrument, so a natural way to enter data into such programs is by assessment across individuals. This is especially true if data are entered and then reentered in a verification pass. When physical

Common Methods of Organizing Data Files

Case Organization	Participant ID	000001
	Assessment 1	
	Assessment 2	
	Assessment 3	
	Assessment *n*	
	Participant ID	000002
	Assessment 1	
	Assessment 2	
	Assessment 3	
	Assessment *n*	
	Participant ID	00000*n*
	Assessment 1	
	Assessment 2	
	Assessment 3	
	Assessment *n*	
Assessment Organization	Assessment 1	
	Participant ID	000001
	Participant ID	000002
	Participant ID	000003
	Participant ID	00000*n*
	Assessment 2	
	Participant ID	000001
	Participant ID	000002
	Participant ID	000003
	Participant ID	00000*n*
	Assessment 3	
	Participant ID	000001
	Participant ID	000002
	Participant ID	000003
	Participant ID	00000*n*

records are filed by case, the data entry task begins with retrieving a given assessment from each client's folder, entering and verifying the data, and then refiling that assessment across participants. So the case organization structure is inefficient when the data needed are assessment specific.

The second method of organizing participant data is by assessment instruments used. Retrieving information from physical files organized in this way, however, poses other problems. When data are collected and stored by instrument, it can be tedious and inefficient to find specific participant information. For instance, a research assistant preparing to do a followup interview may need to get the participant's identification number from one instrument, telephone number from another, and target date of the interview from a third.

Further, file cabinet data storage systems are often idiosyncratic, in that only those staff people who have developed or who heavily use it know how to locate a particular piece of information. Although it is possible for a research team to become familiar with a particular file cabinet system and to make it work for the project, this is not an efficient use of staff time. A complicated system that requires staff to have a lot of knowledge about the research project in order to perform relatively simple tasks, such as data entry or filing, makes it difficult to hire part-time or student help.

In addition, using paper files sometimes poses confidentiality problems. Locking up paper files is one solution to this problem, but there is a good chance that files may be left open or unattended. Also, it may be difficult to organize paper files so that some information is not available to "blind" research staff.

Thus, storing and retrieving data from a file cabinet is time consuming and increases the chance of making data transfer errors. Although most projects need to store huge quantities of data, they typically retrieve discrete bits of information at any one time. Given the widespread availability of personal computers, it becomes obvious that computerized data base storage is the better way.

The remainder of this chapter focuses on the general principles involved in constructing a data base and describes the type of data base that can be used to monitor and enhance participant compliance with the research project protocol. Because several computer platforms are available (e.g., DOS, Windows, UNIX, Macintosh), and for any given platform, several data base programs might exist, the material is presented as platform/program independent.

Determine Your Needs

Before deciding on the structure of the data base program, consider the requirements of each phase of the study, such as screening and recruitment, randomization and assignment to treatment conditions, monitoring treatment attendance, conducting and monitoring posttreatment assessments, and conducting analyses that address the trial hypotheses. Each of these phases involves data collection or manipulation. Although most of the data are collected to address some aspect of the scientific research questions, other data are used to support the process necessary to conduct the trial.

In order to preserve internal validity and conform to the expectations of rigorous scientific practice, the data collected for assessing the trial hypotheses are usually maintained in a distinct data base or as a parent data base with restricted access. This practice allows you to limit information that might influence a research staff member conducting participant assessments. It also helps ensure the confidentiality of the data collected. However, retrieval of selected bits of information from these parent files is both necessary and desirable for monitoring and enhancing client participation. Without cooperative participants, there can be no data with which to address research hypotheses.

Basic Information

The first design consideration in constructing a participant-monitoring data base is to review the data that are routinely being collected for the research project. When staff know what data are available, they can determine which bits of information will be helpful for monitoring participants.

Participant Data

The basic information needed to contact a participant is usually obtained in the initial interview: demographic and residential information such as age, sex, ethnicity, marital status, address, and telephone number along with information that allows staff to determine whether the client is suitable for the study. This type of information is also required to monitor client participation. A Note section allows staff to store anecdotal information concerning those idiosyncracies that they have found helpful in their attempts to contact the participant.

Locator Data

For each participant, the project will also need basic information about at least one locator, that is, a person named by the clients as someone with whom they have a well-established relationship who would likely know their whereabouts at all times. Participants must agree to grant the research staff permission to contact the locator; particulars of the conditions under which contact with the locator can be made are detailed in the informed consent.

