

Managing data for a multicountry longitudinal study: Experience from the WHO Multicentre Growth Reference Study

Adelheid W. Onyango, Alain J. Pinol, and Mercedes de Onis, for the WHO Multicentre Growth Reference Study Group

Abstract

The World Health Organization (WHO) Multicentre Growth Reference (MGRS) data management protocol was designed to create and manage a large data bank of information collected from multiple sites over a period of several years. Data collection and processing instruments were prepared centrally and used in a standardized fashion across sites. The data management system contained internal validation features for timely detection of data errors, and its standard operating procedures stipulated a method of master file updating and correction that maintained a clear trail for data auditing purposes. Each site was responsible for collecting, entering, verifying, and validating data, and for creating site-level master files. Data from the sites were sent to the MGRS Coordinating Centre every month for master file consolidation and more extensive quality control checking. All errors identified at the Coordinating Centre were communicated to the site for correction at source. The protocol imposed transparency on the sites' data management activities but also ensured access to technical help with operation and maintenance of the system. Through the rigorous implementation of what has been a highly demanding protocol, the MGRS has accumulated a large body of very high-quality data.

Key words: Data collection, data processing, database management system, longitudinal study

Introduction

The World Health Organization (WHO) Multicentre Growth Reference Study (MGRS) data management system was set up and operated according to a protocol designed to ensure a high quality of banked data, stored securely against unauthorized manipulation and accidental loss. The data were collected over a period of more than six years (July 1997 to November 2003) and in six different sites with variable levels of data management experience. In this context, using a standardized protocol in all study sites simplified the compilation and maintenance of the central master files at the MGRS Coordinating Centre as well as facilitating Coordinating Centre-to-site and intersite technical support whenever required. The system also imposed transparency to the extent that the Coordinating Centre could replicate and extend key elements of the quality control procedures that sites were expected to carry out as part of the data collection and management protocols. The confidentiality of the study participants was ensured by limiting identification information in the study data files to numbers without names or other information that might identify them beyond the purposes of the study.

The purpose of the present article is to share experience gained in managing the large body of data collected in the MGRS. We describe the data management model, the standard operating procedures (SOPs) used for handling forms and data, the computerized system with its inbuilt quality assurance features, the respective responsibilities of the sites and the Coordinating Centre, data quality checking and cleaning during the data collection phase, and the closure of data management activities in the sites.

General organization of the data management system

The longitudinal and cross-sectional components of the MGRS are described elsewhere in this supplement

The authors are affiliated with the Department of Nutrition, World Health Organization, Geneva.

Please direct queries to: Mercedes de Onis, Study Coordinator, Department of Nutrition, World Health Organization, 20 Avenue Appia, 1211 Geneva 27, Switzerland. Telephone: 41-22-791 3320; fax: 41-22-791 4156; e-mail: deonism@who.int.

Members of the WHO Multicentre Growth Reference Study Group and Acknowledgments are listed at the end of the first paper in this supplement (pp. S13–S14).

[1], where detailed information on the specific data collected from the sample is also provided. In the data management environment, the longitudinal and cross-sectional study components were treated as separate projects with respect to assembling and processing batches of data and creating master files.

Briefly, the longitudinal study data set consists of eight master files, the first of these being the file that describes all screened subjects, regardless of whether or not they were enrolled in the study. Other questionnaires recorded information on the initiation of breastfeeding in the hospital and its continuation at home; baseline demographic and parental characteristics; child feeding, morbidity, and anthropometry during follow-up; and motor development. All enrolled subjects had an end-of-participation form completed indicating when they ended participation and for what reason. The eighth longitudinal study master file contains data from the 12-month study involving refusals and early dropouts who agreed to respond to an interview on the child's first birthday. The cross-sectional data set comprises two essential files: a screening master file with records of all screened subjects, and a survey master file with records of all subjects who responded to the cross-sectional study interview and had their anthropometric measurements taken. One supplemental form was used in Brazil and the United States,

where the mixed-longitudinal design was used [1–3]. In these two sites, some cross-sectional study participants received one or two follow-up visits at which an abbreviated version of the survey questionnaire was used to collect data on anthropometry and intercurrent morbidity. A summary of the types of forms and number of records accumulated by each site up to end of May 2003 is presented in table 1 for the longitudinal study and table 2 for the cross-sectional study.

