Integrating Clinical and Epidemiological Data on Allergic Diseases Across Birth Cohorts: a MeDALL Harmonization Study

Marta Benet, Richard Albang, Mariona Pinart, Cynthia Hohmann, Christina G. Tischer, Isabella Annesi-Maesano, Nour Baïz, Carsten Bindslev-Jensen, Karin C. Lødrup Carlsen, Kai-Hakon Carlsen, Lourdes Cirugeda, Esben Eller, Maria Pia Fantini, Ulrike Gehringer, Beatrix Gerhard, Davide Gori, Eva Hallner, Inger Kull, Jacopo Lenzi, Rosemary McEachan, Eleonora Minina, Isabelle Monas, Silvia Narduzzi, Emily S. Petherick, Daniela Porta, Fanny Rancière, Marie Standl, Maties Torrent, Alet H. Wijga, John Wright, Manolis Kogevinas, Stefano Guerra, Jordi Sunyer, Thomas Keil, Jean Bousquet, Dieter Maier, Josep M. Anto, and Judith Garcia-Aymerich*

Corresponding author: Dr. Judith Garcia Aymerich. Barcelona Institute for Global Health (ISGlobal). Doctor Aiguader 88, 08003 Barcelona, Spain. Phone: +34932147380; Fax: +34932147302; (e-mail: judith.garcia@isglobal.org)

Editorial queries: Marta Benet. Doctor Aiguader 88, 08003 Barcelona, Spain Phone: +34932147307; Fax: +34932147302; (e-mail: marta.benet@isglobal.org).
Author affiliations: ISGlobal, Barcelona, Spain (Marta Benet, Mariona Pinart, Christina G. Tischer, Lourdes Cirugeda, Manolis Kogevinas, Stefano Guerra, Jordi Sunyer, Josep M. Anto, and Judith Garcia-Aymerich); Universitat Pompeu Fabra (UPF), Barcelona, Spain (Marta Benet, Mariona Pinart, Christina G. Tischer, Lourdes Cirugeda, Manolis Kogevinas, Stefano Guerra, Jordi Sunyer, Josep M. Anto, and Judith Garcia-Aymerich); CIBER Epidemiología y Salud Pública (CIBERESP), Barcelona, Spain (Marta Benet, Mariona Pinart, Christina G. Tischer, Lourdes Cirugeda, Maties Torrent, Manolis Kogevinas, Stefano Guerra, Jordi Sunyer, Josep M. Anto, and Judith Garcia-Aymerich); Biomax Informatics AG, Planegg, Germany (Richard Albang, Beatrix Gerhard, Eleonora Minina, and Dieter Maier); IMIM (Hospital del Mar Research Institute), Barcelona, Spain (Mariona Pinart, Manolis Kogevinas, Jordi Sunyer, and Josep M. Anto); Institute for Social Medicine, Epidemiology and Health Economics, Charité - Universitätsmedizin Berlin, Berlin, Germany (Cynthia Hohmann, and Thomas Keil); Epidemiology of Allergic and Respiratory Diseases Department, i-PLESP, INSERM & UPMC, Medical School Saint-Antoine, Paris, France (Isabella Annesi-Maesano, and Nour Baïz); Odense Research Center for Anaphylaxis (ORCA), Odense University Hospital, Dept. of Dermatology and Allergy Center, Odense, Denmark (Carsten Bindslev-Jensen, and Esben Eller); Department of Paediatrics, Oslo University Hospital and University of Oslo, Oslo, Norway (Karin C. Lødrup Carlsen, and Kai-Hakon Carlsen); Department of Biomedical and Neuromotor Sciences, Alma Mater Studiorum – University of Bologna, Bologna, Italy (Maria Pia Fantini, Davide Gori, and Jacopo Lenzi); Institute for Risk Assessment Sciences (IRAS), Utrecht University, Utrecht, The Netherlands (Ulrike Gehring); Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden (Eva Hallner); Centre of Occupational and Environmental
Medicine, Stockholm County Council, Stockholm, Sweden (Eva Hallner); Sachs' Children and Youth Hospital, South General Hospital Stockholm, Stockholm, Sweden (Inger Kull); Department of Clinical Science and Education, Karolinska Institutet, Stockholm, Sweden (Inger Kull); Bradford Institute for Health Research, Bradford Teaching Hospitals NHS Foundation Trust, Bradford UK (Rosemary McEachan, and John Wright); Université Paris Descartes, Sorbonne Paris Cité, EA 4064 Epidémiologie environnementale, Paris, France (Isabelle Momas, and Fanny Rancière); Mairie de Paris, Direction de l’Action Sociale de l’Enfance et de la Santé, Cellule Cohorte, Paris, France (Isabelle Momas); Department of Epidemiology, Lazio Regional Health Service, Rome, Italy (Silvia Narduzzi, and Daniela Porta); School of Sport, Exercise & Health Sciences, Loughborough University, Loughborough UK (Emily S. Petherick); Institute of Epidemiology I, Helmholtz Zentrum München – German Research Center for Environmental Health Neuherberg, Germany (Marie Standl); ib-salut, Area de Salut de Menorca, Spain (Maties Torrent); Centre for Nutrition, Prevention and Health Services, National Institute for Public Health and the Environment, Bilthoven, the Netherlands (Alet H. Wijga); National School of Public Health, Athens, Greece (Manolis Kogevinas); Asthma and Airway Disease Research Center, University of Arizona, Tucson, AZ, USA (Stefano Guerra); MACVIA-France, Contre les MAladies Chroniques pour un Vieillissement Actif en France European Innovation Partnership on Active and Healthy Ageing Reference Site, Montpellier, France (Jean Bousquet); and INSERM U 1168, VIMA: Ageing and chronic diseases Epidemiological and public health approaches, Villejuif, Université Versailles St-Quentin-en-Yvelines, UMR-S 1168, Montigny le Bretonneux, France (Jean Bousquet).
This work was supported by MeDALL (Mechanisms of the Development of ALLergy), a collaborative project conducted within the European Union under the Health Cooperation Work Programme of the 7th Framework programme (grant 261357).

