Quality, quantity and harmony: the DataSHaPER approach to integrating data across bioclinical studies

Published by Oxford University Press on behalf of the International Epidemiological Association *International Journal of Epidemiology* 2010;**39**:1383–1393

doi:10.1093/ije/dyq139

Quality, quantity and harmony: the DataSHaPER approach to integrating data across bioclinical studies

Isabel Fortier,[1,2]* Paul R Burton,[1,3] Paula J Robson,[4] Vincent Ferretti,[5] Julian Little,[1,6] Francois L'Heureux,[1] Mylène Deschênes,[1] Bartha M Knoppers,[1,7] Dany Doiron,[1] Joost C Keers,[8] Pamela Linksted,[9] Jennifer R Harris,[10] Geneviève Lachance,[1] Catherine Boileau,[11] Nancy L Pedersen,[12] Carol M Hamilton,[13] Kristian Hveem,[14] Marilyn J Borugian,[15,16] Richard P Gallagher,[15,16] John McLaughlin,[17] Louise Parker,[18] John D Potter,[19] John Gallacher,[20] Rudolf Kaaks,[21] Bette Liu,[22] Tim Sprosen,[23] Anne Vilain,[1] Susan A Atkinson,[3] Andrea Rengifo,[3] Robin Morton,[9] Andres Metspalu,[24] H Erich Wichmann,[25,26,27] Mark Tremblay,[28,29] Rex L Chisholm,[30] Andrés Garcia-Montero,[31] Hans Hillege,[32] Jan-Eric Litton,[33] Lyle J Palmer,[34] Markus Perola,[35,36] Bruce HR Wolffenbuttel,[37] Leena Peltonen[38] and Thomas J Hudson[5,39,40]

[1]Public Population Project in Genomics (P[3]G), Montreal, QC, Canada, [2]Research Center, University of Montreal Hospital Center (CHUM), Montreal, QC, Canada, [3]Department of Health Sciences and Department of Genetics, University of Leicester, Leicester, UK, [4]Division of Population Health and Information, Alberta Cancer Board, Edmonton, AB, Canada, [5]Ontario Institute for Cancer Research, MaRS Centre, Toronto, ON, Canada, [6]Department of Epidemiology and Community Medicine, University of Ottawa, Ottawa, ON, Canada, [7]Centre of Genomics and Policy, Faculty of Medicine, Department of Human Genetics, McGill University, Montreal, QC, Canada, [8]LifeLines, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands, [9]Generation Scotland, University of Edinburgh, Molecular Medicine Centre, Western, General Hospital, Edinburgh, UK, [10]Division of Epidemiology, The Norwegian Institute of Public Health, Oslo, Norway, [11]Cartagene, Montreal, QC, Canada, [12]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden, [13]PhenX Project, Research Triangle Institute, Research Triangle Park, NC, USA, [14]Department of Public Health, Norwegian University of Science and Technology, Trondheim, Norway, [15]Cancer Control Research, British Columbia Cancer Agency, Vancouver, BC, Canada, [16]Department of Health Care and Epidemiology, University of British Columbia, Vancouver, BC, Canada, [17]Samuel Lunenfeld Research Institute of the Mount Sinai Hospital, Toronto, ON, Canada, [18]Department of Medicine and Department of Paediatrics, Dalhousie University, Halifax, NS, Canada, [19]Cancer Prevention Program, Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA, USA, [20]Department of Epidemiology, Statistics and Public Health, Centre for Health Sciences Research, Cardiff University, Cardiff, UK, [21]Division of Cancer Epidemiology, German Cancer Research Center (DKFZ), Heidelberg, Germany, [22]Cancer Epidemiology Unit, University of Oxford, Oxford, UK, [23]UK Biobank, Stockport, UK, [24]Estonian Genome Project of University of Tartu, Tartu, Estonia, [25]Institute of Epidemiology, Helmholtz Zentrum München, Ludwig-Maximilians-Universität, Munich, Germany, [26]Institute of Medical Informatics, Biometry and Epidemiology, Ludwig-Maximilians-Universität, Munich, Germany, [27]Klinikum Grosshadern, Ludwig-Maximilians-Universität München, Munich, Germany, [28]Department of Pediatrics, Faculty of Medicine, University of Ottawa, Ottawa, ON, Canada, [29]Children's Hospital of Eastern Ontario Research Institute, Ottawa, ON, Canada, [30]The Center for Genetic Medicine, Robert H. Lurie Comprehensive Cancer Center, Northwestern University, Chicago, IL, USA, [31]Banco Nacional de ADN, Universidad de Salamanca, Fundacion Genoma España, Consejería de Sanidad de la Junta de Castilla y León, Spain, [32]Department of Cardiology, University of Groningen and University Medical Center Groningen, Groningen, The Netherlands, [33]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden, [34]Centre for Genetic Epidemiology and Biostatistics, University of Western Australia, Perth, WA, Australia, [35]National Institute for Welfare and Health, Helsinki, Finland, [36]Institute for Molecular Medicine Finland FIMM, University of Helsinki and National Public Health Institute, Helsinki, Finland, [37]Department of Endocrinology and Metabolism, University Medical Centre Groningen, University of Groningen, Groningen, The Netherlands, [38]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK, [39]Department of Medical Biophysics, University of Toronto, Toronto, ON, Canada and [40]Department of Molecular Genetics, University of Toronto, Toronto, ON, Canada

*Corresponding author. Public Population Project in Genomics (P[3]G), 3333 Queen Mary Road, Suite 590, Montreal H3V 1A2, Quebec, Canada. E-mail: ifortier@p3g.org

Accepted 13 July 2010

Background Vast sample sizes are often essential in the quest to disentangle the complex interplay of the genetic, lifestyle, environmental and social factors that determine the aetiology and progression of chronic diseases. The pooling of information between studies is therefore of

central importance to contemporary bioscience. However, there are many technical, ethico-legal and scientific challenges to be overcome if an effective, valid, pooled analysis is to be achieved. Perhaps most critically, any data that are to be analysed in this way must be adequately 'harmonized'. This implies that the collection and recording of information and data must be done in a manner that is sufficiently similar in the different studies to allow valid synthesis to take place.