Preparatory work and system setup

A decentralized data management model was chosen for the study: each site collected, entered, verified, and validated data, and then locally created, updated, and cleaned study master files. Copies of the data files were transferred every month to the Coordinating Centre, where the consolidated study master files were created and updated with incoming data from the sites. Figure 1 illustrates the data flow and summarizes the tasks undertaken by the sites and the Coordinating Centre.

In order for this organizational system to work, the sites followed a common data management protocol, which included the use of centrally prepared data collection forms (questionnaires) and the same data processing system (software and dictionaries). The

TABLE 1. Longitudinal study forms received by May 2003

Form	Brazil (<i>n</i> = 4,801)	Ghana (<i>n</i> = 2,057) ^a	India (<i>n</i> = 692) ^a	Norway (<i>n</i> = 836)	Oman (<i>n</i> = 4,957)	USA (<i>n</i> = 398)	All countries (<i>n</i> = 13,741)
Screening	4,801	538	433	836	4,957	398	11,963
Breastfeeding (hospital)	—	343	353	322	446	237	1,701
Breastfeeding (home)	—	1,241	1,254	1,188	1,221	834	5,738
Baseline	368	351	331	308	328	212	1,898
Follow-up	5,864	6,090	5,435	5,605	5,488	3,843	32,325
12-month visit	101	12	60	41	72	28	314
Motor development	—	2,651	2,623	2,209	2,436	1,726	11,645
End of participation	388	364	234	322	450	232	1,990
All forms	11,522	11,590	10,723	10,831	15,398	7,510	67,574

a. Ghana and India prescreened subjects owing to local circumstances [4, 5] before the MGRS screening interview was administered, hence the difference between the number of subjects and the number of screening forms in these 2 sites.

TABLE 2. Cross-sectional study forms received by May 2003

Form	Brazil (<i>n</i> = 2,292)	Ghana (<i>n</i> = 4,622)	India (<i>n</i> = 3,886)	Norway (<i>n</i> = 5,185)	Oman (<i>n</i> = 4,509)	USA (<i>n</i> = 919)	All countries (<i>n</i> = 21,413)
Screening	2,292	4,622	3,886	5,185	4,509	919	21,413
Cross-sectional survey	487	1,323	1,490	1,387	1,432	562	6,681
Follow-up survey I	450	—	—	—	—	422	872
Follow-up survey II	419	—	—	—	—	364	783
All forms	3,648	5,945	5,376	6,572	5,941	2,267	29,749

electronic dictionaries in the data management system exactly matched the questionnaires and interviewer guides. For example, the interviewer guide specified when contingency questions were to be skipped, and the data entry dictionary had a corresponding rule to skip the variable during data entry in order to reduce unnecessary key punching. To facilitate data entry further, electronic forms were formatted so that the data entry screens matched the pages of each respective data collection form. In sites where questionnaires required translation from English (Brazil, Norway, and Oman), they were translated into the local language and independently back-translated into English to ensure that the content of the questions remained unchanged. The interviewer guides were also translated in these sites.

Before the start of data collection, the data manager of each site participated in a week-long training workshop at the Coordinating Centre in Geneva. The workshop included a presentation of the WHO Good Clinical Practice and data management principles [6], and the DMS/2 data processing software for data entry, verification, validation, and file update. Exercises and dummy runs were organized to ensure that participants clearly understood the SOPs in data management and why it was important to implement them. Some time was devoted to discussing and defining the responsibilities of the sites and the Coordinating Centre with respect to data monitoring, transferring study data, and obtaining help whenever required. Before data collection began, each data manager was given a timetable

with exact dates for monthly data submission to the Coordinating Centre for the duration of the data collection phase.