The sponsor of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the manuscript. The corresponding author had full access to all study data and had final responsibility for the decision to submit for publication. ISGlobal is a member of the CERCA Programme, Generalitat de Catalunya.

Conflict of interests: RA, BG, EM, and DM are employed by Biomax Informatics AG.

Running head: Allergy Data Harmonization. The MeDALL Approach

Running head: Allergy data harmonization. The MeDALL approach

Text word count: 3317

ABSTRACT
International collaborations among birth cohorts to better understand asthma and allergies have increased in the last years. However, differences in definitions and methods preclude direct pooling of original individual participant data. We harmonized data from 14 birth cohorts, with three to 20 follow-ups, from nine European countries, as part of the Mechanisms of the Development of Asthma and Allergies (MeDALL) project. The harmonization process followed six steps: organization of the harmonization panel; identification of variables relevant to MeDALL objectives (candidate variables); proposal of a definition for each candidate variable (reference definition); assessment of the compatibility of each cohort variable to its reference definition (inferential equivalence) and classifications of this inferential equivalence as complete, partial, or impossible; workshop to agree on the reference definitions and classifications of inferential equivalence; and data preparation and delivery through a knowledge management portal. We agreed on
137 reference definitions. The inferential equivalence of 3,551 cohort variables to their corresponding reference definition was classified as complete, partial and impossible for 70%, 15% and 15% of the variables, respectively. A harmonized database was delivered. In birth cohorts of asthma and allergies, the harmonization of data for pooled analyses is feasible and may achieve high inferential comparability. The MeDALL harmonization approach can be used in other collaborative projects.

Keywords [MeSH terms]: allergy, asthma, data accuracy, birth cohorts, data pooling, data sharing, harmonization.

