Mechanisms of Ageing and Development 151 (2015) 31–37

Contents lists available at ScienceDirect

Mechanisms of Ageing and Development

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

The MARK-AGE extended database: data integration and pre-processing

J. Baur[a], T. Kötter[b], M. Moreno-Villanueva[a], T. Sindlinger[a], M.R. Berthold[b], A. Bürkle[a,*], M. Junk[c]

[a] Chair of Molecular Toxicology, University of Konstanz, 78457 Konstanz, Germany
[b] Chair of Bioinformatics and Information Mining, University of Konstanz, 78457 Konstanz, Germany
[c] Department of Mathematics and Statistics, University of Konstanz, 78457 Konstanz, Germany

ARTICLE INFO

Article history:
Received 31 January 2015
Received in revised form 13 May 2015
Accepted 18 May 2015
Available online 21 May 2015

Keywords:
Database
Data entry
Data integration
Data processing
Data extraction
KNIME

ABSTRACT

MARK-AGE is a recently completed European population study, where bioanalytical and anthropometric data were collected from human subjects at a large scale. To facilitate data analysis and mathematical modelling, an extended database had to be constructed, integrating the data sources that were part of the project. This step involved checking, transformation and documentation of data. The success of downstream analysis mainly depends on the preparation and quality of the integrated data. Here, we present the pre-processing steps applied to the MARK-AGE data to ensure high quality and reliability in the MARK-AGE Extended Database. Various kinds of obstacles that arose during the project are highlighted and solutions are presented.

1. Introduction

A database is a structured collection of information that can be accessed by using specific software tools. To extract the information and understand the fundamental structure of collected data, these have to be presented in such a way as to enable efficient knowledge extraction. Various forms of databases have been developed and they are typically categorized on the basis of their function. The most common type is the relational database where the information is stored in various data tables (Codd, 1970), which is commonly used in genomics, proteomics and clinical research where large amount of data have to be stored from each subject or patient (Mackey and Pearson, 2004; Yu and Salomon, 2009).

To enter, organize and select data from a database a database management system (DBMS) is necessary. These programs are specifically designed to enable interaction between user, other applications, and the database itself and cover the following issues:

1. Define, remove and modify the data structure of new or existing database tables.
2. Inserting, modifying, and deleting data.
3. Query data for reports and make them accessible to end-users.
4. Data security and recovery, registering and monitoring of users.

There are many different types of DBMSs, ranging from small systems that run on personal computers to very complex systems that run on mainframes. Structured query language (SQL) was developed in the early 1970s at International Business Machines (IBM) (Boyce and Chamberlin, 1974). It is a standard language to interact with relational databases and is currently the most widely used database language. The usage of SQL includes data insert, query, update and delete, modification and data access control.

In large research projects the sharing of data within a consortium, between different consortia, or with the scientific community at large is essential in order to boost progress during complex data analysis and modelling. An absolute requirement for sharing of a database is that the communicated data are well organized, have been entered in the correct format, carefully checked and validated. Different organizational approaches according such processes are already described for several databases in the ageing research (Craig et al., 2015; De Magalhaes et al., 2005).

During designing a project, strategies need to be implemented in order to prevent errors during data entry. Error prevention strategies, however, cannot guarantee the absence of incorrect or

* Corresponding author. Tel.: +49 7531 884035; fax: +49 07531 884033.
E-mail address: alexander.buerkle@uni-konstanz.de (A. Bürkle).