The complete system installation package was distributed to each data manager. This included the full set of dictionaries for all study questionnaires and the DMS/2 software with documentation for its operation. The dictionary for each form defines its data variables and types, labels, plausible value ranges, and data entry skip-and-fill rules, as well as intervariable cross-checks. The data managers were involved in interviewer training before data collection began in the site to stress the importance of completing the questionnaire forms legibly and according to the instructions in the interviewer guides.

Standard operating procedures at site level

The data management procedures are simple and follow a natural sequence. The data manager had to ensure that each step was successfully completed before moving to the next. Because of the repetitiveness of these operations during data collection, it is easy to miss a problem, therefore the need for rigorous application of the procedures was emphasized during training.

Each form received from interviewers was manually checked for legibility, completeness, and consistency, and any additional coding was done at this stage. All

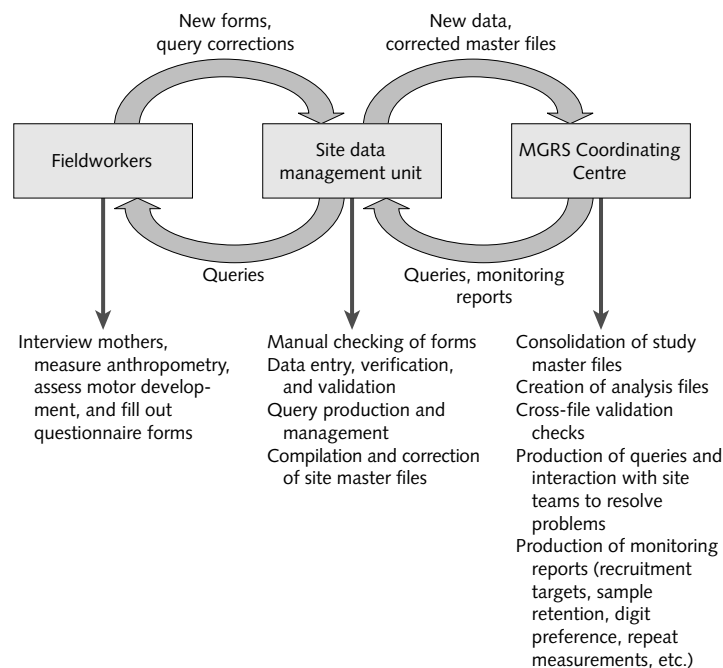


FIG. 1. Data management standard operating procedures and data flow at sites and at the Coordinating Centre

forms received were recorded in the subject form register, a manual or electronic spreadsheet that indicated completed visits and when they had been done for each subject. The subject form register facilitated timely detection and correction of duplicated forms and errors in subject identification, and helped to keep track of missed visits and losses to follow-up. The data managers periodically printed out a computer form register, a replica of the subject form register showing which forms had been accumulated in the master files for each subject. This was a useful double-checking tool for identifying any forms that might have been misplaced between reception at the study center and the data entry unit.

The forms received over one or two days were assembled into a batch that was assigned an identification number. An entry for the batch was added to the batch log register. A cover form was attached to each batch to indicate its source and contents (number and types of form and subject identification numbers). This form provided spaces for recording the dates when the batch was received, entered, verified, validated, and updated to master files. All forms in a batch remained under the same cover until they had been processed through to master file updating. With the batch cover form in order, the batch was sent for data entry. The SOPs specified that data be entered twice, preferably by different data entry operators. For the first entry, the operator activated the relevant electronic dictionaries by indicating which forms were included in the batch. The operators were trained and required to key in data exactly as they appeared on the form. The second entry was done on the same or following day. For this run, the system was set to verification mode, that is, the original entries were hidden and the operator keyed the same data over them. Whenever there was a discrepancy between the original and the verification entry, the system stopped and the operator was required to verify the correct information from the form and enter it.