Abbreviations: DataSHaPER, Data Schema and Harmonization Platform for Epidemiological Research; FP, Framework Program; and MeDALL, Mechanisms of the Development of Asthma and Allergies.

Over 130 birth cohorts with data on asthma and allergy have been initiated in the world over the past 30 years (1). The information gathered by these birth cohorts has already significantly advanced in our understanding of allergy and asthma, particularly during the first years of life (2). However, this data is usually in an isolated, independent database. Although the assessment methods of the data vary, the majority of the birth cohorts followed rigorous methodology, and the resultant data is relatively readily available in electronic format.

Since 2004, the EU Framework Program (FP) for Research and Technological Development FP6-FP7 have funded projects to identify, compare, and evaluate pooling data from existing European birth cohorts (Global Allergy and European Network (GA²LEN), FP6 (3-7), Environmental Health Risks in European Birth Cohorts (ENRIECO), FP7 (8, 9), Developing a Child Cohort Research Strategy for Europe (CHICOS), FP7 (10), and Mechanisms of the Development of ALLergy (MeDALL), FP7 (2, 10-12)). These projects have strengthened the networking capacity of birth
cohorts and produced a large number of joint studies that have frequently used meta-analysis based on cohort original data (1, 12, 13). Though few studies have integrated data from different birth cohorts in single pooled analysis (7, 14), a formal reproducible approach for data harmonization has not been reported.

Several approaches have been proposed to harmonize data from different cohorts (15-21). Among them, the Data Schema and Harmonization Platform for Epidemiological Research (DataSHaPER) project (15) and the Maelstrom Research guidelines (16) have provided guidelines aiming to facilitate rigorous, transparent, and effective harmonization. Other initiatives have proposed methods for collaborative study designs (22) or harmonization of data collection (23). However, no studies have adopted a formal harmonization approach to asthma and allergic diseases despite the well-known complexity in defining and assessing these conditions (1).

Therefore, we report the strategy, process, and results of the harmonization developed during the MeDALL FP7 project (2, 11, 12). We adapted the DataSHaPER approach and capitalized on the experience in previous harmonization efforts by the partners mentioned above (4, 8, 10) and the technological support provided by a knowledge management portal for systems medicine (24).

METHODS

Birth cohorts

The harmonization included questionnaire information from children participating in one of the 14 longitudinal population-based birth cohorts. A total of 47,998 children were recruited from pregnant women or mothers with new born babies in nine
European countries (25). Seven of them recruited children between 1990 and 1998: Asthma Multicenter Infants Cohort Study (AMICS-Menorca), Spain (26), Children Allergy, Milieu, Stockholm, Epidemiology (BAMSE study), Sweden (27,28), Environment and Childhood Asthma study in Oslo (ECA), Norway (29), German Infant Study on the influence of Nutrition Intervention PLUS environmental and genetic influences on allergy development (GINIplus), Germany (30), Influence of life-style factors on the development of the immune system and allergies in East and West Germany Plus the influence of traffic emissions and genetics (LISApplus), Germany (31), Multicenter Allergy Study (MAS), Germany (32), and Prevention and Incidence of Asthma and Mite Allergy (PIAMA), Netherlands (33). Remaining seven cohorts included children recruited between 2003 and 2009: Born in Bradford (BIB), United Kingdom (34), Study of the determinants pre and postnatal of child health and development (EDEN), France (35), Environmental and childhood Project – Sabadell (INMA Sabadell), Spain (26), Pollution and Asthma Risk: An Infant Study (PARIS), France (36), Mother-child cohort in Crete (RHEA), Greece (37), Roma and Bologna Birth Italian Cohorts – Roma (ROBBIC–Roma), Italy (38), and Roma and Bologna Birth Italian Cohorts – Bologna (ROBBIC–Bologna), Italy (38). In all cohorts, parents gave written informed consent and the studies were approved by local ethics review boards.

