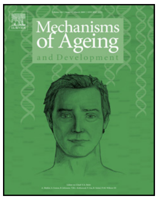
Mechanisms of Ageing and Development 151 (2015) 38–44

Contents lists available at ScienceDirect

Mechanisms of Ageing and Development

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



MARK-AGE data management: Cleaning, exploration and visualization of data

Jennifer Baur[a], Maria Moreno-Villanueva[a], Tobias Kötter[b], Thilo Sindlinger[a], Alexander Bürkle[a,*], Michael R. Berthold[b], Michael Junk[c]

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

ARTICLE INFO

Article history:
Received 30 January 2015
Received in revised form 13 May 2015
Accepted 18 May 2015
Available online 21 May 2015

Keywords:
Data cleaning
Missing data
Batch effects
Outliers
Data visualization

ABSTRACT

Databases are an organized collection of data and necessary to investigate a wide spectrum of research questions. For data evaluation analyzers should be aware of possible data quality problems that can compromise results validity. Therefore data cleaning is an essential part of the data management process, which deals with the identification and correction of errors in order to improve data quality.

In our cross-sectional study, biomarkers of ageing, analytical, anthropometric and demographic data from about 3000 volunteers have been collected in the MARK-AGE database. Although several preventive strategies were applied before data entry, errors like miscoding, missing values, batch problems etc., could not be avoided completely. Such errors can result in misleading information and affect the validity of the performed data analysis.

Here we present an overview of the methods we applied for dealing with errors in the MARK-AGE database. We especially describe our strategies for the detection of missing values, outliers and batch effects and explain how they can be handled to improve data quality. Finally we report about the tools used for data exploration and data sharing between MARK-AGE collaborators.

1. Introduction

Today's high throughput techniques enable the generation of huge data volumes in relatively short time periods. With increasing amounts of aggregated information, the issue of databases gets more and more important. A database is a collection of data in an organized form and can be categorized on the basis of their function. The most common type is the relational database where the information is stored in various data tables. This type is widely used in the fields of genomics and proteomics, where large amount of data must be stored for a single subject (Pearson, 2004; Yu and Salomon, 2010). There are software programs available that enable the user to store, modify and extract information from a database, the so-called database management systems (DBMS). They are especially designed to provide an interaction between user, other applications, and the database itself. Typical software programs that allow the creation and administration of relational databases are Microsoft SQL Server, IBM DB2 or Oracle.

To ensure that data is well organized, entered in the correct format and annotated, a data management plan should be prepared before the beginning of a study. An accurate planning includes not only data handling during the data collection but also after the project is completed. However, best efforts established in project's designs to avoid errors during data collection, cannot prevent from incorrect or incomplete data. Errors in real-world data are common and are to be expected (Orr, 1998; Redman, 1998). Misleading or missing information in databases disables the confirmation of results and conclusions after data interpretation and analysis. Therefore data cleaning is an essential step for the Information Management Chain before storing and analyzing data (Chapman, 2005). Error sources in many cases are not clear detectable and occur in a variety of fashions. Obvious examples are data entry errors, measurement errors or data integration errors (Hellerstein, 2008).

Manual cleaning of data is laborious and time consuming, and in itself prone to errors (Maletic and Marcus, 2000). Data cleaning strategies including the use of machine learning for guided database repair (Yakout et al., 2010), inferring and imputing of missing values (Mayfield et al., 2010) and resolving of inconsistencies using functional dependencies (Fan et al., 2008) have been

[*] 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.2015.05.007
0047-6374/© 2015 Elsevier Ireland Ltd. All rights reserved.

described before. Furthermore Batini and collaborators provide several general methodologies for the improvement of data quality (Batini et al., 2009). After investigating on cleaning purposes, data are of high quality if, they are fit for their intended uses in operations, "decision making and planning." (Juran et al., 1974). Controlled high quality data do not affect accuracy and efficiency of data analysis and can be further processed by users.

The process of transforming data into sensory stimuli and visual images is called data visualisation (Schroeder et al., 2003). Powerful charts, diagrams or maps provide solutions to explore, analyse, and present data. Furthermore data visualisation is an important tool for effective data communication in large research consortia. However data can possess different forms such as numbers, graphs, images, or texts and their visualisation entails some challenges. Finally stored and analyzed data might be requested by other researches. As a result, beneficial strategies for data sharing are also necessary. In the field of ageing research data from databases are typically represented over the internet. One well-known example is the Digital Ageing Atlas where age-related data of different biological levels were collected from the literature, stored and provided on a web page (Craig et al., 2015). Each individual parameter can be requested with a definition according to age and a connection to the raw data. As a broad collection of data increase the chance to obtain powerful results, the Human Ageing Genomic Research combines three databases linking aspects of age-related genetic and evolutionary studies on their web pages (Tacutu et al., 2013). Data between all parts is linked, and users can simply request the information of interest on the user-friendly interface.