The next step was to validate the data. This was done on the basis of the range and consistency rules built into the data entry dictionary. The validation procedure created a query file, from which query sheets were printed. Each query was first checked against the data form, and if it was not a data entry error, i.e., the flagged data were as recorded on the data form, it was sent to the interviewer for investigation. It was necessary in some instances to revisit the respondents to obtain correct information. When data were confirmed to be correct despite the queries, the interviewer indicated this and no correction was made. When corrections were necessary, they were recorded on the data form and the query sheet. The query sheets with corrections were handed back to the data management unit, where the data manager created a correction batch to update the master file.

All master file updates and corrections were carried

out using transaction batch files and correction files, respectively. The procedure for updating master files included a compulsory step in which backup copies of the old master files and the transaction files were saved. The output report was checked after each update to assess whether the procedure had been successfully completed, and if there were any problems, such as duplicate records in the new master file, or if some records in the batch had been rejected in the updating process. The latter occurred if a record with duplicate identification had already been saved in the master file. To correct data errors and delete duplicate or faulty records, correction statements were processed against the master file with the same backup requirements and output listings, as described for adding new records to the master file. Interactive correction of the master files was not permitted, which, together with the careful documentation of queries and corrections, helped to maintain a clear data audit trail. Moreover, in the event of a computer crash, the master files could be recreated by rerunning the transactions and updates in their right order.

The batches were dismantled once the data forms had been processed through to master file updating. At this point, the processing history from the batch cover form was copied into the batch log register, the batch cover form was filed away, and the data forms were stored in the individual subject folders kept in the archiving unit for each study participant.

Standard operating procedures at the Coordinating Centre

The Coordinating Centre had the same software and the same electronic questionnaire dictionaries as the sites, making it possible to replicate some of the site procedures. For the first six months of data management, sites sent their monthly returns in the form of transaction files. The Coordinating Centre replicated the validation and update of these files to evaluate each site's compliance with the SOPs. After the initial period of six months, only master files were transferred to the Coordinating Centre. A log was kept of all data received, and master file update listings were used to double-check that the Coordinating Centre had exact copies of the master files kept in each respective site.

The different master files were combined at the Coordinating Centre to create an analysis file (separate for longitudinal and cross-sectional studies) in which each subject had a single record with aggregate data from separate questionnaires. Data from repeating forms (e.g., the 20 follow-up forms) were reorganized to create only one record per subject instead of having a record for each follow-up visit. The data were thus arranged in suitable format and structure for analyses using standard statistical software programs. This

process also permitted further validation checks for consistency among data originating from different master files, e.g., measurements changing abnormally relative to the chronology of follow-up visits, or visit dates that were inconsistent with the sequence of visits. Derived variables were created from existing variables; for example, in the longitudinal study a feeding compliance indicator was derived from data on breastfeeding and complementary food intake recorded at different follow-up visits.

Descriptive statistics and data plots were also routinely studied to identify data problems. Queries about inconsistent and dubious data were fed back to the site to investigate and implement any required corrections. As with locally generated queries, the interviewers returned to the forms and sometimes to the respondents to verify queried data. Documentation of these queries and the responses to them were kept on file at both the site and the Coordinating Centre. When queries could not be adequately resolved through e-mail correspondence, they were reviewed on site during monitoring visits that were undertaken annually by a member of the Coordinating Centre team.

In addition to quality control checking and interacting with sites to resolve data problems, the Coordinating Centre also produced reports that were used to monitor sample recruitment and retention, and, in the longitudinal study, compliance with MGRS feeding recommendations and smoking restrictions. Detailed monitoring reports were produced periodically to inform the Executive Committee and the Steering Committee of the progress of the study.

During the data collection phase, summary statistics were produced to evaluate each interviewer's digit preference in anthropometric measurements. Significant digit preferences were studied to determine if they might lead to overall biased measurements and were communicated to the site for discussion with the relevant interviewers. The frequency of measurements that had been repeated because the maximum allowable differences between observer pairs had been exceeded was also monitored to assess adherence to the Measurement and Standardization Protocols of the MGRS [7].