Variables

All birth cohorts collected information on participants for a minimum of three and a maximum of 20 follow-up periods (from pregnancy to 20 years of age), see Web Table 1. All birth cohorts followed standardized protocols and included several validated questions regarding the outcome variables such as the International Study
of Asthma and Allergies in Childhood (ISAAC) (39). Birth cohorts followed strict quality control measures before, during, and after data collection to ensure validity of the data collected.

Harmonization process

The harmonization process was adapted from the DataSHaPER project (15) and followed six steps (see Figure 1).

*Step 1: organization of the harmonization panel*, formed by the harmonization coordinators and cohort experts. The harmonization coordinators were in charge of organizing all the process, contacting each cohort, and ensuring active participation of the cohort experts. These included, for each birth cohort, a principal investigator and a statistician or data manager very familiar with the cohort database.

*Step 2: Identification of candidate variables*. The cohort experts identified relevant variables for ongoing and future research objectives within MeDALL. From the identified variables, the harmonization coordinators pre-selected those for which (i) an agreed reference definition was likely to be found or produced by expert consensus, and (ii) enough data was available to provide sufficient power for the envisioned analyses (i.e. at least three cohorts had data available for the variable). The candidate variables were then classified into (i) complex harmonization needed, and (ii) basic harmonization needed (e.g. age, gender, height). A total of 122 variables were preliminary classified as “complex harmonization needed” and were allocated to one of five dimensions: (i) symptoms, (ii) treatment, (iii) environmental exposures, (iv) sociodemographic, and (v) physical activity. (See complete list of variables per dimension in Web Table 2). A total of 28 variables were classified as “basic harmonization needed”. They covered basic demographic variables, early life
risk factors and clinical information, and mostly required adaptation of the units of measurement (see details in Web Table 3).

Step 3: Proposal of a reference definition. The harmonization coordinators proposed a reference definition for each variable based on the validated ISAAC questionnaire (39) and the MeDALL core questionnaires (40). When a reference definition was not available in these sources, the cohort experts were asked to propose one. All proposed reference definitions can be found in the Web Table 2.

Step 4: Inferential equivalence classification of cohort variables to reference definitions. The principal investigator of each cohort identified which question(s) matched each candidate variable in their cohort at the different follow-ups. Most questions (candidate variables) were collected at several periods (e.g., wheezing in last 12 months) and they were considered as many times as they appeared. Then each principal investigator assessed the compatibility (inferential equivalence) of their own variables to the corresponding reference definitions by means of assessing the meaning, format, and data collection procedure of each variable (general pairing rules). Three qualification categories (complete, partial, and impossible) were used, adapted from the ones proposed in the DataSHaPER project (15, 21). A variable was classified as complete if the meaning, format, and standard operating procedures used for data collection allowed the complete construction of the reference definition. A partial qualification was given if the meaning, format, and standard operating procedures used for the data collection allowed the construction of the reference definition, but with an unavoidable loss of information. The inferential equivalence of a variable was classified as impossible when insufficient information existed to construct the reference definition. Further, when a given variable was not included in
a specific cohort the inferential equivalence classification of that variable was considered as missing. To facilitate this task, the harmonization coordinators distributed some examples of candidate variables, reference definitions and classification of inferential equivalence along with the specific pairing rules that could be applied to these example variables. Harmonization coordinators compiled all cohort qualifications prior to a workshop (see next step) for final consensus building.

**Step 5: Consensus agreement workshop.** Harmonization coordinators organized a four-day consensus agreement workshop with the cohort experts to agree on final reference definitions, inferential equivalence classification, and pairing rules for all variables. The rules for discussion were made explicit and agreed by the harmonization panel at the beginning of the workshop e.g. a maximum of ten minutes was assigned for the discussion of a reference definition; if no consensus was reached during that time the proposed reference definition was excluded from the harmonization process and its variable(s) from the final database. Notes were taken during the workshop by different participants and checked by the harmonization coordinators for a post workshop quality control. The final agreed reference definitions can be found in the Web Table 2.