Methods This conceptual article describes the origins, purpose and scientific foundations of the DataSHaPER (DataSchema and Harmonization Platform for Epidemiological Research; http://www.datashaper.org), which has been created by a multidisciplinary consortium of experts that was pulled together and coordinated by three international organizations: P^3G (Public Population Project in Genomics), PHOEBE (Promoting Harmonization of Epidemiological Biobanks in Europe) and CPT (Canadian Partnership for Tomorrow Project).

Results The DataSHaPER provides a flexible, structured approach to the harmonization and pooling of information between studies. Its two primary components, the 'DataSchema' and 'Harmonization Platforms', together support the preparation of effective data-collection protocols and provide a central reference to facilitate harmonization. The DataSHaPER supports both 'prospective' and 'retrospective' harmonization.

Conclusion It is hoped that this article will encourage readers to investigate the project further: the more the research groups and studies are actively involved, the more effective the DataSHaPER programme will ultimately be.

Keywords Data synthesis, data quality, data pooling, harmonization, meta-analysis, DataSHaPER, prospective harmonization, retrospective harmonization

Introduction

Scientific developments in the wake of the Human Genome[1–3] and HapMap[4,5] projects are helping to shape the future of public health and clinical medicine.[6–8] However, while dramatic progress has been made in detecting genetic associations with complex diseases,[9,10] the role of genetic determinants is only a part of a much larger picture. The role of lifestyle, environmental and social factors in modulating the risk and/or progression of chronic diseases has been recognized and explored for many years.[11,12] This is entirely logical even from the perspective of functional genomics: the concept of 'fitness' that is central to natural selection and human evolution has, as its fundamental basis, the interaction between prevailing environment and the genome.[13] This implies that causal pathways leading to disease should be 'expected' to involve gene–environment interactions.[14,15]

It is therefore clear that bioscience needs access to studies that incorporate social, environmental and lifestyle factors as well as genetic determinants.[7,15] Provided that the quality of the information that such studies generate is adequate and that the statistical power of key analyses can be rendered sufficient,[15–18] it will then be possible to successfully pursue a comprehensive investigation of the direct and interactive effects of a broader range of relevant classes of aetiological determinants. However, in the real world, the attainment of adequate statistical power presents a serious challenge. For example, when appropriate account is taken of assessment errors in both determinants and outcomes, sample-size estimates for analyses involving gene–environment interactions comparable in magnitude with the direct genetic effects that have so far been replicated,[9,10,16] typically indicate a requirement for 'tens of thousands of cases'.[16,17,19] This means that even the largest[16,20–22] and best measured[18,23] of contemporary studies will only be able to generate enough cases—or subjects—for the commonest of complex diseases.[16,22] This in turn implies that the analysis of synthesized data across several studies is set to become increasingly important.[15,24]

Such harmonization may be used to support targeted scientific projects,[25–27] and to facilitate synthesis of information among studies[28–34] or data portals.[35–39]

Fortunately, extensive experience already exists in the synthesis of epidemiological studies.[33,40–42] For example, data synthesis was pivotal to the success of the EPIC study (the European Prospective Investigation into Cancer and Nutrition) which starting in the 1990s, recruited more than 500 000 participants via (initially) 22 centres across nine European countries.[28,43] EPIC's focus on nutrition placed heavy demands on sample size, and effective data synthesis across all centres was therefore critical to many of its principal analyses. Although EPIC was designed prospectively as a coordinated consortium of studies, centre-specific questionnaires were used.[28,44] In such a setting, the data synthesis was constrained by the quality[18] of the underlying data and by their compatibility.[45] One of the important achievements of the EPIC project was the development of methods and tools (e.g. EPIC SOFT[43]) to enable calibration and pooling of data that had been collected under different protocols in different centres, so that data synthesis was rendered valid.

However, in common with other major epidemiological consortia—e.g. GenomEUtwin project[30] and EURALIM project[41]—EPIC demonstrated that information synthesis is far from easy. It demands time, resources and rigour.[40,43,44,46] Furthermore, as scientific ambitions and capacities have extended, the sample-size challenge continues to grow,[9,15–18] and the requirement for effective data synthesis has now become a regular necessity.[15,24] Moreover, as different sets of outcome and exposure variables are required for different analyses—and no single study can afford to capture 'all' desired measures—individual studies are necessarily being pooled with different combinations of other studies—as demonstrated, e.g. by the number of different consortia involving studies such as Avon Longitudinal Study of Parents and Children (ALSPAC), EPIC-Norfolk and the 1958 Birth Cohort.[25,26,47,48] This implies that it would be beneficial to supplement consortium-specific approaches to harmonization, calibration and synthesis[29,30,34,41,43] with more generic methods.[49–52]

However, the scientific utility of data synthesis is always constrained by the quantity and quality of the underlying data,[18,53] and by their compatibility between studies.[45,54] The latter implies that the collection and recording of information and data must be carried out in a manner that is sufficiently similar in the different studies to allow valid synthesis to take place. When this is so, 'harmonization' may be said to exist.[53] The fundamental challenge might therefore be viewed as being to increase sample size by synthesizing over an adequate number of studies, but to restrict that synthesis to those studies that are satisfactorily harmonized for the specific outcomes, genetic, environmental and lifestyle factors targeted.[42,54] Two complementary approaches may be adopted to support effective data synthesis. The first one principally targets 'what' is to be synthesized, whereas the other one focuses on 'how' to collect the required information. Thus: (i) core sets of information may be identified to serve as the foundation for a flexible approach to harmonization[51,52,55–57]; or (ii) standard collection devices (questionnaires and standard operating procedures) may be suggested as a required basis for collection of information.[49,58–61]

It is with all of these considerations in mind that the DataSHaPER project (DataSchema and Harmonization Platform for Epidemiological Research) has been launched. The DataSHaPER (http://www.datashaper.org) offers free access to questionnaires and core sets of variables that can be used to support the development of data-collection tools for emerging studies or to serve as a central reference for harmonization between pre-existing studies. The DataSHaPER is an international project that is being developed under the joint umbrellas of P^3G (the Public Population Project in Genomics[50,62]), PHOEBE (Promoting Harmonization of Epidemiological Biobanks in Europe[63]) and CPT (Canadian Partnership for Tomorrow[64]), in collaboration with more than 50 major studies from around the world. This conceptual article describes the motivation, aims and scientific foundation of the DataSHaPER project.