This work describes MARK-AGE data management efforts focusing on strategies such as data retrieval, identification of errors and assurance of data quality. Furthermore we created automatic reports on a web portal using Konstanz Information Miner (KNIME) as interface for data sharing and visualization.

2. Material

2.1. MARK-AGE database

The MARK-AGE database is a relational database and was established using Structured Query Language (SQL) (see Kötter and Moreno-Villanueva et al., this issue) and prepared for usage as described in (see Baur et al., this issue). SQL is a commonly used database language and allows the retrieving of data from a database fast and efficiently. The language can be used not only to create databases but also for updating, retrieving and sharing data with other users. The MARK-AGE database itself consists of analytical, anthropometric and demographic data collected from about 3300 subjects recruited across Europe (see Bürkle et al., this issue).

2.2. Konstanz Information Miner (KNIME)

The Konstanz Information Miner is a modular environment, which enables visual assembly and interactive execution of a data pipeline. It is designed as a teaching, research and collaboration platform, which enables easy integration of new algorithms, data manipulation or visualization methods as new modules or nodes (Berthold et al., 2007).

2.3. KNIME team space

KNIME can be used in a team to share work with other researches, by keeping data files and data analysis workflows in one central shared place. Also metanodes including pre-programmed workflows can be used as a reference to the centrally stored version by all team members in their local workflows. (https://www.knime.org/knime-teamspace).

2.4. KNIME server and web portal

KNIME Server allows storing workflows and accessing them from anywhere via the internet. User access rights control how data is grouped for projects, workgroups or departments. The web portal is the perfect way to distribute preconfigured workflows, created by administrative users, to all end users. (https://www.knime.org/knime-server).

2.5. Programming language R

R is a free available scripting language for statistical computing and the generation of high quality graphs (Ihaka and Gentleman, 1996). A wide choice of pre-programmed R packages can easily be implemented and used for data analysis. KNIME incorporates an R plugin, enabling the use of the R language and its packages in workflows.

3. Results

3.1. Data quality: strategies for cleaning data

Data quality reflects the goodness of the evaluated data. Unfortunately, in spite of efforts put into data entry (see Bürkle et al., this issue), errors still occurred and therefore data cleaning was necessary. In order to visualize and detect errors, automatic standard analyses (Table 1) were performed on entered data. They consisted of histograms, scatter- and boxplots generally used in descriptive statistics. After identification, the MARK-AGE database cleaning strategy includes (1) clearing of missing values, (2) removal of outliers and (3) detection of batch effects. All three types of errors could, if untreated, compromise the conclusions that are drawn from the data.

3.2. Dealing with missing values

Two main scenarios are responsible for missing data in the MARK-AGE database. On the one hand a missing value can occur completely at random because the sample tube was broken, defrosted, lost etc., On the other hand missing data were introduced because specific parameters were only measured in low throughput analysis. In these cases, instead of all recruited subjects, only 300 individuals were measured. A precise definition on the occurrence of missing values is therefore essential to determine the handling strategies. In addition missing data analysis depends on the extent of missing values and their influence on covariates. A large percentage of missing information in a covariate, e.g., females, can lead to group under-representation and requires further investigation. This leads to the effect that selected parts of the database offer different states of missing values and requires adjusted missing value handling. Therefore a general replacement of missing data in the original Database is not possible. The handling of missing values was performed in the individual downstream analysis, using either complete case analysis or a substitution method as described in the following.

If covariates are equally distributed and values are missing completely at random a substitution is not obligatory. In this case a complete case analysis was performed including only data available for all parameters.

To compensate for under-representation of a covariate, the missing values can be replaced with statistical estimates. One popular method is mean substitution, i.e., replacing the missing values

Table 1
Table of automated analyses that were available for the quality checks of parameters.

Report name	Outcome
Scatterplot with age	Scatterplot for a parameter against age with a linear regression line and the Pearson and Spearman correlation coefficients
Boxplot for selected subgroups	Boxplot for a parameter grouped for a pre-defined subgroup (age, recruitment center, gender e.g.)
Histogram	Simple histogram for one parameter
Correlation of two parameters	Scatterplot for two selected parameters and the Spearman correlation coefficient
Batch check recruitment time	Scatterplot of a parameter against the recruitment time
Batch check internal id	Scatterplot of a parameter against the time of biochemical analysis
Empirical distribution function for selected subgroups	Empirical distribution function for selected subgroups of a parameter with the information about the significant differences calculated with bootstrap analysis.

by the average of available values. However, since mean substitution in general cannot be recommended due to biased results, more robust statistical methods like median substitution or multiple imputations should be chosen. Whenever replacement of missing values is necessary, it should be checked if the substitution strategy affects the intended analysis. To accomplish this task, one can start from a representative complete data set in which missing values are introduced randomly. Since the number of missing values can be controlled in this scenario, the effect of the replacement strategy depending on the number of missing values can be analyzed statistically. The resulting information is quite useful for the decision whether the replacement strategy should be applied to a specific data set. The substitution method was adjusted for each individual analysis.