http://dx.doi.org/10.1016/j.mad.2015.05.006
0047-6374/

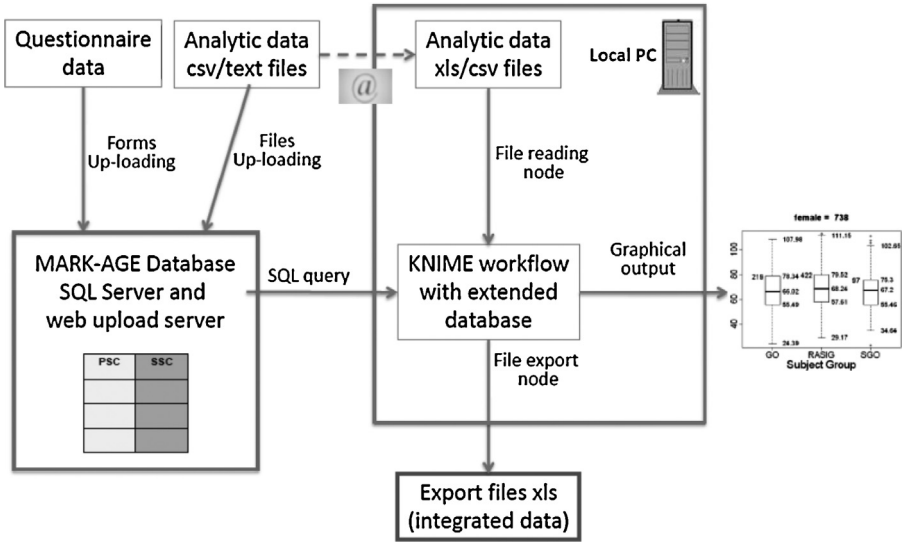

Fig. 1. Scheme of the SQL and KNIME data source fusion.
Data were either uploaded to the SQL database via the internet, or directly implemented in KNIME. To generate a complete set of data both sources were integrated within KNIME on a central place.

incomplete data entry during the project (Van den Broeck et al., 2005). Data that have not been screened and checked for misleading information may produce false results and conclusions. 'Data cleaning', *i.e.* the identification and correction of errors in order to improve data quality before storing and analysing data is therefore an indispensable part of the data management process (Chapman, 2005; Rahm and Do, 2000; Van den Broeck et al., 2005). There are, however, many different error types, and furthermore error sources are not always easy to identify. Typical examples are errors during measurement, data entry or data integration (Hellerstein, 2008). For data evaluation, analyzers must know about possible data quality problems that can compromise the validity of results. This topic is also addressed in the Digital Ageing Atlas, where each entry is connected to the belonging source of raw data, offering the possibility to check original contents (Craig et al., 2015).

Below, it is explained how the MARK-AGE database was prepared and extended within the KNIME data integration platform. Hidden problems in various data sets, as well as handling strategies for misleading data are presented. As the cases mentioned did occur despite careful study design, they may provide some guidance to avoid similar problems in future studies.

2. Materials and methods

2.1. SQL MARK-AGE database

The MARK-AGE database was established using Structured Query Language (SQL) (Kötter and Moreno-Villanueva et al., this issue). SQL is a commonly used database management system for relational databases. The data tables contain the raw data uploaded by MARK-AGE Consortium members. SQL is a rather complex tool for non-experts in informatics. Therefore, KNIME was chosen for querying the SQL database and for further processing the MARK-AGE data additionally (Fig. 1).

2.2. Konstanz information miner (KNIME)

The Konstanz Information Miner (KNIME) is a user-friendly data integration platform, which enables visual assembly and interactive execution of data pipelines. KNIME enables easy integration of new algorithms or visualization methods as modules or nodes. For a clear structuring of big workflows KNIME offers the opportunity to collapse a group of selected nodes into a so-called metanode. In addition, KNIME provides plugins for common programing languages and can be used as database management system (Berthold et al., 2007). To avoid the risk of losing information, the already generated parts of the SQL database tables were opened in KNIME after connecting to the MARK-AGE database server (Fig. 1). KNIME did not provide direct reading access to the original data tables, but to mirrored or joined so-called 'views' that were generated by the SQL DBMS. Thus, this system guarantees a maximal level of safety for the already established SQL based data. Each step performed with KNIME is executed and saved in a dedicated node, which thereby works as documentation platform. As a result, a workflow of nodes was generated performing the complete integration of the MARK-AGE database (Fig. 2).

2.3. KNIME server

The KNIME server allows storage and accessing of workflows *via* the internet. User access rights control how data are grouped for projects, workgroups or departments. The server was used to store and document KNIME workflows of the MARK-AGE database extension process (https://www.knime.org/knime-server).

2.4. Ethical clearance

The MARK-AGE study has been approved by the appropriate ethics committees and has been performed in accordance with the ethical standards laid down in the Declaration of Helsinki. All study subjects gave their informed consent prior to their inclusion in the study. This is described in great detail in Capri et al., this issue.