Minimal Client Management Data

PID _____ Initial Interview _____ Target Interview _____

Last Name _____ Home Phone _____

First Name _____ Employer _____

Address _____ Work Phone _____

City _____ Work Title _____

State _____ Zip Code

Name of Collateral Source and Relationship _____

Soc. Sec.# _____ M F (circle) Ethnicity _____

D.O.B. _____ Yrs of Education _____ GED _____

Interests _____

Student Y N Where _____ FT/PT/NA Major _____

Locators:

Name _____ Name _____

Address _____ Address _____

City _____ State ___ Zip _____ City _____ State ___ Zip _____

Home Phone _____ Home Phone _____

Relationship _____ Relationship _____

Target interview date _____ Comments _____

Scheduled date/time _____

Site (onsite/home visit/telephone/other) _____

Date letter sent _____

Telephone confirmation due date _____

Last date called _____

Outcome of contacts _____

_____

Interviewer ID# _____ Amount paid _____

Interview complete _____ Date of last drink/drug _____

The interviewer should collect sufficient information about the locator to be able to actually contact the person who is named. Also, the information should be verified while the participant is available to make corrections. Keep in mind that life circumstances do change. It is good practice to verify the information on the locator at each contact staff has with the participant. Oftentimes, it is useful to separate the locators from the collaterals, especially when soliciting names from the participant. Someone who may serve well as a locator may know nothing about the participant's drinking (e.g., the grandma with whom the participant keeps in touch).

Collateral Data

Many trials of alcoholism treatments use collateral sources to verify the participant's self-report of drinking. Some check body fluids (blood and urine); others interview someone close to the participant who is likely to know. As with locators, participants must grant the research staff permission to contact the collateral; particulars of the conditions under which contact with the collateral can be made are detailed in the informed consent.

Upon initial contact with the collateral, the research staff should obtain that person's permission to conduct interviews; these are usually conducted over the telephone. As with locators, enough information should be collected on collaterals to actually contact them, and the collateral contact information should be verified with the participant at each contact.

In a file cabinet system, the information identified so far could be considered a hanging file folder. Within this hanging file, which could be labeled "Basics," there are now three folders: participant information, locator information, and collateral information. Thus, similar information of three distinct types is needed, and this information is collected as a routine aspect of the trial. Once it is computer readable, this information is available for monitoring participant compliance. Usually there is no need to collect the data in a separate effort.

Define the Tasks

After basic information on the participant has been collected in the screening interview, the next step is to determine what information will help keep track of where the participant is in the research process. Monitoring the participant's progress over the course of a treatment study begins with the initial contact. Often the most obvious details are the ones that are neglected but prove to be the most useful.

The obvious details here center on the participant's progress from initial contact (which is sometimes over the telephone) to screening, recruitment, baseline assessment, and randomization. Given the complexity of multitreatment service centers, the comprehen-

siveness of research assessments, and participant availability (or lack of it), it is not unusual for this initial process to take several days. Although the completion rates of screening and baseline assessments are usually very high, assuming that a participant will get through the process in a timely manner can lead to several lost applicants during the recruitment phase. It is a better practice to exercise prudence and begin the tracking process at the point of initial contact. Likewise, continue the monitoring process with each participant right through the final exit interview. Doing so will ensure that the trial has sufficient current information available to recontact a former participant should the trial receive additional funding to extend the followup period.

There are many ways to organize information that is useful for monitoring client participation. The method presented here is partitioned along the lines of tasks within the phases of the study and applies to each type of person monitored: participants, locators, and collaterals. The concept is to recognize that people other than the participant are important to track, and that the monitoring activity requires several separate tasks.

Scheduling Appointments

One identifiable task is the scheduling of participant appointments. A file can be created for each phase of the research, that is, baseline, treatment, and followup. Each file would contain four fields: participant identification number (this is the primary key variable that allows linkage to other files in the data base), scheduled date, scheduled time, and completed date.

One advantage of forecasting target dates and times for appointments is that it provides the staff and participants with tangible evidence of the clients' commitment to the project. From a research management perspective, a projected schedule of contact with each participant is an invaluable tool in organizing staff monitoring efforts. Projected schedules allow staff to assess upcoming workloads and plan their activities accordingly. For the staff, such schedules serve as prompts to send participants reminders of upcoming appointments. With a computerized data base, it is possible to automate a considerable amount of the effort involved in routine participant contact. This frees staff time for the specialized techniques designed to reengage noncompliant participants.