As representative example, we selected a data set without missing values from the MARK-AGE database. In (Fig. 1) the spearman correlation of two parameters was determined. The replacement steps were performed randomly and repeated a hundred times. After a replacement of 5% of all data values the correlation differs significantly from the original outcome. This means that replacing more than 5% of missing values would affect the conclusions derived from the initial data. In this case another method than the median imputation must be found.

3.3. Dealing with outliers

Outlier detection is an important task to achieve previous to data analysis. Labelling methods for detecting outliers can be used if the distribution of data sets is difficult to identify or the data cannot be transformed in a proper distribution. We favoured the method based on the interquartile range over other classical approaches, such as standard deviation or z-score, because quartiles are more resistant to extreme values. Therefore an outlier was defined as any data point outside the range between the specified lower and upper quantiles (Fig. 2).

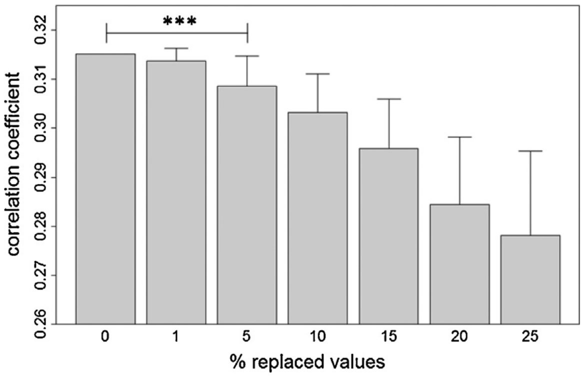


Fig. 1. Overview of the change in correlation between two randomly selected parameters after median imputation. Increasing numbers of data points were randomly removed from the original dataset and replaced by the parameters median. With increasing percentages of replaced data the calculated correlations deviate from the original value. The first significant deviation occurs for this example at 5% (one way ANOVA $p < 0.05$).

A general first check for extreme outliers is the inspection of values exceeding the biological borders, for example a systolic blood pressure of 400. Those suspicious values were clearly included by mistake and either corrected or if not possible excluded from data analysis. In case suspect values were either a legitimate part of the data or the cause was unclear, the data points were only excluded after been identified as outliers by the interquartile range approach.

For standard analysis on inter-parameter correlations and linear modelling analysis 3% quantiles were defined for outlier removal. Because both the upper and lower quantiles were considered, a total of 6% of all data values were excluded for each parameter during analysis. Excluding this amount the data removes the most prominent outliers from the analysis, and the results reflect reliable outcomes. If by excluding data points the total amount of values is decreased too much the analysis can again get distorted. Therefore, the minimal amount of data required for applying this outlier removal strategy was set to 300 data points.

In specific cases when only smaller subgroups of data sets are analyzed it can happen that the selected values offer very high distributions. The quantile limits must then be adapted until the outcomes stay statistically stable. If parameters offer too high distributions or undergo the amount of minimal required values they must be excluded for the specific analysis.

3.4. Dealing with batch effects

A batch effect is the result of systematic error introduced artificially during sample processing. A simple method for detecting batch effects is to perform tests of association between the analyzed values and various experimental variables (e.g., reagent lots, laboratory conditions, personnel differences, processing day or changes in protocols) to see if there are a large number of artificial associations across the data.

In the MARK-AGE study about 3000 subjects were recruited (see Bürkle et al., this issue). In order to obtain the necessary aliquots, body fluids from each subject were processed according to the sampling procedures (see Capri and Moreno-Villanueva et al., this issue) and shipped to the analytic centers. Depending on the assays used, the samples were processed at once, in batches or continuously. To detect batch effects all measured parameters were examined, either for the date of the sampling at the recruitment center or for the processing order or upload day at the analytic center. A straight line that best represents the data on a scatter plot (line of best fit) was used to visualize relationships among variables. In case of batches the maximal rate of change (slope) of the line differs considerably from zero (Fig. 3A), indicating a batch problem. In order to determine not only continuous trends but also discontinuous changes, the points on the line at which changes take place were calculated. The data located between two points were considered a batch.

