European Journal of Clinical Nutrition (2011) 65, S87–S91

0954-3007/11
www.nature.com/ejcn

ORIGINAL ARTICLE

The MSM program: web-based statistics package for estimating usual dietary intake using the Multiple Source Method

U Harttig, J Haubrock, S Knüppel and H Boeing, on behalf of the EFCOVAL Consortium

Department of Epidemiology, German Institute of Human Nutrition, Potsdam-Rehbruecke (DIFE), Nuthetal, Germany

Background/Objectives: The Multiple Source Method (MSM) is a new statistical method for estimating usual dietary intake including episodically consumed foods on the basis of two or more short-term measurements such as 24-h dietary recalls. Optional information regarding habitual use or non-use of a food can be included as a covariate in the model estimating the intake, as well as a parameter for identifying consumers and non-consumers. The objective was to implement the MSM algorithms into an easy-to-use statistical program package.
Subjects/Methods: The implementation was realized as a web-based application using the Perl application framework Catalyst. As the engine for the statistical calculations, the R system was used. To allow simultaneous use of the program by different users, a multiuser system with a resource bag pattern design was implemented.
Results: We established a software program that implements the algorithms of the MSM and allows interactive usage of the method, using standard web technologies. The program is hosted on a website established at the DIFE and can be accessed at https://nugo.dife.de/msm. The communication between users and the program web site is encrypted, securing transmitted data against unauthorized use. Users can interactively import several data sets, define the analysis model, review and export results and graphs. The use of the program is supported by online help and a user guide.
Conclusions: The MSM website provides a program package that allows nutritional scientists to calculate usual dietary intakes by combining short-term and long-term measurements (multiple sources). It promotes simple access to the MSM to estimate usual food intake for individuals and populations.
European Journal of Clinical Nutrition (2011) **65,** S87–S91; doi:10.1038/ejcn.2011.92

Keywords: software; statistical data interpretation; food habits; internet; Perl

Introduction

Methodologies to estimate usual intake from 24-h dietary recalls (24-HDRs) face their limits especially when occasionally or rarely eaten foods are considered (Dodd *et al*., 2006). This is because many consumers do not consume all foods every day, leading to days of zero intakes if the 24-HDR happens to be on a non-consumption day. With two administrations of a 24-HDR, usual intake of such foods is difficult to estimate (Kipnis *et al*., 2009). On this account, an additional food frequency questionnaire that queries the propensity to consume a food over the past year is helpful. It provides valuable information together with repeated 24-HDRs to improve estimates of usual intake. Particularly, such information can be used to identify habitual users of a food for whom an estimate needs to be provided despite zero intake at the 24-HDRs. In addition, the Multiple Source Method (MSM) first estimates the habitual intake of an individual and thereafter derives the moments of the population distribution, such as mean, standard deviation, skewness and kurtosis, from these estimates. The MSM has been developed within the European Food Consumption and Validation (EFCOVAL) Project, an EU-funded collaborative project on dietary assessment methods, aiming to overcome the methodological issues (Dodd *et al*., 2006) when estimating usual dietary intake distributions and to provide a user-friendly interface for routine use by nutritional scientists. MSM assumes that the short-term measurement provided by the 24-HDR provides unbiased

Correspondence: Dr U Harttig, Department of Epidemiology, German Institute of Human Nutrition, Potsdam-Rehbruecke (DIFE), 14558 Nuthetal, Germany.
E-mail: harttig@dife.de

measurements for usual intake and that, after data transformation, these measurements can be decomposed into two independent components (inter- and intraindividual variation). The method itself comprises three steps (Haubrock *et al*., 2011). First, for each individual, the probability of consumption of a food on a day is estimated. Second, for each individual, the usual amount of food intake on days of consumption is estimated. Finally, the usual food intake on all days is calculated by multiplying the probability of consumption of a food with the usual amount of food intake on days of consumption. Although the primary aim of the MSM is to estimate usual intake distributions, the product of these two estimates can be considered as the estimate of individual usual intake.

In addition, the MSM allows to combine repeated short-term measurements such as 24-HDR data with frequency information from a long-term instrument such as a food frequency questionnaire. Covariate information that is thought to be related to the intake of foods can be included within the estimation steps. This enables the MSM to model a possible correlation between the probability of consumption and amount of consumption modeled.

