Erschienen in: Mechanisms of Ageing and Development ; 151 (2015). - S. 26-30
https://dx.doi.org/10.1016/j.mad.2015.03.005

Mechanisms of Ageing and Development 151 (2015) 26–30

Contents lists available at ScienceDirect

Mechanisms of Ageing and Development

journal homepage: www.elsevier.com/locate/mechagedev

The MARK-AGE phenotypic database: Structure and strategy

María Moreno-Villanueva[a,*,1], Tobias Kötter[b,1], Thilo Sindlinger[a], Jennifer Baur[a], Sebastian Oehlke[a], Alexander Bürkle[a], Michael R. Berthold[b]

[a] Molecular Toxicology Group, Department of Biology, University of Konstanz, 78457 Konstanz, Germany
[b] Chair for Bioinformatics and Information Mining, University of Konstanz, Konstanz, Germany

ARTICLE INFO

Article history:
Available online 27 March 2015

Keywords:
Database
Data management

ABSTRACT

In the context of the MARK-AGE study, anthropometric, clinical and social data as well as samples of venous blood, buccal mucosal cells and urine were systematically collected from 3337 volunteers. Information from about 500 standardised questions and about 500 analysed biomarkers needed to be documented per individual. On the one hand handling with such a vast amount of data necessitates the use of appropriate informatics tools and the establishment of a database. On the other hand personal information on subjects obtained as a result of such studies has, of course, to be kept confidential, and therefore the investigators must ensure that the subjects' anonymity will be maintained. Such secrecy obligation implies a well-designed and secure system for data storage. In order to fulfil the demands of the MARK-AGE study we established a phenotypic database for storing information on the study subjects by using a doubly coded system.

1. Introduction

In general a database serves as a data storage device that allows data processes and analysis. Therefore, to design a detailed data model of a database can be beneficial (Teorey et al., 2009). A database, which retains demographic, medical and bioanalytical data obtained from clinical and/or observational studies, is an important source of valuable information for further research. Therefore results of human studies should be made machine-readable for data analysis and data mining and become available to other scientists, thus promoting transparency. Relational databases are a common method for storing repetitive data. A practical explanation of relational database design has been reported before (Wesley, 2000).

MARK-AGE was a population study that comprised 3337 subjects and was conducted to identify a set of biomarkers of ageing, which would measure biological age better than any marker in isolation. Four groups of subjects were recruited, i.e., (1) randomly recruited age-stratified individuals from the general population [RASIG], (2) subjects born from a long-living parent belonging to a family with long living sibling(s) already recruited in the framework of the EC-funded "Genetics of Healthy Ageing (GEHA) project. For genetic reasons such individuals ("GEHA offspring") are expected to age at a slower rate [GO], (3) spouses of GO [SGO], and (4) a small number of patients with progeroid syndromes (see Bürkle and co-workers, this issue). From all subjects enrolled, anthropometric, clinical and social data were collected in a standardised fashion by using questionnaires asking for demographic information (family composition, marital status, education, occupation, and housing conditions), lifestyle (food habits, use of tobacco or alcohol, daily activities), functional status (activities of daily living), cognitive status (STROOP test, 15-picture learning test), health status (present and past diseases, self-perceived health, number and type of prescribed drugs) and mood (ZUNG depression scale). Information of body mass index, waist and hip circumference, blood pressure, heart rate, lung capacity, near vision, chair standing test and handgrip strength were also documented. Additionally MARK-AGE subjects were asked to donate blood after overnight fasting. A part of whole blood was processed to obtain plasma, serum and peripheral blood mononuclear cells (PBMC), another part was sent for blood counts. Buccal mucosal cells and spot urine samples were also collected. Several hundred potential biomarkers of ageing targeting different cellular functions have been measured in MARK-AGE biological material (Table 1). In biological and medical terms the MARK-AGE database contains measurements of classical clinical chemistry parameters and biomarkers giving information about immune system, DNA damage in the nuclear genome, accumulation of some covalently modified proteins and accumulation of oxidative damage in macromolecules. All these have been considered as causative of cellular ageing.