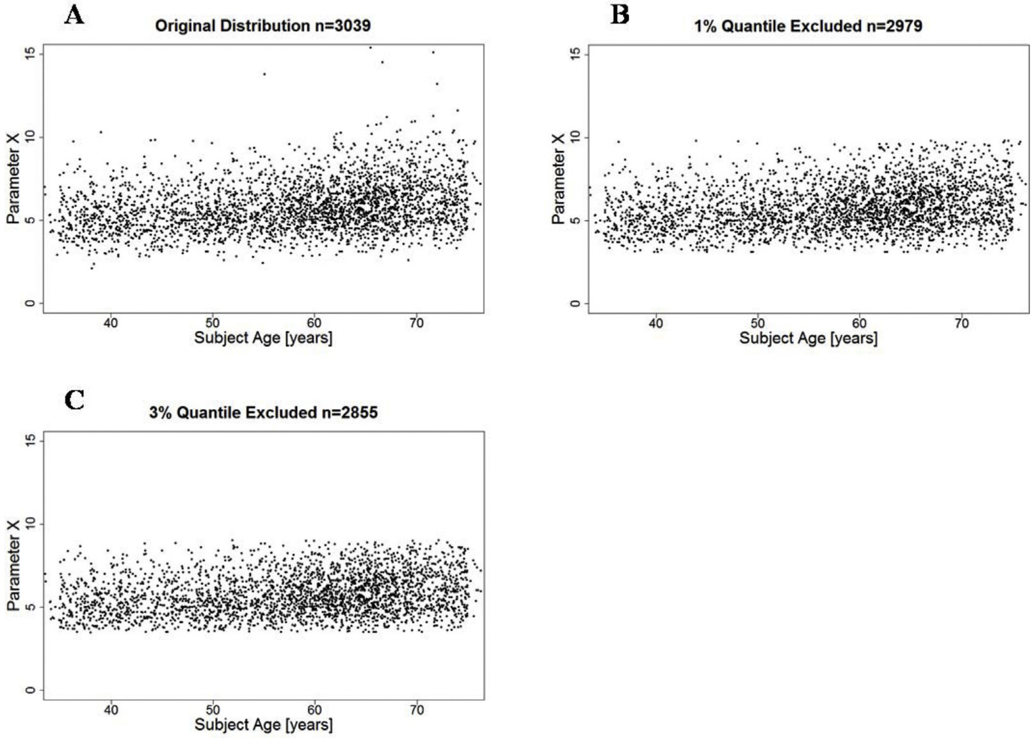


Fig. 2. Scatterplots showing the distribution of a representative parameter and the removal of outliers with the interquartile approach. In the graphs the representative parameter is plotted against the age of the subjects expressed as AgeDaysAsYears. The graphs represent the original distribution (A) of the parameter and after the exclusion of 1% quantiles (B) and 3% quantiles (C). It is apparent that outliers were removed in a stepwise fashion by using this technique.

Since origins of batch effects can be many and varied, a general repair mechanism is not advisable. In fact, any such mechanism is based on a mathematical model of the error and as long as the model does not reflect the nature of the mistake, modifying the data will not improve the quality. If the error model is known, the batch effect was corrected in the database (Fig. 3B). When no error models are available, we excluded suspicious parameters displaying huge batch problems with unexplainable noise from further analysis.

4. Data visualization and sharing

Monitoring and communication of data is an essential step for the successful completion of big data projects. A strategy to get fast and reliable results, which can be distributed under high safety conditions, is required. In the MARK-AGE project KNIME was used as standard tool for (1) data query (2) data visualization and (3) data distribution.

4.1. Data query

For the multifunctional analysis performed in the project different set of parameters and covariates must be requested regularly from the database. Pre-programmed KNIME nodes were used to generate a clear structured tool to filter desired conditions. Fig. 4 shows the standard view of the programmed selection menu for covariates. The user can chose the required group conditions from the assortment. In a second window a list of all available parameters appears. They can either be selected by partner number, work package number, or manually. If for special analysis more selection options were necessary they could easily be attached by the administrator. With this a powerful tool was generated and adjusted for