Harmonization

Standardization is a *sine qua non* of information pooling. However, scientific, technological, ethical, cultural and other constraints make it difficult to impose identical infrastructures and uniform procedures across studies. Furthermore, it is important to recognize that it is not always necessary to use precisely the same methods and tools for data collection in order to achieve valid data integration across studies. Rather, what 'is' crucial is that the information conveyed by each data set is 'inferentially equivalent'. If the 'quality' of the data to be integrated is also adequate, inferential equivalence greatly increases the potential for collaboration between studies and, therefore, the scientific opportunities. The definition of equivalence will vary with the scientific context and must take into account both the primary information collected (e.g. serum cholesterol level) and the qualifying factors that can influence the interpretation of that information (e.g. whether the participant had been fasting prior to sample collection). In some situations, even a small change in the way information is collected can substantially modify scientific compatibility, whereas in others, considerable flexibility can be allowed. Formally, a valid balance must be struck between the use of precisely uniform specifications that render pooling straightforward (e.g. identical questions asked under identical conditions), and the

acceptance of greater flexibility and diversity that may be appropriate and more realistic in a collaborative context (e.g. similar questions, but asked by an interviewer in one study and completed by the participant in another).

In an ideal world, information would be 'prospectively harmonized': emerging studies would make use, where possible, of harmonized questionnaires and standard operating procedures.[53,65] This enhances the potential for future pooling but entails significant challenges—ahead of time—in developing and agreeing to common assessment protocols. However, at the same time, it is important to increase the utility of existing studies by 'retrospectively harmonizing' data that have already been collected, to optimize the subset of information that may legitimately be pooled.[30,53,65] Here, the quantity and quality of information that can be pooled is limited by the heterogeneity intrinsic to the pre-existing differences in study design and conduct.

The DataSHaPER

Concept

The DataSHaPER is both a scientific approach and a practical tool. Originally, the plan had been to develop a standardized questionnaire (or set of questionnaires) with the primary aim of facilitating prospective harmonization of future biobanks. But after some months of work, it became clear that complete standardization was too restrictive and would be of limited applicability to retrospective harmonization. This resulted in a fundamental change of approach that led to the piloting of the concept that is now known as the DataSHaPER. In order to understand the DataSHaPER, an important distinction must be drawn between core 'variables'—the primary units of interest in a statistical analysis—and the specific 'assessment items' that are collected by individual studies (e.g. questions in questionnaires). It is a pre-defined set of 'variables' that serves as the reference for harmonization between studies. This approach provides an appropriate level of flexibility, because a given variable may potentially be constructed using different assessment items in different studies. It is important to note that this does not imply a reduction in scientific rigour; the specific information collected by a given study can only be viewed as harmonized to a particular DataSHaPER variable if the assessment items in that study can be used to generate a 'valid' equivalent to the required variable. This entails a formal scientific evaluation and validation process.

Structurally, the DataSHaPER is a dynamically evolving entity that is built upon two primary components: the DataSchema Platform and the Harmonization Platform. The former incorporates and documents sets of core variables. The latter reflects a step-by-step approach that facilitates estimation of the potential for harmonization and pooling between studies for defined scientific purposes. The web-based application was developed by the DataSHaPER team with the support of experts in ontologies and open-source software. It is written in Java and uses open-source libraries (Spring Framework, Hibernate, Google Web Toolkit, Sesame/Elmo, etc.). The user interface makes extensive use of Asynchronous JavaScript and XML (AJAX) technologies, which provide for enhanced usability. Wherever possible, standard formats and ontologies are used. Thus, key information bearing objects (e.g. DataSchemas and component elements of the Harmonization Platform) are stored using a recognized ontological format[66] to facilitate exchange with other applications. Where relevant, the Generic DataSchema makes use of terms defined in the National Cancer Institute Thesaurus ontology (published on the National Center for Biomedical Ontology BioPortal).[67]

Access to the DataSHaPER application and content is open and free (http://www.datashaper.org). The public website presents the published DataSchemas and offers links to their ontology files. However, to access the DataSchemas under development as well as the results generated by the pre-existing Harmonization Platforms, users need to respond to specific criteria, be authenticated by the DataSHaPER Team and use a username and password.

DataSchema Platform

A DataSchema is a hierarchical structure composed of variables nested within domains, themes and modules (Figure 1). Each DataSchema on the Platform is made up of variables that may be derived from: interview administration; health and risk-factor questionnaires; physical and cognitive measures; medical files; sample collection, handling, processing and banking; biochemical measures, registries (e.g. databases containing deaths, hospitalization episodes and environmental variables) and others. Variables may be of primary scientific interest in their own right or qualifying factors that contribute to the interpretation of other information of primary interest. A variable may be complete in itself [e.g. current smoker (yes/no) or measured weight] or it may derive from one or several others (e.g. body mass index).

The DataSchema platform on the DataSHaPER website contains a comprehensive description of available schemas. Each may include: a list of variables with their definitions and formats; links to relevant ontologies; and access to reference questions/questionnaires, indexes and operating procedures that have been selected[58,68,69] or developed.[70] Where possible, variables have been defined such that they can reliably be constructed from standard questionnaires and classifications (e.g. The International Physical Activity Questionnaire for physical activity[58]). Although a

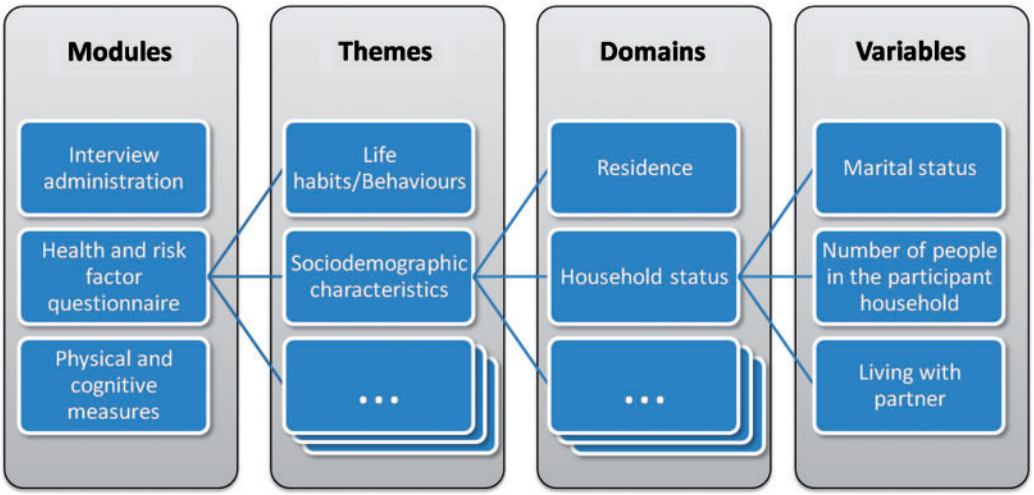

Figure 1 Hierarchical structure of the module, theme and variables related to the 'household status' domain in the Generic DataSchema

DataSchema aims primarily to provide a template for prospective and retrospective harmonization, it also provides a guide to help emerging projects select suitable assessment items and sample collection tools, even when data pooling is not planned.