Participants who were non-consumers of a food according to the long-term data (food frequency questionnaire) and who did not report consumption of this food in the short-term measurement (24-HDR) were defined as true non-consumers. For those true non-consumers, the probability of consumption as well as the food intake is set to zero. Only for consumers, the usual intake is estimated. Other short-term measurements such as 24-h protocols can also be used as long as they are collected on random days.

A comparison of the MSM with other methods of dietary intake is published in this issue (Souverein *et al*., 2011).

To make the new method easily accessible to researchers in the field of nutrition, the algorithms were implemented in a user-friendly web-based statistical program package, which will be reported upon in this paper.

Materials and methods

The MSM program was developed to provide a user-friendly interface for the algorithms of the MSM as part of the work plan in the EFCOVAL Project. The MSM program was designed as a web-based application and implemented with free software that did not require a commercial license and standard Internet technology. The software used for development and deployment was based on the previous experiences in our group.

The web program was implemented using the Perl language (http://www.perl.org). The program was created based on the established application framework Catalyst (http://www.catalystframework.org) and related Perl software packages from the Comprehensive Perl Archive Network repository (http://www.cpan.org).

The MSM algorithm, as described by Haubrock *et al*. (2011), was implemented with the R system (R Development Core Team, 2008), version 2.11.1, as the statistical engine. The communication between the R and the program was facilitated by using the 'Statistics::R' package (http://search.cpan.org/~bricas/Statistics-R-0.04). To support the concurrent use of the program by multiple users and the efficient distribution of computing resources, the program uses a simple multiuser system with a resource bag pattern (round robin) design for the R engine. This procedure allows the creation of multiple R processes (five per default but variable) and therefore concurrent calculations by the program for multiple users.

To implement features that enable and improve the interaction of a user with the program in the web browser, we used the JavaScript library The Yahoo! User Interface Library (YUI) in version 2.8 (http://developer.yahoo.com/yui).

The state of the user interaction, uploaded data and results are tracked and preserved by using sessions on the server side and cookies on the clients' browser side. This functionality, a standard procedure, is provided by the Catalyst application framework.

The deployment of the MSM program was done using a combination of an Apache web server (http://httpd.apache.org) and FastCGI via the mod_fastcgi module (http://www.fastcgi.com/drupal/node/25). This allows flexible deployment, as the MSM program runs persistent under FastCGI control in processes independently of the web server. The Apache web server provides handling of the outside communication and encryption, using the SSL certificate of nugo.dife.de (Figure 1).

Results

MSM program website

The development of the MSM software was based on a three-point concept: allowing easy access through a web-based program, flexible and interactive user interface, and the use of free software and standard software components. On the basis of this concept, we established the MSM program that

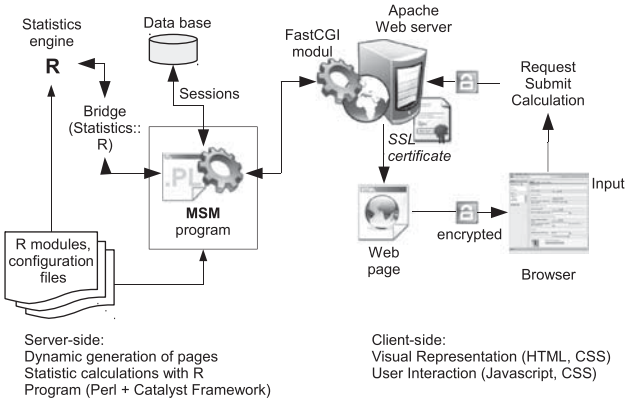

Figure 1 Structure of the MSM web application.

implements the algorithms of the MSM and allows interactive usage of the method. The choice of software tools used for development and the deployment of the application were based on our previous experience with these tools and restricted resources (free software).

The MSM program is a web-based application and can be used at the MSM website, which is hosted by the German Institute for Human Nutrition Potsdam-Rehbruecke (DIFE).