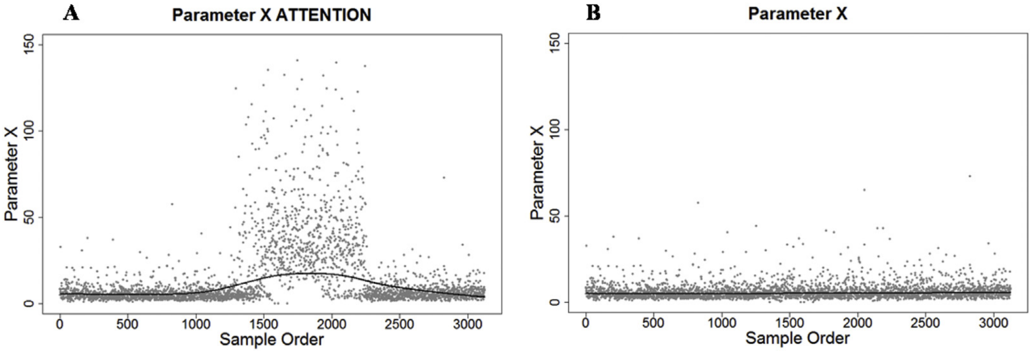
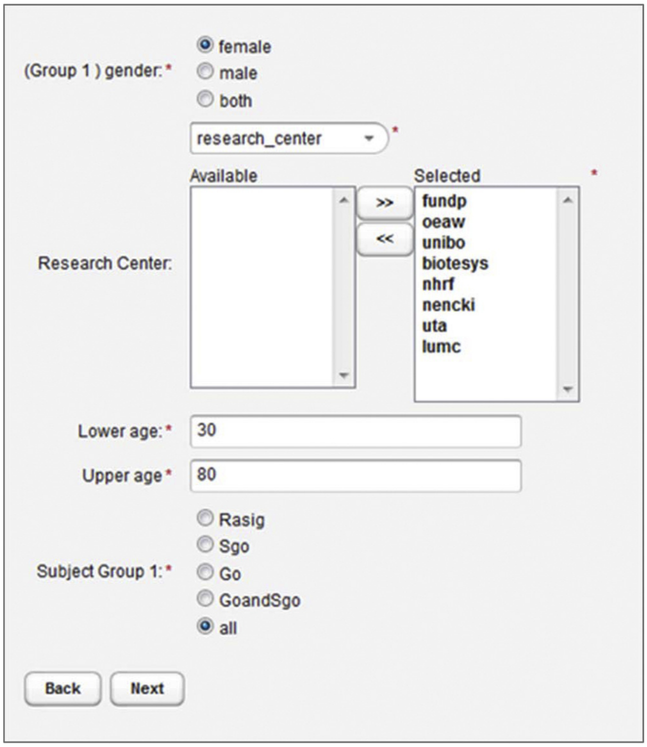


Fig. 3. A representative example for the successful correction of a batch effect. The graphs show scatterplots for a parameter x against the measuring order. The line of best fit (black) offers a slope which is recognized by the first deviation. As in such cases the algorithm automatically adds the term ATTENTION to the graph they can easily be selected from clean parameters. The batch in the example occurs for a defined time because the normalization of the variables was forgotten (A). This was documented by the staff in the respective laboratory. Therefore the normalization could be performed subsequently. After the calculation the parameter is free from any batch effect (B).



(Group 1) gender:* ● female ○ male ○ both

research_center *

Research Center: Available | Selected *: fundp, oeaw, unibo, biotesys, nhrf, nencki, uta, lumc

Lower age:* 30

Upper age* 80

Subject Group 1:* ○ Rasig ○ Sgo ○ Go ○ GoandSgo ● all

Fig. 4. Screenshot of the standard selection menu to query subgroups from the database. The desired conditions can easily be selected in one step.

the specific requirements on the MARK-AGE database for a fast and easy data query.

4.2. Data visualization

Visualization of data in the MARK-AGE project is important for two facts (1) uploaded data must be monitored and controlled for an early recognition and prevention of problems (2) researchers need to extract the biological information from the data.

4.2.1. Data exploration

For data exploration general plotting tools from descriptive statistics were used (Table 1). KNIME workflows were designed to automatically generate the graphs for all parameters available in the database. With those analysis experts in the field could test the data for plausibility and correctness. Also evidence for the underlying distribution and, as described above, for parameters quality could be provided.

4.2.2. Extract biological information

For the extraction of inter-parameter dependencies in the MARK-AGE Database, a tool to visualize correlations was necessary. A typically used numerical correlation matrix, calculated over a whole database, would have been too large for the extraction of significant information. Therefore a KNIME workflow was established automatically generating a network for a pre-selected parameter in the middle, showing the top ten correlating parameters arranged around (Fig. 5). To clearly visualize the dependency between the parameters, information is represented by the length and color of the connection lines, indicating the correlation strength and direction. The available sample number is suggested by the size of the parameter circles. Through this arrangement the user is able to perform plausibility checks for known dependencies at a glance. Thus efficient detection of unknown or rather unexpected dependencies in the data is possible. As an upstream selection menu allows the separation of different subgroups, and additional stratification on the data background is possible. Changes in the correlations between different subgroups give a hint for the interaction of specific body systems or the influence of environmental factors. The identified dependencies can lead to the development of new hypothesis and further detailed analysis.