* Corresponding author. Tel.: +49 7531 884414; fax: +49 7531 884033.
E-mail address: maria.moreno-villanueva@uni-konstanz.de (M. Moreno-Villanueva).
[1] These authors contributed equally to this work.

http://dx.doi.org/10.1016/j.mad.2015.03.005
0047-6374/

Table 1
MARK-AGE metatable containing the necessary information for interpreting and analysing data.

Parameter description
Total of debris particles per millilitre
Number of viable cells per millilitre
% Viable cells
Anti poly(ADP-ribose) antibody stimulated PARP fluorescence intensity
Anti poly(ADP-ribose) antibody basal fluorescence intensity
Number of cells counted for FADU analysis
Poly(ADP-ribose)polymerase
Initial DNA integrity
% DNA repair
DNA integrity after 3.8 Gy
Normed concentration of carotine
DNA amount in sample as reference for norming concentrations
Normed concentration of gluthatione
Normed concentration of vitamin C
Normed concentration of vitamin E
Total analysed spots
Telomere length
% Telomeres shorter than 3 Kb
IgG antibodies specific for influenza A
IgG antibodies specific for influenza B
IgG antibodies specific for measles
IgG antibodies specific for tetanus toxoid
Number of T cells producing INF-g after stimulation with influenza A
Number of T cells producing INF-g after stimulation with influenza B
Number of T cells producing INF-g after stimulation with measles
Number of T cells producing INF-g after stimulation with tetanus toxoid
Number of T cells producing INF-g after stimulation with CMV
Plasma copper
Copper to zinc ratio
Plasma copper eluting with retention time of ceruloplasmin (% of total copper peaks)
Plasma copper eluting with retention time of ceruloplasmin (absolute value in ppb)
Plasma copper not eluting with retention time of ceruloplasmin (% of total copper peaks)
Plasma copper eluting with retention time of ceruloplasmin (absolute value in ppb)
Plasma iron
Plasma iron eluting with retention time of albumin (% of total iron peaks)
Plasma iron eluting with retention time of albumin (absolute value in ppb)
Plasma iron eluting with retention time of transferrin (% of total iron peaks)
Plasma iron eluting with retention time of transferrin (absolute value in ppb)
Background fluorescence (MFI) of secondary antibody in lymphocytes induced by 50 uM Zn
Background fluorescence (MFI) of secondary antibody in monocytes induced by 50 uM Zn
Background fluorescence (MFI) of secondary antibody in PBMCs induced by 50 uM Zn
Metallothionein (MFI) in lymphocytes induced by 50 uM Zn
Metallothionein induction in lymphocytes normalized for blank signal from MFI data
Metallothionein (MFI) in monocytes induced by 50 uM Zn
Metallothionein induction in monocytes normalized for blank signal from MFI data
Metallothionein (MFI) in PBMCs induced by 50 uM Zn
Metallothionein induction in PBMC normalized for blank signal from MFI data
Metallothioneins (MFI) in lymphocytes induced by 24 h treatment with Zn 50 uM without subtraction of background
Metallothioneins (MFI) in monocytes induced by 24 h treatment with Zn 50 uM without subtraction of background

In order to maximise the value of the data collected within MARK-AGE we have established a comprehensive database, which facilitates data storage, sharing and analysing. The MARK-AGE database allows for access and use data by Consortium researches and later on by also external scientists.

2. Material and methods

The MARK-AGE database was established according to European standards.

2.1. Volunteer privacy and confidentiality

Individual subject medical information obtained as a result of this study is considered confidential and any disclosure to third parties is prohibited. On all documents subjects must be identified only by a numerical code, and never by their names. Technicians in charge of processing the samples did not have access to the central database located in another city. By implementing this procedure the privacy and confidentiality of the uploaded data is guaranteed.

2.2. SQL-database

The Structured Query Language (SQL) is a powerful tool for interacting with relational database systems. SQL enables users to perform complicated data analysis using simple syntax and structure (Jamison, 2003). The database consists of several management, meta and data tables. The data tables represent the various sections of the questionnaires as well as the uploaded analytical results. The management table stores the user information as well as an indication which data have been entered for which subject. The meta tables contain all additional information of the bioanalytical parameters e.g., short and long name of a given parameter, description and membership to the Work Packages and MARK-AGE consortium members (Supplementary information).