**Step 6: Data preparation and delivery.** Each cohort provided the harmonized variables following the decisions agreed on during the workshop to the knowledge management portal.

The MeDALL partner Biomax, a bioinformatics company with experience in systems medicine (24, 41, 42), provided dedicated technological support during all the steps. Biomax developed a knowledge management portal for the project ([https://ssl.biomax.de/medall](https://ssl.biomax.de/medall)) that stores, manages, structures, and provides project-
specific knowledge, allowing flexible data harmonization and integration. After the harmonization process, all the data was integrated in the portal where different algorithmic checks were performed to ensure data quality. Cleaning rules included checks for completeness (e.g., were data available for all participants?), availability of required data (e.g., were all variables provided as defined in the metadata?), data type check (e.g., float, thesaurus), consistency checks of the collected data across follow-up periods (e.g., if a “yes” answer occurred to “doctor diagnosis of asthma ever” then the same question in the following examinations were set to “yes”), and outlier detection. Units and codes were automatically converted and unified based on a unit ontology and code mapping. We also included logical checks where possible, e.g. stated gender was confirmed with chromosomal information data. Finally, data quality descriptors were provided as summary statistics including completeness, coverage (availability at different ages), and standard statistics of data distribution.

Statistical analysis

We calculated the proportion of variables classified as complete, partial, and impossible among the total that required complex harmonization, and stratified these results by cohort and harmonization step (before or after the workshop).

The Cohen’s kappa coefficient was calculated to evaluate the agreement between the qualifications done by each cohort before the workshop and the qualifications resulting from it. This coefficient was calculated overall, by cohort, by domains, and by variables.

RESULTS

Reference definitions
A total of 122 reference definitions were proposed for discussion in the consensus agreement workshop, during which some reference definitions were changed for clarification, variable merging (i.e. combining two or more definitions in one), or creation of new reference definitions. We finally harmonized 137 reference definitions (see Web Table 2 for all proposed reference definitions together with modifications) and classified the inferential equivalence to the reference definition of 3,551 variables collected across the multiple follow-ups of the 17 cohorts.

Pairing rules

During the harmonization workshop, we agreed on the pairing rules to classify the inferential equivalence of each variable to its reference definition. For example, a variable would result in a complete qualification if differences to the reference definition consisted of: (i) minor additional answer categories e.g. having the explicit missing option don’t know or don’t answer; or (ii) equivalent methods of data generation e.g. telephone interview vs paper questionnaire. A partial qualification would result if: (i) minor language differences were found e.g. single synonym not covered; or (ii) minor part of the definition was not asked e.g. “had an asthma attack” instead of “ever had an asthma attack”. Finally, an impossible qualification would result if: (i) questions asked about different time frames e.g. “at least two weeks” instead of “at least six months”; (ii) variables had strongly more restrictive definitions e.g. asking for a specific allergic reaction instead of asking for an allergic reaction in general; or (iii) different methods of data generation had been used e.g. physical activity from an accelerometer vs questionnaire data. Pairing rules did not include any consideration with regard to data distribution for each variable (e.g., mean values, missing values, or outliers). Table 1 shows an example of how a variable was
harmonized including the reference definition agreed during the workshop, the definitions available in different cohorts or periods, and a set of pairing rules. All harmonization results are stored in the knowledge management portal and can be provided upon request.

(Table 1 here)

Inferential equivalence classification of variables

Before the workshop, cohort experts classified their variables according to their inferential equivalence (Table 2). Among 3,551 variables included, 2,206 variables (62%) were qualified as complete, 1,243 (35%) as partial, and 102 (3%) as impossible. After the workshop, 2,481 (70%) of the 3,551 variables were qualified as complete, 550 (15%) as partial, and 520 (15%) as impossible (Table 2). Variables that were not available (missing) in a given cohort or period were not included in the denominator, as their inferential equivalence (complete, partial, or impossible) remains unknown. Web Table 4 shows the inferential equivalence for all variables according to cohort and period. Figure 2 shows the distribution of final inferential equivalence classification according to the five variable dimensions mentioned above. The symptoms dimension was the closest to the overall classification with 73% of the variables in this dimension classified as complete, 13% as partial, and 14% as impossible. The proportion of variables classified as complete was high (79%) in the environmental exposures dimension and low in the treatment (57%) and physical activity (29%) dimensions. Almost 60% of variables in the physical activity dimension were classified as impossible. Final classifications for all included variables are available in Web Figures 1 to 14. All variables, and their inferential equivalence classifications, have been integrated in the final MeDALL database in
order to provide researchers with additional information to conduct sensitivity analyses/test miss-classification.