The website can be accessed through the address https://nugo.dife.de/msm/. The web server and the browser both encrypt the data sent between them, securing transmitted data and results against unauthorized access. This secure connection is specially marked by the browser. The alternative address http://nugo.dife.de/msm gets automatically redirected to the secure connection.

The website provides a short introduction and the link to the MSM program. Upon entry, the MSM program shows an overview with the four main steps of using the program (Figure 2) and gives access to the User Guide of the program. The four steps are as follows: data upload, analysis setup, calculation and results review.

In step one, 'Data upload', the user can import tabular data (character delimited text files with ASCII encoding) into the program system. The data structure requires at least one identifier column and a column with the response data with at least two data sets per individual. The data set can contain data for multiple analyses such as different food groups. In this case, the groups must be identified by a separate column. Additional data such as consumption frequencies and other individual covariate data related to the usual intake, such as age and gender, must be included in the same data set. In step two, 'Setup', the user defines the analysis model by specifying the response variables and the covariates contained in the data set. Information on consumption frequency from other long-term measurements such as food frequency questionnaires or external sources can also be specified. The model is submitted and the calculation is started in step 3. The duration of the analysis depends on the number of data sets, recalls per individual, the number of separate groups to be analyzed and the work load of the MSM server. To conserve computing resources of the web server, we impose a 60-min limit for calculations, which is sufficient to analyze large data sets. Table 1 lists the projected duration of analyses with different sized data sets. Calculations with multiple groups tend to run longer, as data preparation and generation of the output files are relatively more time consuming.

After the calculation is completed, the results can be reviewed in step 4. To assist with reviewing the results, the program shows univariate statistics and a plot of the resulting distributions on the 'Results' page. The univariate statistics reports the four moments, mean, standard deviation, kurtosis and skewness, as well as the 5th–95th percentiles of the distribution. The graph shows a density plot to illustrate the shape of the intake distribution with and without MSM analysis.

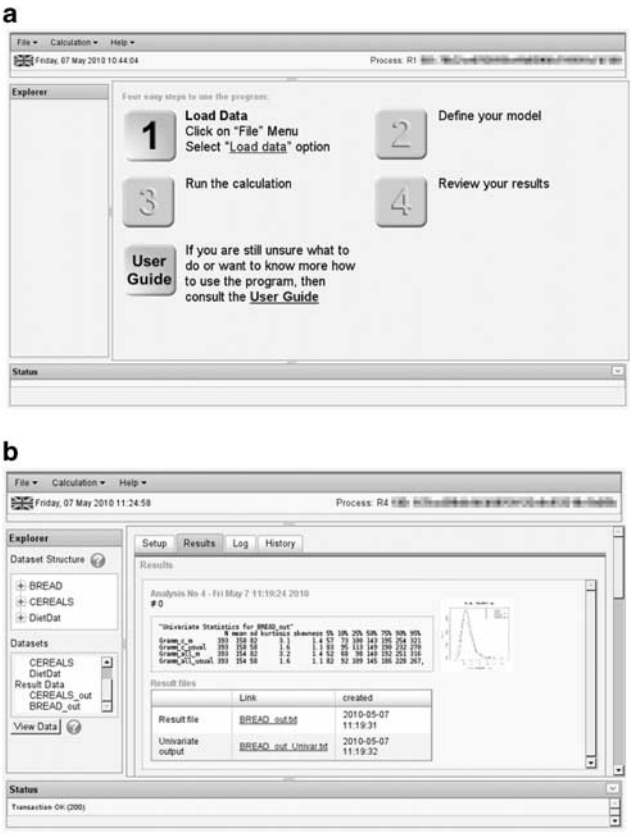

Figure 2 (**a**) Screenshot of the entry page of the MSM program. (**b**) Screenshot of the result page of the MSM program.

Table 1 Analysis of duration for differently sized data sets

Individuals	*Replicate measurements*	*Groups*	*Observations (lines of text file)*	*Duration (min)*
2000	7	1	14 000	0.4
10 000	7	1	70 000	2.0
20 000	7	1	140 000	5.1
50 000	7	1	350 000	21.2
2000	3	3	18 000	1.2
10 000	3	3	90 000	6.1
20 000	3	3	180 000	13.1

Abbreviation: MSM, Multiple Source Method.
Average duration of MSM analyses using the MSM web application. Duration numbers are mean of duplicate measurements in minutes.