The 'Generic DataSchema' is the first schema to have been developed under the DataSHaPER project. It is aimed at supporting the construction of general-purpose baseline questionnaires for use in large cohorts enrolling middle-aged participants. Its construction was a collaborative effort involving investigators from more than 25 international cohorts in 14 countries. Its structure and contents were determined at a series of international consensus workshops held over 2 years (2006–08) with iterative rounds of comments and feedback between meetings. The contents were chosen so as to provide a core data set with broad international applicability. Ethnic and cultural specificity was therefore minimized and the schema was chosen so as to be simple enough to encourage widespread use, yet comprehensive enough to support meaningful research. Detailed selection criteria for individual variables are listed in Box 1.

The Generic DataSchema contains 3 modules, 13 themes, 45 domains and more than 180 variables. As an illustrative example of its content (Figure 1), the theme 'sociodemographic characteristics' contains the domain 'household status' (defined as a social unit comprised of one or more individuals living together in the same dwelling, all of whom need not be related) which in turn includes three variables: (i) 'marital status (currently married; yes/no)'; (ii) 'living with a partner in a common household (yes/no)'; and (iii) 'number of people who live with the participant in the same household (number)'.

Box 1 Criteria for selecting individual variables in the Generic DataSchema

(i) The variable is of substantial relevance to genomic or public health research.
(ii) The variable may potentially be used for a wide range of research questions in a variety of populations.
(iii) Each level of response to a categorical variable is of high enough prevalence to ensure that sufficient power can potentially be obtained.
(iv) The assessment items required to generate the variable can be obtained in a valid way and reliably be assessed, where possible using standard questions and/or indexes.
(v) The assessment items required to generate the variable can be collected in a manner that entails no major burden to participants.
(vi) The assessment items required to generate the variable can be collected at acceptable cost.
(vii) A variable may be selected if it is of primary interest in its own right, is a qualifying variable that modifies the interpretation of other variables, or is viewed as being a potentially important confounder, or indicator of potential bias.

A fundamental aim was to restrict the Generic DataSchema to a limited number of variables identified as key by consensus.

Early versions of the Generic DataSchema were used by several large population-based studies to help create their data-collection tools. These included the LifeLines[71] (The Netherlands) and LifeGene[72] (Sweden) Projects as well as the five cohorts in the CPT Project[64] and the Canadian Longitudinal Study on Aging.[73]

Harmonization Platform

It is the Harmonization Platform that enables a DataSchema to be used as a basis for harmonization in a specific scientific context. It provides a rigorous approach to a three-step process that entails: (i) the development of rules providing a formal assessment of the potential for each individual study to generate each of the variables in the DataSchema; (ii) the application of these rules to determine and tabulate the ability of each study to generate each variable, thereby identifying the information that 'can' be shared; (iii) where a variable can be constructed by a given study, the development and application of a processing algorithm enabling that study to generate the required variable in an appropriate form.

The compatibility of variables is formally assessed on a three-level scale of matching quality: 'complete', 'partial' or 'impossible' (e.g. see Table 1). This process is referred to as 'pairing'. Rules generated for variable pairing are context specific and will vary according to each harmonization project. Rule creation and pairing are both systematic processes based on protocols involving iteration between domain experts, research assistants and a validation panel. The whole procedure is subject to appropriate quality assurance.

The first use of the Harmonization Platform was in association with the Generic DataSchema. Pairing rules were therefore developed for all the variables in that schema. As an illustrative example, Table 2 details the rules created for the variable 'Current quantity of red wine consumed (number of glasses of red wine/week)'. Using such pairing rules, the potential to harmonize 50 large population-based studies (each including at least 10 000 healthy participants) has now been explored for 'all' variables in the DataSchema: additional studies joined the collaboration to enable this formal evaluation to take place. In combination, these 50 collaborating studies have recruited or plan to recruit a total of approximately 5.4 million participants.

The detailed results of the full pairing analysis will form the basis of a second paper to follow. For the purposes of the present conceptual article, we will therefore do no more than provide a brief illustration of the nature of the results to be anticipated. For example, using the specific variable considered in Table 2 ('Average number of glasses of red wine consumed by the participant per week'), 7 (14%) of the 50 studies generated a complete match, 3 (6%) a partial match and 38 (76%) an impossible match. In the particular case being considered, therefore, information from approximately 873 900 participants might potentially be co-analysed for the variable of interest; i.e. from those studies that provide a complete match. In contrast, when the variable 'Current quantity of wine consumed' was considered (with no specification of red or white wine), 21 (42%) studies provided a complete match (1.8 million participants). As another example, when the variable 'measured weight' was investigated, 36 (72%) studies (3.6 million participants) provided a complete match. According to the pairing rules in this setting, in order that it might be considered a 'complete match', the weight of the participant had necessarily to be 'measured' at least once by a trained nurse/interviewer with a standard device. Where, weight was 'reported' by the participant, it was viewed only as a 'partial match'.

However, in order to answer a real scientific question, the pairing statuses of more than one variable must usually be considered simultaneously. For example, if harmonized information is required on 'Current quantity of wine consumed', 'Body Mass Index' and 'Current Tobacco Smoker', a total of 12 studies provide a complete match for all three variables (approximately 1 million participants). At the same time, additional issues must also be taken into account. These include ethico-legal constraints on access to data or samples, the compatibility of different study designs and protocols, and the distribution of missing data. Consideration of such issues is fundamental to scientific rigour in using the DataSHaPER.