For example, treatment attendance is a leading indicator of subsequent participation in posttreatment assessment (Del Boca et al. 1995). Thus, a file that summarizes client participation in the treatment phase of the study is important. In the case of treatment dropouts, such information can be used to flag the need for specialized text in the letter sent to the participant prior to the initial posttreatment followup assessment. Such specialized let-

Sample Participant Schedule Files
Baseline Assessment

PID	Scheduled date	Scheduled time	Completed date
0700001	10/12/98	1:30 pm	10/12/98
0700002	10/13/98	10:00 am	10/13/98
0700003	10/16/98	9:30 am	10/16/98
0700004	10/17/98	9:30 am	10/18/98
0700005	10/18/98	2:00 pm	10/18/98

First Followup Assessment

PID	Scheduled date	Scheduled time	Completed date
0700001	12/12/98	1:30 pm	00/00/00
0700002	12/13/98	10:00 am	00/00/00
0700003	12/16/98	9:30 am	00/00/00
0700004	12/17/98	9:30 am	00/00/00
0700005	12/18/98	2:00 pm	00/00/00

Second Followup Assessment

PID	Scheduled date	Scheduled time	Completed date
0700001	12/12/98	1:30 pm	00/00/00
0700002	12/13/98	10:00 am	00/00/00
0700003	12/16/98	9:30 am	00/00/00
0700004	12/17/98	9:30 am	00/00/00
0700005	12/18/98	2:00 pm	00/00/00

Third Followup Assessment

PID	Scheduled date	Scheduled time	Completed date
0700001	12/12/98	1:30 pm	00/00/00
0700002	12/13/98	10:00 am	00/00/00
0700003	12/16/98	9:30 am	00/00/00
0700004	12/17/98	9:30 am	00/00/00
0700005	12/18/98	2:00 pm	00/00/00

Sample Treatment Attendance File

PID	Code
0700001	1
	1
0700002	3
	1
0700003	1
	1
0700004	2
	3
0700005	2
	1

ters can acknowledge participants' dropping out of treatment and inform them that it is nonetheless important to the study that they participate in the posttreatment assessments.

Clearly, a participant's attendance during treatment is just the type of information that could bias the research staff who conduct the posttreatment assessments. However, using the data base tool described in this chapter eliminates the concern that "blind" research interviewers would have access to information about the treatment history of participants. A common feature of data base programs (password lock) can restrict access to these files so the information can be hidden from inappropriate exposure to research assistants. In the Project MATCH protocol, part of the function of Project Coordinators was to monitor treatment participation and to initiate specialized letters to treatment dropouts at the appropriate time. Research interviewers could thus maintain their ignorance of a participant's treatment experience.

It is equally important to construct similar schedule files for the collaterals. In Project MATCH, contact with the collateral source did not occur as frequently as contact with the participant. This made it all the more important to project the contact schedule so that the relationship with the collateral could be maintained with timely reminders.

Reminder Letters and Telephone Calls

One method of enhancing participant compliance is the use of letter reminders. In terms of the electronic data base, two things are suggested. The first is to create in the data base a letter schedule file. Only a few fields are required, because dates on which the reminder letters should be sent can be cued from the scheduled dates in the participant (or collateral) schedule file. Client ID (collateral ID), date letter sent, and status code should be sufficient.

Second, in some data bases, the text of the letters can be entered as a file. Some data bases also have a mail merge feature that allows staff to take mailing addresses from the participant (or collateral) demographic file, put these together with a form letter, and print the mailing envelopes as well. In other data bases, it is possible to

Sample Participant Reminder for Telephone Calls File

PID	3 month		6 month		9 month		12 month		15 month	
700001	12/4/92	1	2/8/93	1	4/9/93	1	6/8/93	1	8/9/93	1
700002	12/6/92	1	2/10/93	2	4/10/93	1	6/10/93	1	8/11/93	1
700003	12/11/92	1	2/10/93	1	4/11/93	1	6/12/93	2	8/14/93	1
700004	12/15/92	1	2/15/93	1	4/16/93	1	6/17/93	1	8/18/93	1
700005	12/25/92	1	3/1/93	0	5/9/93	1	7/8/93	1	9/9/93	1

retrieve relevant information for each participant, such as name, address, and pertinent dates, but that information must be exported to a text processor in order to merge with the form letters.