2.3. Hardware and software

Selected hardware and software components (see below) were suitable to achieve an electronic data processing system that meets the demands of security in data transfer and storage, usability for Consortium partners and high availability and reliability of services.

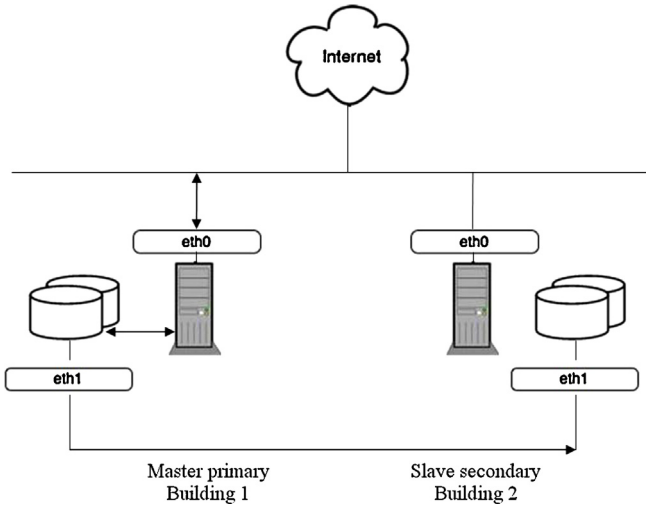

Fig. 1. Database system. The database system is a high availability (HA) cluster. The HA cluster consists of two running servers. At each point in time only one server acts as the master. In this figure, the server in building 1 is the master that supplies all necessary services for data processing and storage via eth0 LAN adapter to and from the internet. All storage operations to the database are mirrored via eth1 LAN adapter and a point-to-point connection to the inactive server (slave). In case of service failure, master and slave will swap roles to restore availability of service.

2.4. Hardware description

As the central hardware component the PRIMERGY TX150 S6 Server (Fujitsu–Siemens) was chosen. This is a middle-sized reliable hardware platform with a pair of internal mirrored hard dick drives (raid 1). Two such servers are used as a high availability cluster (Fig. 1). In a high availability cluster each server surveys the availability of services like web server or file storage on the other servers. If the active server fails to supply a service another server automatically takes over the active part and carries on to supply all services without delay. We selected two locations in two distant buildings on the campus of the University of Konstanz with sufficient network connectors in order to avoid running the servers in the same place. In cooperation with the University Computing Centre a dedicated LAN point-to-point connection was installed based on an existing fibre-optic cable for data exchange between the two servers in the two buildings. So even if one location were damaged severely, e.g., by fire, the second server would take over.

2.5. Software description

In order to guarantee excellent usability for the MARK-AGE Consortium members in transferring data, it was decided to use a standard web interface technique consisting of a web server and a scripting language on the server side. Therefore, a partner only needed a PC with internet connection and a web-browser for data input and transfer. The servers were delivered with the operating system Debian Linux and standard packages. The following was installed or configured on both cluster servers:

- A distributed replicated block device (DRBD) system (to get the storage operations mirrored).
- Via the dedicated LAN point-to-point connection).
- The heartbeat system (to get the server-to-server services survey).
- The relational database system PostgreSQL (as the storage layer).
- Apache Web server system (as the input layer).
- Modules of server side scripting language PHP (as the processing layer).

3. MARK-AGE database features

3.1. Data coding system

The major task for the information systems within the MARK-AGE project was to establish services for entry, storage and retrieval the phenotypic data from 3337 subjects. A general requirement in the project is to separate three types of data, i.e., subject identifying data, biographical data and bioanalytical data. To achieve this it was decided to connect the first and latter two types only by identifiers but store them separately. The identifiers we called subject codes (SC). Fig. 2 provides a summary of SC and data flow.

Personal data and results from bioanalytical measurements were entered into the central database only using a subject code.