Discussion

The DataSHaPER was originally launched under the P^3G[50,62] and PHOEBE[63] initiatives in response to requests from the members of both consortia for guidelines to support the construction of questionnaires to facilitate prospective harmonization of large population-based studies. But the overall focus evolved, and rapidly subsumed the critical need for tools to support retrospective synthesis of information between existing/legacy studies. As the nascent project progressed, it became clear that one of the primary needs of the scientific community was to have access to comprehensive documentation of the potential to synthesize data across subgroup of studies. It was also recognized that such documentation needed

Table 1 Example of pairing results for the variables under the domain 'Individual history of diabetes'

Individual history of diabetes	Study A	Study B	Study C	Study D	Study E
Occurrence of diabetes	Complete	Complete	Complete	Complete	Complete
Type of diabetes	Impossible	Impossible	Complete	Partial	Impossible
Onset of diabetes	Partial	Complete	Complete	Complete	Impossible

Table 2 Pairing rules used for the variable 'Current quantity of red wine consumed' and specific questions asked by exemplar studies

Current quantity of red wine consumed
(http://www.datashaper.org/Datashaper.html#dataschemasTab$RELEASE$GENERIC_1$GENERIC_2$GENERIC_14$GENERIC_55$GENERIC_266)
Definition: average number of glasses of red wine consumed by the participant per week; *Unit:* glasses of red wine/week; *Format:* open; *Type:* integer

Complete: *the meaning and format of the question or questions of the questionnaire allow the construction of the variable as described (definition and format)*
Rules specific to 'Current quantity of red wine consumed':
- the number of drinks can be collected per day, week, month, etc;
- the size (millilitre, ounces, etc.) of the glasses or drinks can be specified or not;
- the information can be collected for the entire week or for specific periods covering the whole week (week-days and week-end; Monday to Sunday, etc);
- the question must target the current consumption (over the past 12 months or more contemporaneously).

Exemplar question: Study A
In a typical week, how many glasses of red wine (6 ounces) do you drink per day?
[___] Number of drinks per day

Exemplar question: Study B
In general, how many glasses of red wine do you drink per day over a week and weekend?
Week: [___] Number/day
Weekend: [___] Number/day

Partial: *the meaning and the format of the question or questions of the questionnaire could allow the construction of the variable as described, but with an unavoidable loss of information*
Rules specific to 'Current quantity of red wine consumed':
- partial if categories are used instead of continuous variables.

Exemplar question: Study C
In a typical week, how many glasses of red wine do you drink per day?
☐ 1–3
☐ 4–6
☐ 7–9
☐ 10 or more

Impossible: *there is no information or insufficient information in the questionnaire to allow the construction of the variable as described*
Rules specific to 'Current quantity of red wine consumed':
- impossible if only wine is mentioned without distinction between types of wine (red, white);
- impossible if relevant information is collected only for the consumption in the past (before the past 12 months);
- impossible if relevant information is collected at the same time for the current and the past consumption without distinction between the two.

Exemplar question: Study D
In a typical week over the past 3 years, how many drinks of wine did you drink per day?
[___] Number of drinks per day

Exemplar question: Study E
Over the last 12 months, how many glasses of beer, wine and/or spirits altogether did you usually drink?
Total Number of glasses per day:
☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6
☐ 7 ☐ 8 ☐ 9 ☐ 10 or more

to include descriptions of the procedures used to collect data and to target both generic and specialized data collected by studies using various designs. The network of the DataSHaPER collaborators has therefore extended over time and now includes, e.g., scientists working in disease-oriented networks of studies such as Genecure (chronic kidney diseases).[74] Clearly, the ongoing development of new DataSchemas and Harmonization Platforms will reflect the interests and needs of the scientific teams using and developing them. As illustrative examples, future DataSHaPERs may focus on particular conditions (e.g. stroke, type 2 diabetes), social and lifestyle factors (e.g. nutrition, environmental pollutants), or specific population subgroups (e.g. newborn, elderly).

Documenting the potential to synthesize information across studies is critical and should foster collaboration, but it is only a step in the process leading to the final statistical analyses making use of synthesized data sets. In its recent development, the structure and web interface of the DataSHaPER is thus being consolidated in order to facilitate complementarity with other tools and approaches to harmonization, data access, processing, pooling and analysis (e.g. PhenX[49], dbGaP[36], DataSHIELD,[75] OBiBa[76] and SAIL[77]). It is the access to such integrated suites of tools that will ultimately facilitate the generation of new scientific discoveries using large-scale synthesized data sets across networks of studies.

The question 'What would constitute the ultimate proof of success or failure of the DataSHaPER

Box 2 Development of a DataSHaPER: how to ensure rigour of the process and formal validation of the outcomes

DataSchema

Process—the development of a DataSchema should be scientifically driven and based upon iterative review and consensus methodologies.[78,79]

Outcome (list of core variables)—the structure and specific content of each schema has to be formally evaluated. This involves, e.g., the assessment of the content by external groups of experts and systematic comparison with current practice or relevant gold standards and guidelines. Furthermore, for the schemas underpinning data-collection devices (e.g. questionnaires), formal validation of these resultant tools should be undertaken.

Harmonization Platform

Process—methodical quality control should be implemented through all of the harmonization process. This should include systematic validation of the pairing rules that have been developed and analysis of the agreement between the pairing classification achieved by different staff and an independent control panel.

Outcome (database including pairing results)—formal assessment of the impact of participant studies and variables characteristics on the pairing results should be undertaken to define factors influencing the potential for synthesis. Factors targeted include: (i) study characteristics such as design of the study (e.g. cohort or case–control study), nature of the population (e.g. minimal age at recruitment, sex, place of residence) and the procedural methods used to collect information (e.g. paper- or computer-based questionnaire); and (ii) variable characteristics such as whether the variable has a quantitative or a categorical format, and whether the information defining a particular variable relates to the participant or to his/her family.

Final set of synthesized 'data'

As for any database generated by a stand-alone study, once it has been created, the final product of the harmonization process (a synthesized database including data from all participating studies) should be subjected to standard data validation procedures including appropriate range checking and tests of internal validity.

Box 3 Tools provided by the DataSHaPER (http://www.datashaper.org)

For emerging studies

– Lists of core variables useful in the development of information collection tools relevant in specific scientific contexts.
– Exemplar questionnaires and standard operation procedures enabling collection of these core variables.