A telephone call schedule file should also be created in the data base. As with the letter schedule file, only a few fields are needed. Participant ID, telephone number(s), date and time of contact, and a status code should be sufficient. Create a similar file for the collateral.

Compensation and Incentives

In treatment research, it is useful to offer participants some type of compensation for the time they spend involved in research assessments. Under current federal efforts to increase the participation of women and minority populations, it is also becoming common for research projects to offer compensation as a means of reducing barriers for these groups. For example, transportation fees and childcare arrangements might be offered to participants on a case-by-case basis.

Sample Participant Compensation File

PID	To date	Current	Childcare	Travel	Total
0700001	0	75	0	2.5	77.5
0700002	100	25	0	0	125.0
0700003	125	50	0	2.5	177.5
0700004	175	10	0	0	185.0
0700005	185	50	0	0	235.0

In addition, monetary incentives are sometimes offered to noncompliant clients to tip the balance in favor of participation. A common practice in such cases is to offer recalcitrant clients the sum total of compensation that would have been paid to them had they cooperated up to the assessment in question.

From a management perspective, having the ability to track cash outlays on a microlevel helps in monitoring project budget expenditures. It also helps ensure that participants are compensated in a timely fashion and with the correct amounts. Further, it is important to be aware of potential problems associated with preserving confidentiality when checks are used. The best solution is to have cash in hand, which can be paid immediately to participants upon completion of the appropriate session.

In a trial with rolling recruitment, where client assessments are staggered on a week-to-week basis, a participant compensation file also allows the research staff to plan for upcoming assessments by

identifying the correct amount that each participant is due at that point in the trial.

Choosing a Computer Data Base

After taking all of the foregoing into consideration, the next step is the selection of a data base program. Because of the desire to limit access to the scientific data, a participant monitoring data base will probably be constructed to operate independently. This principle also holds when the parent data already exist in a data base.

The most elemental choice to make is between programs that use either a "flat file" or a "relational file" organizational structure. Without being too technical, flat files string all fields onto one record. The result is a very long record for each participant and a large degree of duplication. The total file would soon become very large and unwieldy, even for virtual space. Further, whenever one piece of information changes, no matter how small the item, the change would be made anywhere the field appeared, perhaps a dozen places. Even minor changes could require significant time, and the risk of transcription errors would likewise increase. Data bases with flat file structures are generally to be avoided.

A relational structure allows data to be maintained in small, distinct files analogous to folders. Data that have logical similarity are grouped together. In relational data bases, each piece of information need be stored in only one place. The relational structure allows changes to be made with greater ease, since only the folder containing the change need be accessed and updated. New information is added to existing folders or new folders can be created in the data base.

A data base is like a file cabinet. Information organized by topic resides in folders, and the data base is analogous to a file cabinet drawer. When the drawer is opened, there is access to the topic folders the drawer contains. A relational data base is the best tool for creating this type of data base.

The Relational Data Base

The following describes some of the generic features that make a relational data base an effective tool for monitoring participant compliance.

Folders

The data in a folder are usually presented as rows and columns. The columns are called fields (or variables), and the rows are called records. Each record in the folder contains the same set of fields, and each field contains the same type of information. Records can be the collection of information about one participant. In order to allow records containing data about one participant across folders,

there must be a unique identifier for each participant, and this identifier must be a field in each record.

Primary Key Field

In a relational data base, the information in multiple folders is linked by assigning a unique identifier called a primary key field. The participant identification number is a good example of a primary key field. By specifying the PID as the primary key, data from all of the folders in the data base can be retrieved under a case organization. The primary key field facilitates using the information contained in the data base in a very flexible manner.

Query

A query is a question that is asked about the data in a given data base. The query function allows answers to come from records in any number of folders in the data base. In essence, the query is the means whereby pieces of data in separate folders are related to one another. The query function is common to all relational data bases. Different terms might be used, but in order to effectively use the relational structure of the data base, there must be a function that allows the user to identify and bring together information from different folders.

When a query is defined, users describe the set of records that they want. The records might be drawn from several folders. As an example, suppose that it is Friday and the staff would like to contact all participants who are due for followup assessments between Monday and Wednesday of the upcoming week. The task is to query the data base for a listing of participants whose assessments are due in that period. The needed information is identified as participant name, due date and time, date of last assessment, telephone number, best time to call, and any notations that previous interviewers might have entered into the data base. Staff also need the name, address, and telephone number of the collateral so that this information can be checked with the participant to ensure that it remains current and that no changes have occurred that would prohibit contacting the collateral.