Closure of data management activities

After data collection was completed in a given site, a period of about six months was dedicated to in-depth data quality checking and master file cleaning. The Coordinating Centre produced detailed validation reports, descriptive statistics, and plots from the site's master files. For the longitudinal study, each anthropometric measurement was plotted for each individual child from birth to the end of his or her participation. These plots were examined individually for any questionable patterns. Query lists from these

analyses were sent to the site for investigation and correction or confirmation as required. As with the data collection process, the site data manager prepared correction batches to update the master files. The updated master files were then sent to the Coordinating Centre, and this iterative quality assurance process continued until the site and the Coordinating Centre were satisfied that all identifiable problems had been detected and corrected.

At this point, a team from the Coordinating Centre carried out a data management closure visit with the following objectives: to clean up any outstanding data problems and document those that could not be resolved; to certify the site's adherence to the data management SOPs; and to produce the final site data set, list closure analyses, and archive all study materials.

Any pending data errors were corrected during the visit, and those that could not be resolved were documented, such as observations flagged as out of the probable range but confirmed to be correct. Clerical procedures for handling data collection forms as well as computer procedures for handling the data in the electronic files were reviewed and documented. Finally, a set of descriptive analyses was run on the final data set, and the results were reviewed with the site team. An inventory of all study materials was made, and the location of their storage and their retention period were discussed with the site team and documented. The final site master files were archived, and copies of the same were sent to the Coordinating Centre for inclusion in the MGRS master data set.

Closure of data management activities meant that master files were henceforward frozen and therefore not to be changed by either the site or the Coordinating Centre. Any problems identified thereafter could only be dealt with and documented at the analysis stage. The final master file copies and other study documentation were kept in read-only format at the Coordinating Centre with CD-ROM backups of the same.

Discussion

Among the criteria applied in selecting sites for the MGRS were the existence of local expertise and the capacity to implement the study. The need to have personnel with adequate skills and computing facilities for data management was integral to this criterion. In addition, the data manager from each site received specific training in implementing the MGRS data management protocol and using the centrally prepared computing system. Each site had at least two computers dedicated exclusively to data management. The staff involved had variable data management experience, but since all sites used a standard package, those that experienced problems received technical support from the Coordinating Centre or the data managers from other sites whenever

required. For example, the first data manager in Ghana left before the study began, and her replacement was trained on-site by the Norwegian data manager.

The setup and operation of the system were designed to ensure the accumulation of high-quality data and to secure them against unauthorized manipulation and accidental loss. We chose to decentralize data handling rather than use a centralized model in which all data would have been sent to the Coordinating Centre for entry, verification, validation, and creation of the primary master files [8]. The chosen organizational model had the advantage of fostering capacity building in the sites and provided a framework for intersite technical support. The decentralized system also kept the questionnaire forms close to the data sources, which minimized the risk that data would be damaged or lost and made the process of verifying queried data efficient, especially when it was necessary to revisit the respondents. Few problems were experienced over the years in transferring data through the Internet, and these were minor, easily resolved ones, such as corrupted files. The sites kept their reporting schedules throughout data collection, which facilitated the Coordinating Centre's task of ensuring that the central master files were up-to-date with the site master files. It also helped the timely detection of problems that could only be revealed when data from separate master files were combined in the analysis file, so sites could be alerted to investigate them within weeks of the initial data entry.

Data management in the longitudinal study in Brazil proceeded differently from the process described in this article, because the site began data collection well ahead of the others (in July 1997) and served to pilot test the MGRS protocol and questionnaires [1, 2]. This head start also explains why Brazil did not have forms for breastfeeding and motor development data (table 1), as decisions to collect and record these data were taken after this site had initiated data collection [1]. The first data management workshop was conducted in Geneva

in November 1998, by which time Brazil was in the second year of the longitudinal follow-up. The data from that site had therefore to be converted from Epi Info [9] to the DMS/2 system for incorporation into the MGRS master files. The conversion process was achieved with the Coordinating Centre's assistance over the Internet and a site visit to Brazil by the Norwegian data manager. Once the data files were converted to DMS/2, they were subjected to the same in-depth validation and quality checking that was standard for the other MGRS sites. The cross-sectional study data were collected using the centrally prepared questionnaires and computer processed using the standard package.