3. Results

3.1. Data entry

To guarantee the blinding of the study, questionnaire data (comprising descriptors of the respective subject such as age and gender) and bioanalytical data were stored separately. Questionnaires were uploaded to the database under a primary subject code (PSC) and data from biochemical analyses of samples were recoded

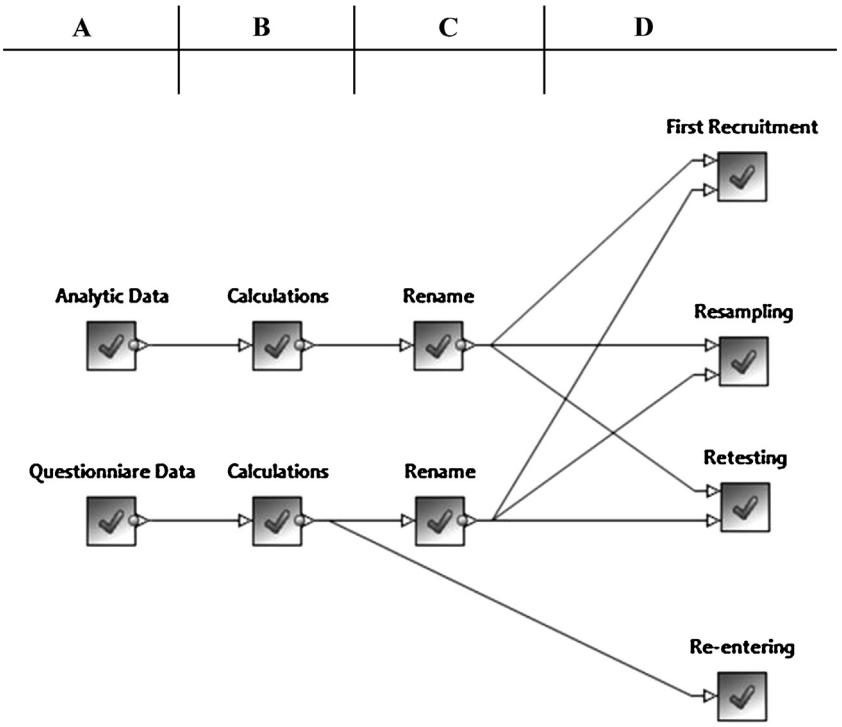

Fig. 2. Overview of the KNIME preparation workflow for the extended database.
Information was separately implemented for bioanalytical and questionnaire data (A) Subsequently, specific calculations (B) and the renaming (C) of the parameters were performed. In the last step, subgroups from several recruitment phases were separated (D).

and entered into the database under the secondary subject code (SSC) (see Bürkle et al., this issue). To join both types of information, a translation table (termed 'PSC-to-SSC' table) was established during the project (see Moreno-Villanueva and Kötter et al., this issue).

3.1.1. *Entry of* xls *and* csv *data tables*

For logistic reasons not all of the collected data from biochemical analysis were uploaded to the SQL database *via* the website interfaces. Therefore *xls* or *csv* tables containing the secondary subject code with the respective data were integrated with the already established database using KNIME, thus, leading to the formation of an 'Extended Database' (Fig. 1). The tables, manually generated by the analyzers offered problems with the subject codes as some of those were either invalid due to typos or multiple usage. To solve this problem, a KNIME workflow was established that automatically identifies and rejects all rows of a table containing invalid or multiple subject codes. Thereby each incoming *xls* or *csv* file was checked separately, first for the uniqueness of SSCs by counting and second for validity by comparing all available codes with the PSC-to-SSC table from the SQL database, containing all valid subject code information (Fig. 4). Invalid entries were separated and documented for communication to the partners. Applying this procedure, a total of 19 data tables were added to the Extended Database containing on average $1.3\% \pm 1.2$ invalid and $0.4\% \pm 0.8$ multiply entered subject codes. The high values of standard deviations indicate that some tables displayed more errors in coding than others. Additionally, the amount of miss-entered subject codes indirectly reflects the data quality. Another scenario occurring in the data tables was the mix-up of codes and respective bioanalytical values in the laboratories. Those cases could only be identified if they led to outliers in some downstream analysis, for example if a female subject is attributed values that applied to males only.