For network of studies to be prospectively or retrospectively harmonized

– A scientific method, web-based platform and provision of expertise for: (i) the definition of core set of variables to be shared; and (ii) the development and application of the harmonization process.

approach' needs to be addressed. Such proof will necessarily accumulate over time, and will involve two fundamental elements: (i) ratification of the basic DataSHaPER approach; and (ii) confirmation of the quality of each individual DataSHaPER as they are developed and/or extended. An important indication of the former would be provided by the widespread use of our tools. However, the ultimate proof of principle will necessarily be based on the generation of replicable scientific findings by researchers using the approach. But, for such evidence to accumulate it will be essential to assure the quality of each individual DataSHaPER (see Box 2). Even if the fundamental approach is sound, its success will depend critically on how individual DataSHaPERs are constructed and used. It seems likely that if consistency and quality are to be assured in the global development of the approach, it will be necessary for new DataSHaPERs to be formally endorsed by a central advisory team.

The novelty of the DataSHaPER is not in the scientific challenges or solutions being addressed and proposed: similar projects have been embarked upon before. However, the DataSHaPER provides access to useful tools (see Box 3) and has critical advantages. The approach aims to be generic, flexible and can be used both prospectively and retrospectively. Furthermore, the web interface can easily be updated as new DataSchema and Harmonization Platforms are added and thus, provides a good potential for constant improvement of the content. Finally, the DataSHaPER has emerged as a common approach to

the concrete need to document the potential to synthesize data across biobanks and cohort studies. However, the scientific utility of any synthesized data set depends on the quality of data to be pooled and on the rigour of the harmonization and synthesis process. The DataSHaPER can make a valuable contribution. However, if it is to be successful, it must continue to evolve and it must be used both widely and wisely.

Funding

Genome Canada and Genome Quebec (The Public Population Project in Genomics); Canadian Partnership Against Cancer (CPT); European FP6 (LSHG-CT-2006- 518418 to Promoting Harmonization of Epidemiological Biobanks in Europe); Medical Research Council Project Grant (G0601625; methods programme in genetic epidemiology at the University of Leicester that focuses on genetic statistics and large-scale data harmonization and pooling); Wellcome Trust Supplementary Grant (086160/Z/08/A); Leverhulme Trust Research Fellowship (RF/9/RFG/2009/0062); National Institute for Health Research (Leicester Biomedical Research Unit in Cardiovascular Science); German Federal Ministry of Education and Research (BMBF) in the context of the German National Genome Research Network (NGFN-2 and NGFN-plus) (to E.W.); German Federal Ministry of Education and Research (BMBF) (Model attempt for networking in German research consortiadevelopment of a common concept for biobanks); European Framework 7 (Biobanking and Biomolecular Resources Research Infrastructure); J.L. is a Canada Research Chair in Human Genome Epidemiology.

Acknowledgements

We would like to thank all of the additional studies and biobanking experts that provided advice and information on the development of the DataSHaPER, and are now part of the ongoing collaboration that is taking the DataSHaPER project forward. The funders had no role in study design, data collection and analysis, the decision to publish, or preparation of the manuscript.

Conflict of interest: None declared.

KEY MESSAGES

- Large-scale data pooling and meta-analysis are central to modern bioscience.
- If the data from two studies are sufficiently similar for a valid synthesized analysis, the two studies may be said to be harmonized in the particular scientific context that applies.
- The DataSHaPER (DataSchema and Harmonization Platform for Epidemiological Research) provides a flexible, but structured, approach to harmonization and data synthesis.
- A DataSchema provides a selected set of core variables to be shared between studies while the Harmonization Platform contains rules that determine whether the particular data items collected by a given study can be used to create each DataSchema variable.
- The DataSHaPER may be used prospectively, as a source of harmonized questions for new studies, or retrospectively as a structured framework for harmonizing existing/legacy studies.
- Access to the DataSHaPER application and content is open and free through its public website at: http://www.datashaper.org/. To access the Harmonization Platform (for retrospective harmonization), users must register with the DataSHaPER Team.

References

[1] Venter JC, Adams MD, Myers EW *et al*. The sequence of the human genome. *Science* 2001;**291:**1304–51.

[2] Lander ES, Linton LM, Birren B *et al*. Initial sequencing and analysis of the human genome. *Nature* 2001;**409:** 860–921.

[3] Little J, Khoury MJ, Bradley L *et al*. The human genome project is complete. How do we develop a handle for the pump? *Am J Epidemiol* 2003;**157:**667–73.

[4] The International HapMap Project Consortium. The International HapMap Project. *Nature* 2003;**426:**789–96.

[5] Altshuler D, Brooks L, Chakravarti A, Collins F, Daly M, Donnelly P. A haplotype map of the human genome. *Nature* 2005;**437:**1299–320.

[6] Collins FS. Shattuck lecture—medical and societal consequences of the Human Genome Project. *N Engl J Med* 1999;**341:**28–37.

[7] Collins FS, Green ED, Guttmacher AE, Guyer MS. A vision for the future of genomics research. *Nature* 2003; **422:**835–47.

[8] Davey Smith G, Ebrahim S, Lewis S, Hansell AL, Palmer LJ, Burton PR. Genetic epidemiology and public health:hope, hype, and future prospects. *Lancet* 2005;**366:** 1484–98.

[9] Manolio TA, Brooks LD, Collins FS. A HapMap harvest of insights into the genetics of common disease. *J Clin Invest* 2008;**118:**1590–605.

[10] Hindorff LA, Sethupathy P, Junkins HA *et al*. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci USA* 2009;**106:**9362–67.

[11] Pott P. *Chirurgical Observations Relative to the Cataract, the Polypus of the Nose, the Cancer of the Scrotum, the Different Kinds of Ruptures, and the Mortification of the Toes and Feet*. London: L. Hawes, W. Clarke, and R. Collins, 1775.

[12] Doll R, Hill AB. Smoking and carcinoma of the lung; preliminary report. *Br Med J* 1950;**2:**739–48.

[13] Via S, Lande R. Genotype-environment interaction and the evolution of phenotypic plasticity. *Evolution* 1985;**39:** 505–22.

[14] Willer CJ, Speliotes EK, Loos RJ *et al*. Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. *Nat Genet* 2009;**41:**25–34.

[15] Hunter DJ. Gene-environment interactions in human diseases. *Nat Rev* 2005;**6:**287–98.

[16] Burton PR, Hansell AL, Fortier I *et al*. Size matters: just how big is BIG? Quantifying realistic sample size requirements for human genome epidemiology. *Int J Epidemiol* 2009;**38:**263–73.