The inbuilt range and consistency checks of the computerized system, as well as the ongoing data monitoring routines at the Coordinating Centre, were highly effective in detecting data errors, and since data cleaning kept pace with data collection and entry, most problems were detected and corrected soon after the data had been computerized at the site or received at the Coordinating Centre. The emphasis on keeping a clear audit trail in all data handling also helped in identifying sources of problems and taking appropriate measures to strengthen the quality assurance system.

The multiple tiers of data checking steps may have been too labor-intensive for some sites, but where routine might lead to errors being overlooked, the possibility of their detection was provided by the next checking level. Individual anthropometry plots were checked at the end of data collection in each site, and hence a few data errors were detected long after the data were collected. These very few obviously erroneous measurements that could not be corrected were excluded from the analysis file. Overall, however, the site and Coordinating Centre data management teams implemented the data management protocol with a high degree of rigor, and the MGRS data set is of very high quality.

References

1. de Onis M, Garza C, Victora CG, Onyango AW, Frongillo EA, Martinez J, for the WHO Multicentre Growth Reference Study Group. The WHO Multicentre Growth Reference Study: planning, study design, and methodology. *Food Nutr Bull* 2004;25(1)(suppl 1):S15–26.
2. Araujo CL, Albernaz E, Tomasi E, Victora CG, for the WHO Multicentre Growth Reference Study Group. Implementation of the WHO Multicentre Growth Reference Study in Brazil. *Food Nutr Bull* 2004;25(1)(suppl 1):S53–9.
3. Dewey KG, Cohen RJ, Nommsen-Rivers LA, Heinig MJ, for the WHO Multicentre Growth Reference Study Group. Implementation of the WHO Multicentre Growth Reference Study in the United States. *Food Nutr Bull* 2004;25(1)(suppl 1):S84–9.
4. Lartey A, Owusu WB, Sagoe-Moses I, Gomez V, Sagoe-Moses C, for the WHO Multicentre Growth Reference Study Group. Implementation of the WHO Multicentre Growth Reference Study in Ghana. *Food Nutr Bull* 2004;25(1)(suppl 1):S60–5.
5. Bhandari N, Taneja S, Rongsen T, Chetia J, Sharma P, Bahl R, Kashyap DK, Bhan MK, for the WHO Multicentre Growth Reference Study Group. Implementation of the WHO Multicentre Growth Reference Study in India. *Food Nutr Bull* 2004;25(1)(suppl 1):S66–71.
6. World Health Organization. Guidelines for good clinical practice (GCP) for trials on pharmaceutical products. Geneva: World Health Organization, 1995.
7. de Onis M, Onyango AW, Van den Broeck J, Chumlea WC, Martorell R, for the WHO Multicentre Growth Ref-

- erence Study Group. Measurement and standardization protocols for anthropometry used in the construction of a new international growth reference. *Food Nutr Bull* 2004;25(1)(suppl 1):S27–36.
8. Pinol A, Bergel E, Chaisiri K, Diaz E, Gandeh M. Managing data for a randomised controlled clinical trial: experience from the WHO Antenatal Care Trial. *Paediatr Perinat Epidemiol* 1998;12(suppl 2):142–55.
 9. Dean AG, Dean JA, Coulombier D, Brendel KA, Smith DC, Burton AH, Dicker RC, Sullivan K, Fagan RF, Arner TG. Epi Info, version 6.04a, a word processing, database, and statistics program for public health on IBM-compatible microcomputers. Atlanta, Ga, USA: Centers for Disease Control and Prevention, July 1996.