Agreement between inferential equivalence classification before and during the workshop

The overall agreement between the inferential equivalence classification assigned to all variables before the workshop by the cohort principal investigator and the final qualifications agreed during the workshop was 0.49, ranging from 0.32 in PIAMA to 0.76 in PARIS birth cohorts (Table 2). In general, agreement was higher for variables from cohorts recruiting children between 2003 and 2009 than for those recruiting children between 1990 and 1998. A fair to moderate agreement was obtained for all five dimensions (0.40 to 0.50) (data on agreement by dimension and for each individual variable is available from the authors upon request).

(Table 2 here)

DISCUSSION

Main findings

The present MeDALL harmonization study shows that harmonization of databases from different European asthma and allergy birth cohorts is feasible and successful following and adapting the steps reported by the DataSHaPER (15, 21) group. After six months of preparation and a four-day workshop we have agreed on 137 reference definitions and classified their inferential equivalence to 3,551 cohort variables. More than two thirds of the harmonized variables were classified as complete and the remaining 30 per cent were either partial, or impossible.

Comparison with similar initiatives
This manuscript supports and extends previous and ongoing initiatives on data
harmonization (15-21, 43-46). Two novel features of our harmonization process are:
(i) the consensus workshop as a key step that allows discussion and agreement on
all reference definitions and inferential equivalences, and (ii) the broad spectrum of
harmonized exposures and outcomes, not driven by a single specific research
questions, but integrated to eventually answer multiple research questions (47, 48)
including omics (49).

Our findings support the importance of undertaking the harmonization exercise at the
beginning of a large collaborative project. Actually, it is common to undertake several
harmonization efforts of the same variables at multiple occasions for different
analysis involving different actors and implying a substantial waste of time and lack of
reliability. The moderate agreement in variable qualification before and after the
workshop (overall kappa coefficient of 0.49) may result from numerous aspects that
an individual expert would consider differently when thinking alone than in group
discussions. These aspects may include conceptual complexity of the involved
variables, differences in the wording and formatting of the questions and even the
iterative nature of the harmonization process itself. In this sense, the kappa values
should not be interpreted as a measure of the quality of the first inferential
equivalence classification but as a marker of the complexity and the necessity of the
harmonization process. Our approach ensures a more efficient use of time and
financial resources, improves reliability of results of pooled analysis within the
MeDALL project, and allows performing meta-analyses with other project’s data with
a clear frame on how variables have been defined (50). In general, no significant
differences in results have been found between meta- and pooled analyses although
pooled analysis exhibits higher precision of estimates (51-53). Since a big limitation
to pooling data is heterogeneity, a harmonization process, as the one reported here, will facilitate also pooled strategies in the future.