Included on the output page are also some notes from the log file that indicate special procedures performed by the MSM, if the structure of the original data distribution make this necessary. The result files with the summary statistics and individual estimates can be downloaded for further processing. The individual estimates data file shows, per individual, the measured response on the recorded days, the consumer status used during analysis, the mean of response for consumption days, the mean of response for all days, the resulting intake estimate for consumption days and the

intake estimate for all days. The complete log file for the calculation can be reviewed on the 'Log' page or stored locally as a record of the analysis setup. During a session, the program stores the information regarding all result files produced so far. These can be accessed on the 'History' page. For an in-depth description of input and output of the program, we refer the reader to the MSM User Guide.

In this first version of the MSM web application, we do not provide a dedicated functionality for generating standard errors and confidence limits for the estimates. These standard errors could be generated by using bootstrapping procedures. For the future, we are planning to work on an extension of the program (batch job submission system), which will allow long running analyses that are required for bootstrapping.

Data security and internal data handling

Data security in the MSM program is provided by a three-step approach: (1) encryption to protect the data transfer between user and the program, (2) encapsulation to separate data from multiple users and (3) temporary storage for the duration of user interaction/session with the program.

The communication between a user's browser and the web server at the DIFE, which hosts the MSM program, is encrypted using the standard SSL/TLS protocol. This encryption prevents eavesdropping by third parties on data sent to and from the web server. To ensure that the web server with which the browser communicates is the correct one, the web server identifies itself by using a certificate issued by an established Certification Authority, which is checked by the browser.

The MSM program uses a session mechanism, a standard procedure for managing user communication with a web-based application or website. Using a unique session ID, all input and output data files are uniquely assigned to the correct user. This mechanism enables the program to encapsulate the data of a user and to separate different users from each other. Users can only see and analyze their own data, tagged with their own session ID, even if multiple users are using the program at the same time.

Uploaded data files and the resulting data files are stored only temporarily in designated directories to which only the program has access to. The storage is cleaned after the session has expired. Data sets used by the R statistical engine are stored only in the memory of the respective R process and are deleted after an R process has ended.

Visits to the website and its use are recorded by the web server using the IP of the requesting computer. But as we do not track the analyses run with the MSM program outside the session mechanism, individual analyses cannot be identified and connected to a specific user by an external observer.

Discussion

On the basis of the new statistical procedure for estimating individual and population-wide dietary intake, the MSM, we have created a web-based program package for general use that implements the MSM algorithms. Some of existing methods are also provided to the scientific community in software form. For the NCI method (Tooze *et al*., 2006), SAS macros are available (http://riskfactor.cancer.gov/diet/usualintakes/macros.html). For the ISU/Nusser method (Nusser *et al*., 1996), a standalone program package for personal computer called Software for Intake Distribution Estimation (PC-Side) is also available (http://cssm.iastate.edu/software/side.html). In case of the SAS macro, an SAS system with its license costs and familiarity with the SAS programming language is necessary but not always available to nutritional researchers. The Internet, and with it the world-wide web, has proven to be one of the easiest way of disseminating information and knowledge. As we intended not just to provide a MSM implementation but a simple access to the new method as service to nutritional scientists, we decided to use the web infrastructure. We achieved the goal to provide common and guided access to the MSM by implementing a web-based application that can be used by means of any modern browser. As browser software is ubiquitous, this approach guarantees easy access to the method. Hosting the MSM program on a server has also the advantage for the user that always the most recent version of the program is available, including new features that were not available at the start of the website. The transfer of data to the server can be perceived as a disadvantage of a web-based application because of the potential of misuse of this data. However, specific measures of a security setup of the application will minimize this risk substantially.

For the MSM calculations, we used the statistical programming language R. R was chosen as it is a free software environment for statistical computing and graphics, avoiding large licensing cost, which would be necessary if commercial software such as SAS would have been used. R runs on a wide variety of computing platforms and is widely used in computation of intensive areas of life science, such as bioinformatics. The Bioconductor Project for the analysis and comprehension of genomic data (Gentleman *et al*., 2004), for instance, is based largely on R software. The capabilities and openness of R has spawned a large user community with its own peer-reviewed journal (http://journal.r-project.org/).