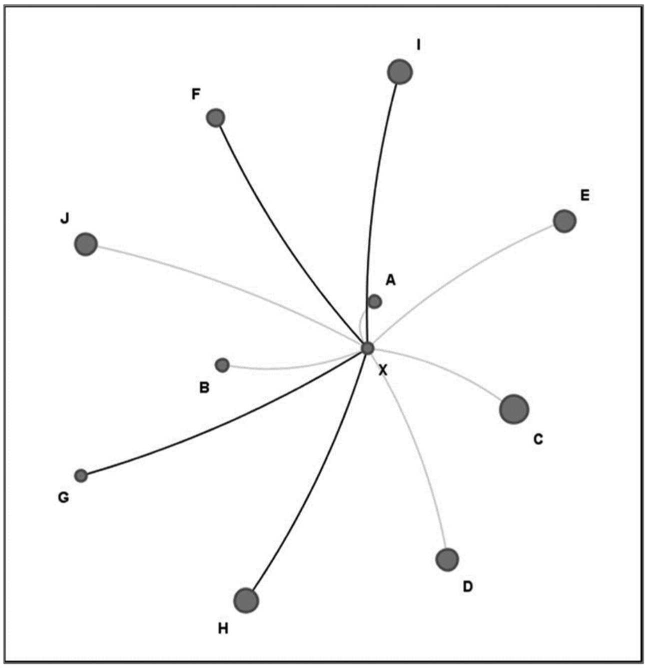


Fig. 5. Network analysis to check for correlations. The automated analysis presents the 10 best correlating parameters (A–J) for the parameter X, selected by the user. The color of the lines reflects the direction of the correlation (dark grey positive correlation, light grey negative correlation). The length of the lines indicates the strength of the correlation (the longer the lines the weaker is the correlation). The size of the circles stand for the amount of data considered for the calculation (the larger the diameter the more data is available). Which subgroups should be considered in the analysis is defined in the selection menu (see Fig. 4).

4.2.3. Data sharing

In order to provide MARK-AGE partners with graphical results based on the ongoing analysis the KNIME team space and the KNIME WebPortal were established.

4.2.3.1. KNIME team space. To share the database, metanodes and analysis workflows, the KNIME team space was used. Workflows were placed on a dedicated MARK-AGE server. Access to the flows was restricted to team members responsible for the coordination of the MARK-AGE analysis. These members could not only access the uploaded flows but also provide other colleagues with self-generated workflows. The workflows provided could be downloaded as a local copy Original versions, however, could only be modified by their authors. Thus, the KNIME team space provided a collaborative information exchange in a documented manner and under high-level security conditions.

4.2.3.2. KNIME WebPortal. The KNIME team space was restricted to members responsible for the coordination of MARK-AGE analysis. Therefore the KNIME WebPortal (Fig. 6) was used, enabling all project partners to receive analysis-feedback on own measurements. The KNIME WebPortal was connected to the MARK-AGE server and workflows were organized in specified folders. Users could select the desired analysis conditions from a menu (Fig. 4) and pre-programmed, automatic algorithms run in the background. Afterwards users could download graphical results either as pdf,

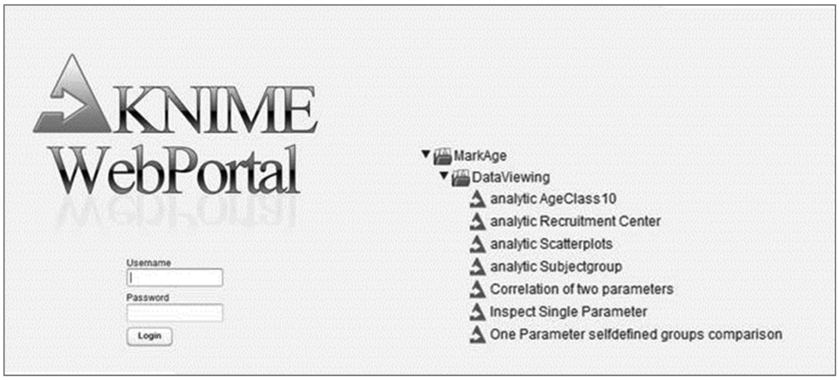


Fig. 6. Screenshot of the KNIME WebPortal user interface with the list of available analysis that could be selected. Only activated users can log in and receive analysis from the automatic reports.

xls, pptx, or word document (Fig. 7). This strategy ensures the blinding of MARK-AGE project, since the subjects codes linking the subject with the analysis results remain protected.