Table 1
List of general modifications performed during construction of the Extended Database.

Recalculation of non-SI units to SI units
Normalizations on parameters
Normalizations on parameters with values measured by another partner
Removal of subjects with values beyond the boundaries of measuring
Calculation of ratios on parameters measured
Calculation of the means of duplicate measures
Normalizations for batch effects
Exclusion values from ineligible samples (*e.g.*, frozen samples inadvertently thawed during shipment or storage)
Calculation of established scores like BMI and HOMA index
Calculation of new scores developed during the MARK-AGE study

3.1.2. *Addition of data columns*

After data entry, some bioanalytical researchers ('analyzers') requested calculations like normalizations and corrections on their specific parameters. Therefore, a KNIME workflow was generated performing the steps requested, separately for each requesting analyzer (Fig. 3). If normalizations were performed on parameters, new columns were appended to the extended database in order to maintain the original data. The Workflow not only documents the performed modifications but also automatically renews the calculations each time new data were uploaded.

Based on existing parameters, compound parameters such as ratios between individual parameters were calculated and appended to the Extended Database, like body mass index (BMI) (Must and Anderson, 2006) or Homeostasis Model Assessment (HOMA) index (Matthews et al., 1985). Such compound parameters have either been published already or were newly designed by the MARK-AGE Consortium members, like the 'Nutrition Score'.

Table 1 shows a list of general calculation methods performed. The established KNIME workflow is used as documentation base and organized in a way that new columns could easily be added by the user.

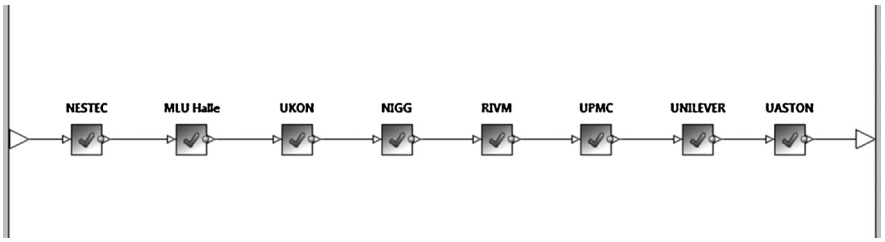

Fig. 3. Schematic overview of the calculation workflow.
Metanodes are used to perform the calculation steps separated for each project partner. The clear structure simplifies the usage and visualizes each step performed.