[17] Garcia-Closas M, Lubin JH. Power and sample size calculations in case-control studies of gene-environment interactions: comments on different approaches. *Am J Epidemiol* 1999;**149:**689–92.

[18] Wong MY, Day NE, Luan JA, Chan KP, Wareham NJ. The detection of gene-environment interaction for continuous traits: should we deal with measurement error by bigger studies or better measurement? *Int J Epidemiol* 2003;**32:** 51–57.

[19] Garcia-Closas M, Rothman N, Lubin J. Misclassification in case-control studies of gene-environment interactions: assessment of bias and sample size. *Cancer Epidemiol Biomarkers Prev* 1999;**8:**1043–50.

[20] Chen Z, Lee L, Chen J *et al*. Cohort profile: the Kadoorie Study of Chronic Disease in China (KSCDC). *Int J Epidemiol* 2005;**34:**1243–49.

[21] Peakman TC, Elliott P. The UK Biobank sample handling and storage validation studies. *Int J Epidemiol* 2008; **37(Suppl. 1):**i2–6.

[22] Manolio TA, Bailey-Wilson JE, Collins FS. Genes, environment and the value of prospective cohort studies. *Nat Rev* 2006;**7:**812–20.

[23] Ness AR. The Avon Longitudinal Study of Parents and Children (ALSPAC)—a resource for the study of the environmental determinants of childhood obesity. *Eur J Endocrinol* 2004;**151(Suppl. 3):**U141–49.

[24] Khoury MJ, Millikan R, Little J, Gwinn M. The emergence of epidemiology in the genomics age. *Int J Epidemiol* 2004;**33:**936–44.

[25] Repapi E, Sayers I, Wain LV *et al*. Genome-wide association study identifies five loci associated with lung function. *Nat Genet* 2009;**42:**36–44.

[26] Newton-Cheh C, Johnson T, Gateva V *et al*. Genome-wide association study identifies eight loci associated with blood pressure. *Nat Genet* 2009;**41:**666–76.

[27] Campbell PT, Jacobs ET, Ulrich CM *et al*. Case-control study of overweight, obesity, and colorectal cancer risk, overall and by tumor microsatellite instability status. *J Natl Cancer Inst* 2010;**102:**391–400.

[28] Riboli E, Kaaks R. The EPIC Project: rationale and study design. European Prospective Investigation into Cancer and Nutrition. *Int J Epidemiol* 1997;**26(Suppl 1):**S6–14.

[29] Naess O, Sogaard AJ, Arnesen E *et al*. Cohort Profile: Cohort of Norway (CONOR). *Int J Epidemiol* 2008;**37:**481–85.

[30] Litton JE, Muilu J, Bjorklund A, Leinonen A, Pedersen NL. Data modeling and data communication in GenomEUtwin. *Twin Res* 2003;**6:**383–90.

[31] Yuille M, van Ommen GJ, Brechot C *et al*. Biobanking for Europe. *Brief Bioinform* 2008;**9:**14–24.

[32] Piccinin A, Hoffer S. *Integrative Analysis of Longitudinal Studies on Aging: Collaborative Research Networks, Meta-Analysis, and Optimizing Future Studies. Handbook of Cognitive Aging: Interdisciplinary Perspectives*. London: Sage Publications, 2008.

[33] Thompson A. Thinking big: large-scale collaborative research in observational epidemiology. *Eur J Epidemiol* 2009;**24:**727–31.

[34] European Network for Genetic and Genomic Epidemiology (ENGAGE). 2010 [cited 15 June 2010]. http://www.euengage.org/ (3 August 2010, date last accessed).

[35] CESSDA. *Council of European Social Science Data Archives*. 2010 [cited 14 June 2010]. http://www.cessda.org/ (3 August 2010, date last accessed).

[36] Mailman MD, Feolo M, Jin Y *et al*. The NCBI dbGaP database of genotypes and phenotypes. *Nat Genet* 2007; **39:**1181–86.

[37] European Bioinformatics Institute. 2010 [cited 16 June 2010 16]. http://www.ebi.ac.uk/ (3 August 2010, date last accessed).

[38] Thorisson GA, Muilu J, Brookes AJ. Genotype-phenotype databases: challenges and solutions for the post-genomic era. *Nat Rev* 2009;**10:**9–18.

[39] Flicek P, Birney E. *The European Genotype Archive: Background and Implementation. Document version 2.0*. Cambridge: EBI and Sanger Institute, 2007. http://www.ebi.ac.uk/ega/bcms/ega/Documents/EGA_whitepaper.pdf (3 August 2010, date last accessed).

[40] Friedenreich CM, Slimani N, Riboli E. Measurement of past diet: review of previous and proposed methods. *Epidemiol Rev* 1992;**14:**177–96.

[41] Beer-Borst S, Morabia A, Hercberg S *et al*. Obesity and other health determinants across Europe: the EURALIM project. *J Epidemiol Community Health* 2000;**54:**424–30.

[42] Blettner M, Sauerbrei W, Schlehofer B, Scheuchenpflug T, Friedenreich C. Traditional reviews, meta-analyses and pooled analyses in epidemiology. *Int J Epidemiol* 1999; **28:**1–9.

[43] Slimani N, Deharveng G, Charrondière RU *et al*. Structure of the standardized computerized 24-h diet recall interview used as reference method in the 22 centers participating in the EPIC project. *Comput Methods Programs Biomed* 1999;**58:**251–66.

[44] Kaaks R, Slimani N, Riboli E. Pilot phase studies on the accuracy of dietary intake measurements in the EPIC project: overall evaluation of results. European Prospective Investigation into Cancer and Nutrition. *Int J Epidemiol* 1997;**26(Suppl. 1):**S26–36.

[45] Friedenreich CM. Improving long-term recall in epidemiologic studies. *Epidemiology* 1994;**5:**1–4.

[46] Pols M, Peeters P, Ocke M, Slimani N, Bueno-de-Mesquita H, Collette H. Estimation of reproducibility and relative validity of the questions included in the EPIC Physical Activity Questionnaire. *Int J Epidemiol* 1997;**26(Suppl. 1):**S181–89.

[47] Frayling TM, Timpson NJ, Weedon MN *et al*. A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. *Science* 2007;**316:**889–94.

[48] Lindgren CM, Heid IM, Randall JC *et al*. Genome-wide association scan meta-analysis identifies three loci influencing adiposity and fat distribution. *PLoS Genet* 2009;**5:** e1000508.