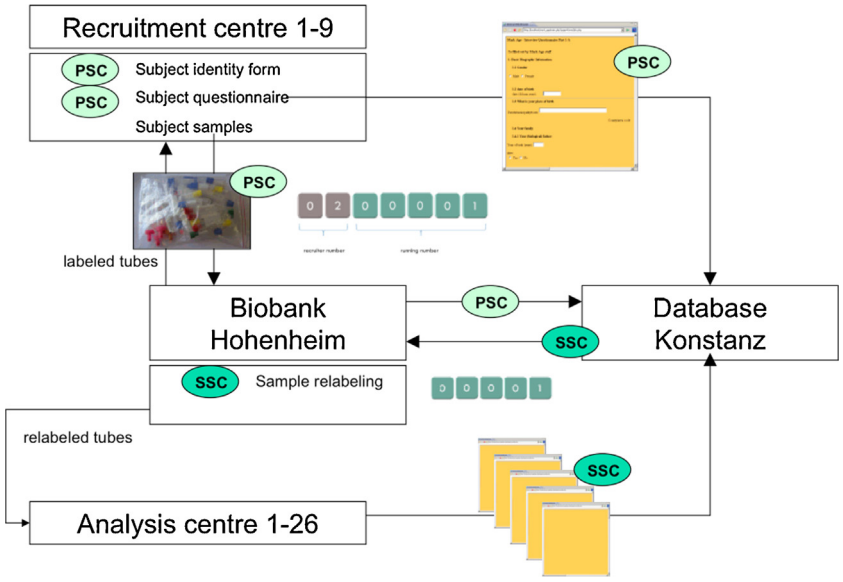

Fig. 2. Data flow and subject codes within MARK-AGE. Recruitment centres generated unique primary subject codes (PSC) for coding of both, electronic questionnaire information and biological material. Biobank staff introduced the PSC into a database system service, which converted it in a randomly generated secondary subject code (SSC). Biobank staff then re-labelled any tubes containing biological material using the SSC. Bioanalytical laboratories uploaded their data by using the SSC.

Coding of recruited individuals was performed directly at each recruitment centre by assigning a unique alphanumeric code (primary subject code, PSC) to each subject. The PSC consistsed of 7 digits; the first and the second digits identified the recruitment centre; the third and fourth identified recruitment phases and the last three digits are running numbers with exception of "TRY-Phase" coding (see below). Biological samples were re-coded at the MARK-AGE Biobank, by assigning a secondary subject code (SSC). This code was generated automatically at the Biobank and was passed on to the Coordinating Centre only. The SSC consisted of 5 digits. The first 4 digits were generated in a random but unique manner; the last digit was a checksum. Only SSC-coded biological material was distributed to MARK-AGE members for bioanalytical measurements. Subject-related data including biographical data were entered into the database by the recruitment centres whereas bioanalytical data were entered only using a secondary code; both types of data were connected by the PSC to SSC coding table.

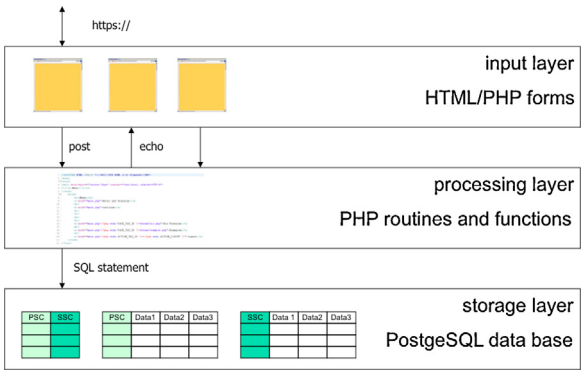

Fig. 3. Data input was based on three software layers on the MARK-AGE cluster server. The database system was contacted from the internet by a web browser via the secure hypertext transfer protocol (https). After filling out a HTML form and clicking the send button, the data were passed to the processing layer by the post command. Then PHP scripts could inspect data and allow or reject input if values were beyond thresholds. If allowed, the data were inserted into the database by SQL statements from the processing layer.