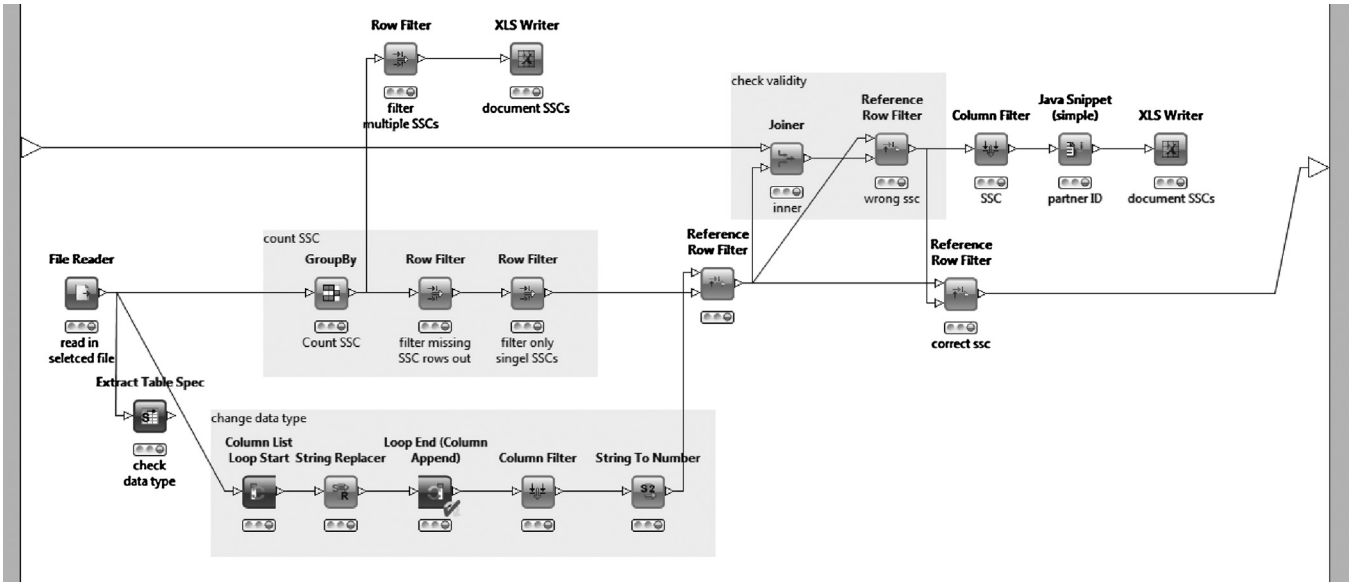

Fig. 4. Representative schematic overview of the KNIME file append flow.
Three steps were performed to check the incoming files from the MARK-AGE partners. SSCs were counted (middle box) and controled for validity (upper box). In addition the file format was checked (lower box) and converted if necessary. Invalid or mulitple codes were excluded and documented in xls files via the red xls writer nodes.

3.2. Data pre-processing

3.2.1. Re-naming of parameters

The column names of the SQL database tables had been designed using short terms and the underscore sign. Parameter names, however, must be usable for headings in graphs *etc*. The re-naming of the parameters was performed with a standard KNIME node that automatically translates the column names, according to a reference table containing original column names as well as new names. As a central place to store both kinds of information was necessary, they were implemented in the metatable established during the project (see Kötter and Moreno-Villanueva et al., this issue). Corrected names contain the more readable full-length name or, if too long, the standard abbreviation as well as the unit in which the parameter was measured.

3.2.2. Joining of analytic and questionnaire data

A necessary step to work with the extended database, was the joining of separately stored bioanalytical and questionnaire data. As questionnaires were stored under the PSC and bioanalytical measurements under the SSC, a simple joining procedure was performed using the 'PSC-to-SSC' translation table. Upon checking of the data, however, a discrepancy in the numbers of subjects between bioanalytical and questionnaire data was noted.

As the questionnaires were divided in six electronic parts, the PSC had to be entered up to six times separately. Therefore, in some cases different PSC codes with typos or reversed digits appeared for a single subject. The database recognized each newly entered, differing PSC as a new subject and generated a new entity. As a result, for one subject more than one SSC could be available, and the complete questionnaire information from one subject could be dispersed as well (Table 2). Consequences of this problem may well have affected any phase of recruitment.

Table 2
Chart illustrating the PSC entering problem.

	PSC	SSC	quest1	quest2	quest3	quest4	quest5	analysis
1	0200123	12345	x		x	x	x	x
2	0200128	12346		x	x	x		
3	02rt123	12347	x	x	x			
4	02RT123	12348				x	x	

Row 1 and 2 reflect entries for the same subject in the first recruitment round. When questionnaire two was entered the interviewer transposed an 8 with the 3 at the last digit by mistake. Because the second code is unknown to the system it generates a new row reflecting a new person by mistake. The same problem occurs if indicators from further recruitment rounds were transposed (rt –> tr) or indicated differently with lower or upper letters for one subject (row 3 and 4).

3.2.2.1. First recruitment. For the main recruitment phase, fixing erroneous multiple PSC insertion was not done. As there is no valid possibility to assemble the different entries belonging to one subject, all cases where only parts of questionnaires had been entered were excluded from further analysis.

3.2.2.2. Subsequent recruitment phases. Additional recruitment rounds, termed 're-sampling' or 're-testing', were performed during

Table 3
Overview of achieved subject numbers after database extension.

	Recruited subjects	Valid subjects for analytical data analysis
First recruitment	3228	3048
Re-sampling	98	97
Re-testing	410	391

the project. In these phases only a subset of subjects was examined and data were entered again. For identifying purposes, an altered PSC code, with a separate identifier consisting of two letters was used. If those letters were mixed up during various entries, again different PSCs were generated for one subject. Fortunately enough, those mis-entries could be retraced, according to the remaining parts of the code. As a result, a KNIME workflow was established removing the identifier and writing all separately stored information into one valid PSC with a correct indicator.