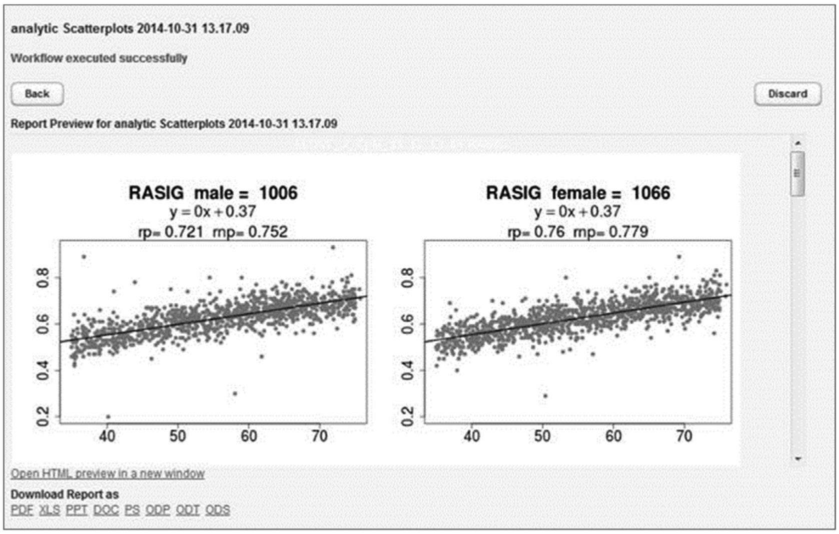
5. Discussion

In this paper we describe how the data cleaning and visualization processes were performed during the MARK-AGE project. Problems were identified especially with regards to outliers, missing values and batch effects. Some of our observations represent already known problems that can occur in databases, but published handling strategies cannot be used directly on MARK-AGE data. Therefore adjustments on the specific project's requirements are reported. In addition, new developments were described that could work for prevention and as prototype examples in upcoming ageing studies.

As mentioned under Section 2, the MARK-AGE database contains several types of information, i.e., (1) values from demographic data, (2) values from anthropometric measurements and (3) values from analytic data (see Bürkle et al., this issue). During data analysis a main challenge was to deal and prepare the different kind of data for reliable analysis. While information on subjects was collected in questionnaires during interviews, bioanalytical data were obtained from biological material analysed in the corresponding laboratories. Both, questionnaires and analytic data are error-prone and even our best efforts in the project design could not prevent such errors. Several circumstances can lead to missing values, outliers or batch effects, which in turn can significantly impact statistical analysis. Therefore strategies for cleaning data are necessary. However, there is no standard strategy available, and the procedures to be used it mainly depend on the type of data and type of analysis. In order to check the effectiveness of data cleaning strategies, Dasu and Loh introduced the concept of *statistical distortion*. They argued that data cleaning strategies could have an impact on results and no longer represent the real process that generates the data. Therefore, 'cleaner' data does not necessarily mean more useful or usable data (Dasu and Loh, 2012).

In case of missing values, most statistical procedures automatically exclude the subjects concerned. This leads to a reduction of data available for performing statistical analysis. As a result, outcomes may not be statistically significant due to lack of statistical power. Furthermore missing values can also cause misleading results by introducing bias. Due to the nature of the MARK-AGE project in some cases the reason for missing values was difficult to identify. Often it is not obvious if missing data will cause a problem. In some cases results might be affected, while others stay unchanged.

Like for missing values, variations that arise from outliers can influence the statistical outcome and the reliability of data. Most criteria identifying possible outliers are effective if data possess a


analytic Scatterplots 2014-10-31 13.17.09

Workflow executed successfully

Report Preview for analytic Scatterplots 2014-10-31 13.17.09

Fig. 7. A representative example for the outcome of an analysis on the KNIME WebPortal. The workflow runs in the background and provides the user with graphs and tables. Obtained results can be saved in one of the listed file formats (bottom left corner).

normal distribution. Although these methods are powerful, it may be problematic to apply them to non-normally distributed data or small sample sizes without information about their characteristics. Outliers can occur due to biological variance or technical reasons. In the MARK-AGE project the inter-quartile approach was used in order to eliminate the most severe cases. However, we might exclude the most interesting subjects that offer a real biological outlier. Unfortunately in some cases it is not possible to determine the background of such an outlier. By contrast, batch effects are a source of non-biological variation. Although there are methods available to fit different slopes or steps in a distribution this should not be applied blindly to each situation. Furthermore, during the analysis of batch effects it is essential to check whether outliers and batch effects offer overlapping problems.

Due to the huge amount and different type of data generated during the MARK-AGE project, a manual database exploration would have been time consuming. However automated data exploration tools cannot be applied in each case. We concluded that here is a need for useful and powerful tools that automate the data cleaning process, or at least assist manual procedures. However, automated methods can only be part of the procedure. There is a ongoing need for the development of new tools to assist this process, especially for the use in best practice routines.