3.2. *From "TRY" to "REAL" phase*

A detailed explanation and a practical demonstration of all the above procedures including the processing of completed questionnaires were provided to all researchers/staff involved in data entry, thus minimizing the risk of operator errors. Nevertheless database management staff was given a time period of three months for getting familiar with the system by simulating data entry. During this "TRY" phase of the MARK-AGE project all activities foreseen were rehearsed, from recruitment to analysis. Each recruitment centre sampled 10 volunteers who did not belong to the MARK-AGE target populations in order to verify the reliable execution of all the steps in each of the standard operating procedures (SOPs) (see Moreno–Villanueva and Capri and co-workers, this issue), including data entry. Subjects of the "TRY" phase were identified by a different primary subject code to avoid any possible confusion with 'real' subjects to be examined afterwards. This special PSC had the following structure: xxTRYyy. The two first digits "xx" identified the recruitment centre and two last digits "yy" identified the subject. This TRY phase was extremely useful as it enabled us to correct threshold values in the input and processing layer.

The procedure of entering data in the database was tested and rehearsed repeatedly by the partners, in order to prevent any problems during the phase of active recruitment. With the onset of the "TRY" phase the data input started. Each recruitment centre processed 10 "TRY-subjects", and at the end of the TRY phase the information of 80 "TRY-subjects" from 8 recruitment centres had been saved in the database (50 subjects in December, 20 subjects in January and 10 subjects in February). The entry of questionnaires from 'real' subjects started in March 2009.

Thanks to the close and very intense co-operation between all partners involved in recruitment, all possible sources for mistakes and incoherencies associated with electronic data storage,

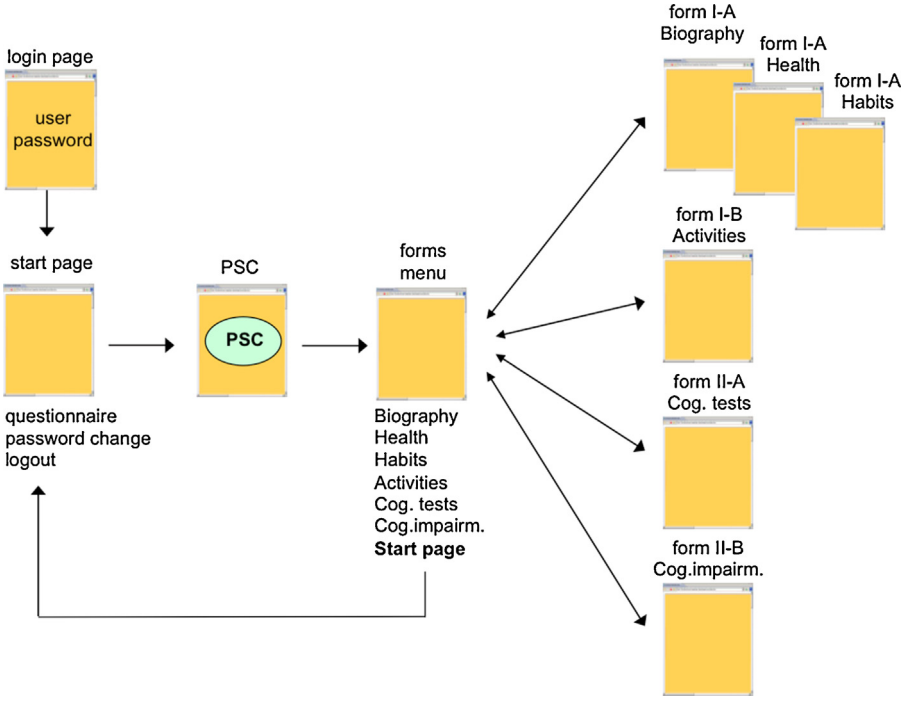

Fig. 4. Forms for questionnaire data input for the recruitment centres. PHP framework allowed login, introduction of subject code and data input in a multiple session manner. There were six questionnaire forms, which covered various parts of the interview. The interviewers used the PHP framework forms to transfer the information collected in the questionnaires to the database. After saving data in the database the same form could not be re-selected.

performed via a secure internet link, could be identified and fixed.