3.2.3. Filtering data

According to the design of MARK-AGE, criteria for the enrolment of subjects had been established. Subjects had to fall in the age range 35–74.9 years; and furthermore, positivity for hepatitis B or C was an exclusion criterion. For logistic reasons, some subjects were entered into the database that did not satisfy these criteria. General filters were established in a KNIME node that exclude and document those subjects.

Two exceptions had to be considered during the setup of the algorithm. The 'spouses of GEHA Offspring' [SGO] group, were allowed to exceed the defined age range as their number was rather low. Furthermore subjects recruited during the re-testing phase obviously had different age requirements: If a person was 73 at the time of the first recruitment, 3 years later the age exceed the above-mentioned limit, which, however, was accepted in this case.

After the data cleaning steps mentioned above it was possible to determine the number of valid entries of subjects that participated in the study. Therefore, we defined requirements necessary for a subject to be considered in standard analytical analysis. A 'valid subject' required at least one bioanalytical parameter measured successfully (beside hepatitis analysis, where positivity was an exclusion criterion) and the fully entered part of the questionnaires containing age and gender information while a recruited subject requires one part of the questionnaires or blood sampling collections entered into the database. Table 3 shows the number of recruited and valid subjects established with the explained cleaning steps and definitions.

3.3. Corrections of coupled parameters

Beside corrections on single parameters, corrections of coupled parameters were also set up.

3.3.1. ATC codes

As drug names vary from country to country, the standardized Anatomical Therapeutic Chemical (ATC) classification system was used to clearly indicate the drug intake of a subject. An ATC code consists of 5 levels defined by a specific order of letters and numbers (WHO, 2013). In the MARK-AGE database, typos in these codes occurred: for example, the number 0 and the letter O were used synonymously or the coding system was not maintained. An algorithm was set up correcting the typos and extracting the invalid codes, which should subsequently be corrected by the responsible recruitment centres. To extract information about the disease, level 3 ATC codes of all MARK-AGE subjects were grouped and a disease translation table was generated (Table 4).

Table 4
Overview of the available ATC codes in the MARK-AGE population. The established classification system uses the ATC code until level 3.

level 1	level 2	level 3	Disease indication
A	02	A	Gastrointestinal
A	02	B	Male sex hormones
A	03	A	Male sex hormones
A	04–05	A	Gastrointestinal
A	05	B	Gastrointestinal
A	06	A	Gastrointestinal
A	07	A	Infections disease
A	07	E	Antiinflammatories/antirheumatic compounds
A	07	F	Gastrointestinal
A	08	A	Diabetes
A	09	A	Gastrointestinal
A	10	A,B	Diabetes
A	11	A,C,D,G,H	Vitamins
B	01–02	A	Thrombosis/coagulation disorders
B	03	B	Vitamins
C	01	A-E	Cardiac disease
C	02	A,C,D	Hypertension cluster
C	03	A,B,C,E	Hypertension cluster
C	04	A	Thrombosis/coagulation disorders
C	07	A,B,C	Hypertension cluster
C	08	C	Hypertension cluster
C	08	D	Cardiac disease
C	09	A,B,C,D,X	Hypertension cluster
C	10	A,B	Lipid metabolism
D	05	A	Skin diseases
D	07	A	Antiinflammatories/antirheumatic compounds
D	07	C	Infections disease
D	11	A	Skin diseases
H	02	A	Antiinflammatories/antirheumatic compounds
H	03	A,B,C	Thyroid disorders
J	01	A,C,F,X	Infections disease
J	05	A	Infections disease
L	01	A,B,X	Cancer therapy
M	01	A,C	Antiinflammatories/antirheumatic compounds
M	02	A	Pain
M	04	A	Gout
M	05	B	Bone disease
M	09	A	Antiinflammatories/antirheumatic compounds
N	02	A,B	Pain
N	03	A	CNS disorders other than depression
N	04	B	CNS disorders other than depression
N	05	A	CNS disorders other than depression
N	05	B	Depression/anxiolytics
N	06	A	Depression/anxiolytics
N	06	B,D	CNS disorders other than depression
N	07	X	CNS disorders other than depression
P	01	B	Infections disease
R	03	A,B,D	Lung/bronchial disorders
R	05	C,D	Lung/bronchial disorders
S	01	A	Infections disease
S	01	B	Antiinflammatories/antirheumatic compounds
S	01	E,F,X	Eye disease
V	01	A	Immune system disorders