The MSM algorithm was packaged into a web-based, interactive program that makes use of the Perl-based Catalyst application framework. This framework allowed rapid development of the MSM program within the human and financial resource constraints as many procedures that are necessary for web-based applications, such as user and session handling, are already available as components.

The MSM program website at DIFE (https://nugo.dife.de/msm) allows an interactive access to the features of MSM. It enables nutritional scientists to analyze their dietary assessment data with an advanced statistical method that combines short-term and long-term assessment data. The security and data protection setup of the program and the

website ensures the confidentiality of the data processed. The use of the program is supported by online help, a program user guide, as well as direct support in case of problems from the maintainer of the program.

The website and application are managed and maintained by the Department of Epidemiology of the German Institute for Human Nutrition Potsdam-Rehbruecke (DIFE). The development and use of the MSM is part of the analysis strategy of the departments for estimating dietary intake. Therefore, the work with the MSM and its further refinement will continue in the future. The MSM will be a topic in other projects such as the European Food Safety Agency (EFSA) Project ETUI (Usual intake estimation: statistical methods, data and criteria for application by EFSA, http://www.efsa.europa.eu/en/scdocs/scdoc/86e.htm). In this project, statistical experts will critically evaluate and review algorithms and implementations and provide scientific transparency. Therefore, the MSM website and program will continue to be improved and will be available to its users in the future. The developers also intend to add new features that increase the scientific value of the application, such as facilities to generate standard errors of the estimates.

We hope that the MSM program helps to promote the use of advanced statistical techniques to estimate usual dietary intake for individuals and populations, especially for foods with a sizeable proportion of non-consumers.

Conflict of interest

The authors declare no conflict of interest.

Acknowledgements

This work was created under the auspices of the EFCOVAL Consortium. UH was the author of the manuscript and was involved in Software development. JH was involved in the MSM Method development and implementation, as well as program testing. SK contributed to MSM Method evaluation and program testing. HB was involved in MSM Method creation and development. This document reflects only our views, and the Community is not liable for any use that may be made of the information contained therein. The MSM was conceived and developed by our colleague Dr Kurt Hoffmann who unexpectedly passed away in August 2007. All authors and colleagues remember him for his works and substantive contribution to this article respectfully. We also thank Wolfgang Bernigau for his invaluable help with implementing the MSM algorithms and its testing. The Community funding under the Sixth Framework Program for the EFCOVAL project is acknowledged (FOOD-CT-2006-022895).

References

Dodd KW, Guenther PM, Freedman LS, Subar AF, Kipnis V, Midthune D *et al*. (2006). Statistical methods for estimating usual intake of nutrients and foods: a review of the theory. *J Am Diet Assoc* **106**, 1640–1650.

Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S *et al*. (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* **5**, R80.

Haubrock J, Nöthlings U, Volatier JL, Dekkers AL, Ocke M, Harttig U *et al*. (2011). Estimating usual food intake distributions by using the Multiple Source Method (MSM). *J Nutr* **141**, 914–920.

Kipnis V, Midthune D, Buckman DW, Dodd KW, Guenther PM, Krebs-Smith SM *et al*. (2009). Modeling data with excess zeros and measurement error: application to evaluating relationships between episodically consumed foods and health outcomes. *Biometrics* **65**, 1003–1010.

Nusser SM, Carriquiry AL, Dodd KW, Fuller WA (1996). A semiparametric transformation approach to estimating usual daily intake distributions. *J Am Statist Assoc* **91**, 1440–1449.

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing: Vienna, Austria. ISBN 3-900051-07-0, http://www.R-project.org.

Souverein OW, Dekkers AL, Geelen A, Haubrock J, de Vries JH, Ocké MC *et al*. (2011). Comparing four methods to estimate usual intake distributions. *Euro J Clin Nutr* **65** (Suppl 1), S92–S101.

Tooze JA, Midthune D, Dodd KW, Freedman LS, Krebs-Smith SM, Subar AF *et al*. (2006). A new statistical method for estimating the usual intake of episodically consumed foods with application to their distribution. *J Am Diet Assoc* **106**, 1575–1587.