3.3. Data input

Two methods were available for uploading bioanalytical results to the database. The web form based data upload focused on flexibility and was based on the data upload framework that had been developed for the questionnaires. These web forms allowed the upload of analytic results per SSC and thus did not support the upload of mass data, whereas the second focussed on mass data to support the reporting of hundreds of data points per secondary subject code (SSC) at once (Fig. 3). The data file constraints were as follow:

1. The file had to be provided as a text file (".txt").
2. Single values had to be separated by a tab or a semicolon.
3. The first row had to have a distinct column name.
4. No special characters (such as "[,],!,/,;" etc) could be used in the column names with the exception of "_".
5. Column names had to start with an alphabetical character.
6. Empty columns had to be avoided.
7. One column had to contain the SSC.
8. The name of column that contained the SSC had to be specified in the upload dialogue.
9. Non-numeric values such as "not tested", "not applicable", "n/a", "nd" etc., were not supported within numerical columns. The field should be left blank.

3.4. Questionnaires

Each section of the questionnaires was represented by a web page that was build using a custom programmed PHP framework (Fig. 4). The framework has been developed to provide an easy creation of the input forms, batch uploading of analytical results, as well as a validation of the input data, e.g., the range of a numerical field or a mandatory input field using regular expressions. Additional validation was implemented in the database by enforcing not null as well as unique key constraints where necessary. In addition each set of data contained a timestamp recording the time the data set had been saved in the database.

3.5. Bioanalytical data

Each data, e.g., the results from the questionnaires or the bioanalytical results, could be entered only once per subject by the person entitled. In order to edit a set of data, a recruiter needed special permission from the administrator, which was necessary for enabling the reviewing and possibly re-entering of data for the requested subject into the database.

The web framework further ensured that users could enter and review only data they are entitled to, e.g., questionnaires or analytical data from their own laboratory. Therefore, each user had his own login, which is coupled with a certain role, e.g., recruitment centre, bioanalytical laboratory or Biobank.

3.6. Metadata

Several hundred potential biomarkers of ageing targeting different cellular functions have been measured in MARK-AGE biological material. In order to facilitate data extraction a metadata table (metastable) has been created. Metadata are used to describe digital data in order to provide relevant information about one or more aspects of the data. Metadata is often called data about data or information about information (Guenther and Radebaugh, 2004). MARK-AGE metatable contains information about parameter description, parameter short name, parameter name, parameter unit, type of biological material, method used for measurement, calculations performed, number of analysed probands, count female, count male, count RASIG, count GO, count SGO, partner ID, WP number, DB table name and comments (Table 1).

4. Conclusion

Human studies are of very high relevance in biomedical research. Therefore, the design of a study, the procedures performed and results obtained need to be machine-readable in order to facilitate data analysis and data mining. Furthermore, there is a growing interest in sharing research data within the scientific community. However, effective sharing of data requires not only sharing results but also additional information collected in a meta table containing the necessary information for interpreting and analysing data. Within MARK-AGE we have established a phenotypic database able to accommodate all different types of data generated during the project. Aside from the necessary infrastructure, a double-code system and data transfer and data sharing strategies were implemented in accordance with the specifications of MARK-AGE project.

Acknowledgements

We wish to thank the European Commission for financial support through the FP7 large-scale integrating project "European Study to Establish Biomarkers of Human Ageing" (MARK-AGE; grant agreement no.: 200880) and all MARK-AGE Consortium partners for the excellent collaboration.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.mad.2015.03.005.

References

Guenther, R., Radebaugh, J., 2004. Understanding Metadata National Information Standards Organization. NISO Press, Bethesda MD.
Jamison, D.C., 2003. Structured query language (SQL) fundamentals. Curr. Protoc. Bioinf., Chapter 9:Unit 9.2.
Teorey, T.J., Lightstone, S.S., et al., 2009. Database Design: Know It All, 1st ed. Morgan Kaufmann Publishers, Burlington MA.
Wesley, D., 2000. Relational database design. J. Insur. Med. 32 (2), 63–70.