3.3.2. Blood parameters in standard units

For each subject a general blood count was performed during the study. As standard blood analysis devices and their reporting format vary in different countries, the upload tables were generated such that the number and the unit for a single parameter were entered in two different columns. In this process, the unit had to be selected from a drop-down menu. In order to standardize the parameter units an algorithm had to be established re-calculating all values according to the international metric system (SI units) (Table 5). Large unstained cells (LUC) and platelet distribution width (PDW) were only measured in one laboratory and therefore, excluded from the analysis.

3.3.3. ZUNG scale scoring

The self-rating depression scale (Zung, 1965) was used to monitor if the subjects suffer from depressive disorders (see Bürkle

Table 5
List of blood count parameters analyzed and units used.

Parameter	Long name	Unit
MCH	mean corpuscular haemoglobin	picogram
MCHC	mean corpuscular haemoglobin concentration	g/dl
MCV	mean cell volume	femtoliters
HCT	haematocrit	%
RDW	red cell distribution width	%
HGB	haemoglobin	g/dl
HDW	haemoglobin distribution width	g/dl
RBC	red blood cells	million/μl
WBC	white blood cells	thousand/μl
Neutrophils	neutrophils	thousand/μl
Eosinophils	eosinophils	number/μl
Basophils	basophils	number/μl
Monocytes	monocytes	number/μl
Lymphocytes	lymphocytes	number/μl
Platelets	platelets	thousand/μl
MPV	mean platelet volume	femtoliters

et al., this issue). Twenty questions had to be answered by using the following options: a little of the time, some of the time, good part of the time, most of the time. Each question was scored afterwards to calculate the depression status value. Gaps were not allowed in this system but the subjects had the choice to select 'na' (not applicable) if he or she did not wish to answer the question. Therefore, it occurred that several questions were not usable for the rating system. As already published (Shrive et al., 2006) these gaps were filled with the mean of the points from the individual subject.

3.3.4. *Calculating uniform time intervals*

Information requested on food and beverage intake was entered with a tabular system, in which the subject has to specify the amount consumed either per day, week or month. To work with this data in a standardized fashion all intakes were recalculated to a weekly and monthly indication with an established algorithm.

4. Conclusion

In this paper, we describe the necessary steps that were performed on the MARK-AGE raw data to generate a database for the final user. The tools used present methods to detect and handle problems hidden in the data structure of collected raw data. Problems we reported should be used to implement preventive strategies in new aging research projects. Additionally KNIME is introduced as web based tool for the development of an easy-to-handle data communication platform.

Even our best efforts invested in the design of the project could not guarantee complete prevention of errors or problems related with data entry into the database. During the project, error sources in the growing database were identified, and thus, data quality improved continuously. Some of the problematic effects were identified shortly after the launch of the project whereas others were hidden and only detected after analysis. The fact that an international project like MARK-AGE involves several countries with divergent standards and guidelines made it even more difficult to establish a standardized working system. Since the creation of large European databases for the analysis of biological systemic effects will continue to be a relevant task, strategies to generate reliable databases of highest quality are necessary. So far publications covering this aspect have been rare. With our above description of essential steps on the MARK-AGE database, we provide relevant information from our hands-on experience on frequent error sources and ideas for preventive solution strategies.

Investing sufficient time and manpower in the construction phase of a project and its database is essential and avoids high costs, delays and quality loss. In particular, a realistic estimate of the required funds for accomplishing the crucial programming work in the early phase of the project is of utmost importance. These programming tasks include the implementation of a database structure and the establishment of an appropriate backup system, the design of web-interfaces for data entry with suitable consistency checks like value restrictions (avoiding free entry frames whenever possible), the selection and implementation of error-correcting subject codes, and the implementation of a framework for subsequent analysis. KNIME was chosen for data integration and retrieval because it is easier to handle for non-IT-experts and directly provides analysis tools and elaborate reporting features. Even non-experts can efficiently work with this system after a training period of approximately one week, based on the user-friendly interface and intuitive node structure.