[49] Stover PJ, Harlan WR, Hammond JA, Hendershot T, Hamilton CM. PhenX: a toolkit for interdisciplinary genetics research. *Curr Opin Lipidol* 2010;**21:**136–40.

[50] Knoppers BM, Fortier I, Legault D, Burton P. The Public Population Project in Genomics (P3G): a proof of concept? *Eur J Hum Genet* 2008;**16:**664–65.

[51] Tolonen H, Kuulasmaa K, Laatikainen T, Wolf H. European Health Risk Monitoring Project. *Recommendation for Indicators, International Collaboration, Protocol and Manual of Operations for Chronic Disease Risk Factor Surveys*. Helsinki: Finnish National Public Health Institute, 2002.

[52] Cella D, Yount S, Rothrock N *et al*. The Patient-Reported Outcomes Measurement Information System (PROMIS): progress of an NIH Roadmap cooperative group during its first two years. *Med Care* 2007; **45(5 Suppl. 1):**S3–11.

[53] Burton P, Fortier I, Deschenes M, Hansell A, Palmer L. Biobanks and biobank harmonization. In: Davey Smith G, Burton P, Palmer L (eds). *An Introduction to Genetic Epidemiology*. Bristol: Policy Press, 2010.

[54] Friedenreich CM. Methods for pooled analyses of epidemiologic studies. *Epidemiology* 1993;**4:**295–302.

[55] North West Hub for Trial Methodology Research. *COMET Initiative*. 2010 [cited 14 June 2010]. http://www.liv.ac.uk/nwhtmr/comet/comet.htm (3 August 2010, date last accessed).

[56] HALCyon. *Healthy Aging across the Life Course*. 2010 [cited 15 June 2010]. http://www.halcyon.ac.uk/ (3 August 2010, date last accessed).

[57] Cornelis MC, Agrawal A, Cole JW *et al*. The gene, environment association studies consortium (GENEVA): maximizing the knowledge obtained from GWAS by collaboration across studies of multiple conditions. *Genet Epidemiol* 2010;**34:**364–72.

[58] Craig CL, Marshall AL, Sjostrom M *et al*. International physical activity questionnaire: 12-country reliability and validity. *Med Sci Sports Exerc* 2003;**35:**1381–95.

[59] Warren CW, Lee J, Lea V *et al*. Evolution of the Global Tobacco Surveillance System (GTSS) 1998–2008. *Glob Health Promot* 2009;**16(Suppl. 2):**4–37.

[60] Rose GA, Blackburn H. Cardiovascular survey methods. *East Afr Med J* 1969;**46:**220–27.

[61] National Centre for Social Research. *Survey Resources Network*. 2009 [cited 14 June 2010]. http://surveynet.ac.uk/ (3 August 2010, date last accessed).

[62] Public Population Project in Genomics (P3G). 2010 [cited 14 June 2010]. http://www.p3g.org/ (3 August 2010, date last accessed).

[63] Promoting Harmonisation of Epidemiological Biobanks in Europe (PHOEBE). 2010 [cited 16 June 2010]. http://www.phoebe-eu.org (3 August 2010, date last accessed).

[64] Borugian MJ, Robson P, Fortier I *et al*. The Canadian Partnership for Tomorrow Project: building a pan-Canadian research platform for disease prevention. *CMAJ* 2010;**182:**1197–201.

[65] Burton PR, Fortier I, Knoppers BM. The global emergence of epidemiological biobanks: opportunities and challenges. Building the evidence for using genetic information to improve health and prevent disease. In: Khoury MJ, Gwinn M, Bradley L, Little J, Higgins JP, Ioannidis JP (eds). *Human Genome Epidemiology (Second Edition)*. New York: Oxford University Press, 2010, pp. 77–99.

[66] OWL Working Group. *OWL Web Ontology Language Reference.* 2009 [cited 19 June 2010]. http://www.w3.org/TR/owl-ref/ (3 August 2010, date last accessed).

[67] National Center for Biomedical Ontology (NCBO). *NCBO Bioportal*. 2010 [cited 21 June 2010]. http://bioportal.bioontology.org/ (date last accessed).

[68] World Health Organisation. *International Statistical Classification of Diseases and Health Related Problems (The) ICD-10 Second Edition*. Geneva: WHO Press, 2003.

[69] World Health Organisation. *Anatomical Therapeutic Chemical Classification System with Defined Daily Doses*. 2009 [cited 3 April 2009]. http://www.whocc.no/atcddd/ (3 August 2010, date last accessed).

[70] DataSHaPER. *The DataSHaPER Project*. 2009 [cited 19 November 2009]. http://www.datashaper.org/ (3 August 2010, date last accessed).

[71] Stolk RP, Rosmalen JG, Postma DS *et al*. Universal risk factors for multifactorial diseases: LifeLines: a three-generation population-based study. *Eur J Epidemiol* 2008;**23:** 67–74.

[72] LifeGene Biobank. 2010. https://www.lifegene.se (3 August 2010, date last accessed).

[73] Raina PS, Wolfson C, Kirkland SA *et al*. The Canadian longitudinal study on aging (CLSA). *Can J Aging* 2009; **28:**221–29.

[74] GENECURE. *GENomic StratEgies for Treatment and Prevention of Cardiovascular Death in Uraemia and End-stage REnal Disease*. 2007 [cited 21 June 2010]. http://www.genecure.eu/ (3 August 2010, date last accessed).

[75] Wolfson M, Wallace SE, Masca N *et al*. DataSHIELD: resolving a conflict in contemporary bioscience—performing a pooled analysis of individual-level data without sharing the data. *Int J Epidemiol* 2010;**39:**1372–82.

[76] OBiBa. *Open Source Software for Biobanks*. 2009 [cited 19 November 2009]. http://www.obiba.org/ (2 August 2010, date last accessed).

[77] SIMBioMS. *SAIL*. 2010 [cited 24 June 2010]. http://www.ebi.ac.uk/Tools/sail/ (2 August 2010, date last accessed).

[78] Dalkey N, Helmer O. An experimental application of the delphi method to the use of experts. *Manage Sci* 1963;**9:** 458–67.

[79] Glaser EM. Using behavioral science strategies for defining the state-of-the-art. *J Appl Behav Sci* 1980;**16:**79–92.