Last but not least, data sharing is indispensable in large research consortia such as MARK-AGE. Budin-Ljøsne and colleagues describe the challenges of sharing data based on their experiences in the European Network for Genetic and Genomic Epidemiology, ENGAGE (Budin-Ljosne et al., 2014). The science community expects authors to share research results, therefore there is a need for data archiving especially when the research deals with health issues or public policy formation. Data archiving means the storing of large amounts of data to be accessed from central locations. During the MARK-AGE project the KNIME WebPortal was used to share results within the Consortium. Access to the WebPortal is restricted to MARK-AGE members but this could be opened to the scientific community in the future. If this happens it is necessary for each user to understand the underlying facts about the database development (see Baur et al., Kötter and Moreno-Villanueva et al., this issue) and possible quality issues described in this work. Especially for biogerontologists not only the Database opening but also upcoming publication on the MARK-AGE data will be of great interest. The interpretation of those results as well as the development of new hypotheses will mainly rely on the quality and the confiding work on the database itself. Information on the age-dependent parameters identified in the MARK-AGE project might also be linked with other databases like the HAGR (Tacutu et al., 2013). But even prior to that, the conclusions reached during the data cleaning and visualization steps in MARK-AGE can be used to prevent errors in upcoming studies investigating in the human ageing process.

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.

References

Batini, C., Cappiello, C., Francalanci, C., 2009. Methodologies for data quality assessment and improvement. ACM Comput. Surv. 41, http://dx.doi.org/10.1145/1541880.1541883

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. Studies in Classification, Data Analysis, and Knowledge. Springer-Verlag, Heidelberg-Berlin.

Budin-Ljosne, I., Isaeva, J., Knoppers, B.M., Tasse, A.M., Shen, H.Y., McCarthy, M.I., Harris, J.R., 2014. Data sharing in large research consortia: experiences and recommendations from ENGAGE. Eur. J. Hum. Genet. 22 (3), 317–321.

Chapman, A.D., 2005. Principles and methods of data cleaning – primary species and species-occurrence data. In: Version 1.0, Report for the Global Biodiversity Information Facility. Global Biodiversity Information Facility, Copenhagen.

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.

Dasu, T. and Loh J. M. (2012). Statistical Distortion: Consequences of Data Cleaning. Proceedings of the VLDB Endowment. Z. M. Özsoyoğlu. Istanbul, Turkey, VLDB2012. 5: 1674-1683.

Fan, W., Geerts, F., Xibei, J., Kementsietsides, A., 2008. Conditional functional dependencies for capturing data inconsistencies. ACM Trans. Database Syst. 33 (2), 1–48.

Hellerstein, J. M. (2008). Quantitative Data Cleaning for Large Databases. Survey for the United Nations Economic Commission for Europe. http://db.cs.berkeley.edu/jmh

Ihaka, R., Gentleman, R., 1996. R: a language for data analysis and graphics. J. Comput. Graph. Stat. 5 (3), 299–314.

Juran, J.M., Gryna, F.M., Bingham, R.S., 1974. Quality Control Handbook, 3rd edition. McGraw-Hill, New York, ISBN 0070331758.

Maletic, J.I., Marcus, A., 2000. Data Cleansing: Beyond Integrity Analysis. Massachusetts Institute of Technology, Boston, pp. 200–209.

Mayfield, C., Neville, J., Prabhaker, S., 2010. ERACER: a database approach for statistical inference and data cleaning. SIGMOD, 75–86.

Orr, K., 1998. Data Quality and Systems Theory. CACM 41 (2), 66–71.

Pearson, A. A. (2004). Relational Databases. Current protocols in bioinformatic. supplement 7: 9.4. 1-9.4.25.

Redman, T., 1998. The impact of poor data quality on the typical enterprise. CACM 41 (2), 79–82.

Schroeder, W., Martin, K., Lorensen, B., 2003. The visualisation toolkit. In: A System for Guided Data Repair. SIGMOD. Prentice Hall PTR, New Jersey, pp. 1223–1226.

Tacutu, R., Craig, T., Budovsky, A., Wuttke, D., Lehmann, G., Taranukha, D., Costa, J., Fraifeld, V.E., De Magalhaes, J.P., 2013. Human Ageing Genomic Resources: integrated databases and tools for the biology and genetics of ageing. Nucleic Acids Res. 41, D1027–D1033.

Yakout, M., Elmagarmid, A.K., Neville, J., Ouzzani, M., 2010. GDR: a system for guided data repair SIGMOD 10. In: Proceedings of the 2010 ACM SIGMOD International Conference on Management of data ACM, NewYork, pp. 1223–1226.

Yu, K., Salomon, R., 2010. Peptidedepot: flexible relaitonal database for visual analysis of quantitative proteomic data and integration of existing protein information. Proteomics 9 (23), 5350–5358.