Without sufficient knowledge about the data background and the ways data were prepared, analysis can cause misleading results. The documentation of the extended MARK-AGE database construction is completed and a detailed description is available for users. Therefore the work presented is also necessary for upcoming publications presenting results on the MARK-AGE data. If parts of published data were to be integrated in other aging research databases like the Digital Ageing Atlas (Craig et al., 2015) it would be necessary to know about data source and data preparation strategies. Lastly, at the time the MARK-AGE database could be made publicly available, it would be necessary that each user is familiar with the data background.

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). Furthermore we wish to thank all MARK-AGE Consortium members for their efforts to make this work possible. Our special thanks go to Lothar Gasteiger, Thorsten Meinl and Peter Burger for the support regarding hardware and software tools.

References

Berthold, M.R., Cebron, N., Dill, F., Gabriel, T.R., Kötter, T., Meinl, T., Ohl, P., Sieb, C., Thiel, K., Wiswedel, B., 2007. KNIME: the konstanz information miner. In: Studies in Classification Data Analysis and knowledge organization. Heidelberg-Berlin, Springer-Verlag.

Chamberlin, D.D., Boyce, R.F., 1974. SEQUEL: a structured english query language, SIGFIDET, 47 Proceedings of the 1974 ACM SIGFIDET workshop on Data description, access and control, 249-264.

Chapman, A.D., 2005. Principles and Methods of Data Cleaning Occurrence Data. Version 1.0. Report for the Global Biodiversity Information Facility. Copenhagen, 1–72.

Codd, E.F. (1970). Relational Model of Data for Large Share Data Banks, Communications of the ACM, 13:6.

Craig, T., Smelick, C., Tacutu, R., Wuttke, D., Wood, S.H., Stanley, H., Janssens, G., Savitskaya, E., Moskalev, A., Arking, R., De Magalhaes, J.P., 2015. The digital ageing atlas: integrating the diversity of age-related changes into a unified resource. Nucleic Acids Res. 43, D873–D878.

De Magalhaes, J.P., Costa, J., Toussaint, O., 2005. HAGR: the human ageing genomic resources. Nucleic Acids Res. 33, D537–D543.

Hellerstein, J.M., 2008, Quantitative data cleaning for large databases, Survey for the United Nations Economic Commission for Europe (UNECE), http://db.cs.berkeley.edu/jmh

Mackey, J., Pearson, W.R., 2004. Using relational databases for improved sequence similarity searching and large-scale Genomic Analyses. In: Current Protocols in Bioinformatics. John Wiley & Sons, Inc, Chapter 9, Unit 9.4.

Matthews, D.R., Hosker, J.P., Rudenski, A.S., Naylor, B.A., Treacher, D.F., Turner, R.C., 1985. Homeostasis model assessment: insulin resistance and beta-cell function from fasting plasma glucose and insulin concentrations in man. Diabetologia 28 (7), 412–419.

Must, A., Anderson, S.E., 2006. Body mass index in children and adolescents: considerations for population-based applications. Int. J. Obesity 30 (4), 590–594.

Rahm, E. and Do, H.H. 2000. Data cleaning: problems and current approaches. bulletin of the IEEE computer society technical committee on data engineering, 23, 4, 3–13.
Shrive, F.M., Stuart, H., Quan, H., Ghali, W.A., 2006. Dealing with missing data in a multi-question depression scale: a comparison of imputation methods. BMC Med. Res. Methodol. 13 (6), 57.
Van den Broeck, J., Cunningham, S.A., Eeckels, R., Herbst, K., 2005. Data cleaning: detecting, diagnosing, and editing data abnormalities. PLoS Med. 2 (10), e267.
WHO, 2013. Collaborating centre for drug statistics methodology, guidelines for atc classification and ddd assignment. Oslo, 2012.
Yu, K., Salomon, R., 2009. Peptidedepot: flexible relational database for visual analysis of quantitative proteomic data and integration of existing protein information. Proteomics 9 (23), 5350–5358.
Zung, W.W.K., 1965. A Self-Rating Depression Scale. Arch. Gen. Psychiatr. 12, 36–70.