The query function allows the user to select each of these fields. The date and time schedule fields are located in one folder, while the participant's name, telephone, best time, and notes are located in another folder. The fields for the collateral information reside in a third folder. The query function allows the user to bring all these fields together and display the fields participant by participant. Keep in mind that the query function of any relational data base is quite flexible, and the appearance of the display can be customized to a project's particular needs.

One feature to look for in a relational data base is the capacity to transmit changes made to data in the query as updates to the

source folders. For example, if staff contact the participant to confirm an assessment appointment and find that the participant no longer is on speaking terms with the collateral, the staff member can solicit a new collateral source and update the data base form. Clearly, this is a time-saving feature. Then, when it is time to conduct the next collateral assessment, the collateral is current.

There are many examples of questions that can be translated into queries. For instance, a researcher may ask: Which participants are due in March, which participants are overdue, or which participant interviews were completed in March. Reports can easily be generated for each of these requests. These queries are useful in performing the day-to-day functions of a research project, such as scheduling or contacting participants, assigning participants to staff members, and printing the data necessary for participant tracking.

When generating queries, it is important to clearly formulate questions and to know what types of information are available before attempting to translate these questions. As noted by Brunner et al. (1992), common obstacles in query formulation are "poor knowledge of the data base's constituent structure" and lack of skill in translating a general question into a correct query.

Reports produced from queries benefit the long-term planning and maintenance of the research project. For example, reports can focus on the projected workload or the distribution of the workload among staff members. Information that details characteristics of the study that are not necessary for its day-to-day or long-term maintenance can also be generated. For instance, demographic information (i.e., all participants born before 1930) may be reported.

Forms

Data base programs allow users to display data in different layouts. A form is a layout for entering, changing, and viewing records in a folder. This common feature of a data base allows the user to create a display that is best suited to the task in hand. It is often desirable, for example, to have the option of displaying data in a spreadsheet format so that information is visible and easy to access. However, forms allow users to do things not available in the default spreadsheet, such as include lists of values to choose from, display error messages for incorrectly entered data, fill in data, display check-off boxes, and show the results of calculations.

Reports

The report function of a data base is the means for printing information selected from records in a customized layout. Like a form, a report allows the user to manipulate records in a number of different ways without altering organizational structure of the

folders. Generally, reports allow users to display data from fields, the results of calculations, graphs, pictures, or even other forms or reports.

Passwords

The password feature makes it easy to protect participant confidentiality. Research staff often have a password assigned that allows access to a specific level of information in the data base. In addition, fields can be created so that all information entered in the field is invisible unless the field is specifically activated.

Note Fields

Note fields are special fields within folders that also allow users to hide confidential (Social Security Numbers) or sensitive (AIDS test results) information from the view of casual onlookers who are not permitted access. It also hides specific information in a folder from those who have access to other, less sensitive information in that folder.

Mail Merge

A data base should also permit users to retrieve information for the purpose of printing letters or postcards. By merging text from the data base file into a previously written document, letters or postcards can be generated easily and quickly. Letters can be written for a variety of compliance enhancement tasks, such as confirming assessment appointments, contacting participants who fail to attend scheduled appointments, and contacting participants who cannot be reached by telephone. Personalized participant letters can be tailored to specific functions. The use of personalized correspondence has been demonstrated to be more effective than standardized correspondence in stimulating compliance to an assessment schedule (Curry et al. 1993).

Updating

The use of the form function should simplify the task of updating the data in folders. That leaves the scheduling and execution of timely updates to consider, and these are administrative matters. However, sometimes a research project will have multiple recruitment sites and maintain a separate data base for each site. In this case, consideration must be given to whether the data base allows concatenation of two or more data bases to form a master data base.

Other Features

It is helpful if the data base allows deleted fields to remain in the folder unless the folder is reorganized to remove them. Fields that have been deleted sometimes later prove useful, and having a way to reinstate the field is often of great benefit.

It is important to select a data base that allows fields to be continually modified without reentering data. For example, an address field of 25 characters might be created, but after entering several records, it is discovered that some participant addresses are longer than 25 characters. The data base program should make it simple to alter the field without having to reenter data. Similarly, the data base should allow for alteration in field type from numeric to character.