Strengths and limitations

Strengths of the present work are the very rigorous process applied, which allows others to reproduce the approach, the use of a technological support (the MeDALL knowledge management portal) that includes all reference definitions, variables, and codification, and all the expert knowledge used in order to take decisions. Existing long term collaboration of most birth cohorts starting with the GA\textsuperscript{2}LEN initiative (3, 4), and continued through the ENRIECO (8) and CHICOS (10) projects were fundamental to this commitment and to establish a birth-cohort alliance in the Human Early-Life Exposome project (HELIX) (54) which links all environmental hazards that mothers and children are exposed to, to the health, growth, and development of children. Harmonized data based from these cohorts increase the range of exposures, increases the sample size, and thus the statistical power of the study and allows for a more detailed stratification. Therefore, a collaborative project with harmonized data (either performing pooled or meta-analyses) will increase the reproducibility, reliability, and validity of its results (49). The harmonization process involved a panel of multidisciplinary experts including medical, epidemiological, psychological, biostatistical, data management, and IT experts. Following harmonization, all MeDALL partners agreed to keep up-to-date and active the data management portal so that several research studies could be conducted and their results published (50, 55, 56). Finally, three additional MeDALL birth cohorts that did not participate in the first harmonization process could be harmonized thanks to the detailed harmonization reports available in the knowledge management portal.
(details available from Biomax upon request), thus supporting the reproducibility of our approach.

We encountered several limitations while harmonizing the MeDALL data. First, the cross-cultural differences have been challenging occasionally, with some of the symptom definitions reflecting the subtle differences between the languages involved in this large European collaboration (e.g. wheezing in German cannot be translated directly but is translated in three words: Giemen, Pfeifen, Brummen). Second, the cohorts were heterogeneous regarding the spectrum and assessment methods of environmental and psychosocial exposures. For instance, some of the cohorts had more detailed questions on indoor environment than others (26-31, 33, 36, 38) while others focused on psychological factors (26-28, 30-32). Of note, some exposures and diseases could not be harmonized due to the large heterogeneity or lack of data. Thus the new common database after the MeDALL harmonization work does not yet include all but a large set of all core variables on asthma and allergy and on the most prevalent exposures and risk factors.

Third, we did not assess the influence of using harmonized variables on the validity of previous studies using the same variables, which is an area deserving attention in future research. Finally, our study did not consider country differences in intellectual property rights or ethical rules and regulations, which fall beyond the scope of a data harmonization exercise.

Conclusions

We have shown that data harmonization from different birth cohort and periods with cross-cultural differences is feasible and may achieve high comparability by using a predefined strategy, a technological support, and commitments from all involved
members. We encourage other collaborative projects to adopt and execute similar harmonization strategies either by accessing our reference definitions, detailed pairing rules, and examples for variables on allergic symptoms, diseases, and risk factors in children, or by taking advantage of the lessons learned and detailed stepwise description of the defined procedures. Further evidence is needed on the effects of the data harmonization process in the validity of study results.

REFERENCES


5 Bousquet J, Burney PG, Zuberbier T, et al. GA2LEN (Global Allergy and Asthma European Network) addresses the allergy and asthma 'epidemic'. *Allergy*. 2009;64(7):969-977.


Porta D, Fantini MP, on behalf of the GASPII and Co.N.ER Study Groups. Prospective cohort studies of newborns in Italy to evaluate the role of environmental and genetic characteristics on common childhood disorders. *Ital J Pediatr.* 2006;32:350–357.


Table 1: Example of a Reference Definition $^a$ and Pairing Rules to Classify the Inferential Equivalence of Each Original Cohort Variable $^b$ to This Reference Definition, as Part of the Harmonization Process of Asthma and Allergy Data in 14 Birth Cohorts from Nine European Countries

<table>
<thead>
<tr>
<th>Inferential equivalence classification (qualification)</th>
<th>Definition provided by birth cohorts</th>
<th>Pairing rules</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complete</td>
<td>Has your child had wheezing or <strong>whistling</strong> in the chest during or after exercise in the last 12 months</td>
<td>Synonyms for “wheezing” are accepted as they are language and cultural specific</td>
</tr>
<tr>
<td></td>
<td>Has your child ever had wheeziness when playing or when outdoors with/without having a cold?</td>
<td>The timing of wheezing relative to exercise can be either during or after it.</td>
</tr>
<tr>
<td></td>
<td>Has your child had wheeziness when playing or when outdoors with/without having a cold after the age of one year?</td>
<td>All questions not specifying “in the last 12 months” but where the “12 months” are respected due to the follow-up time frames, have been considered as “complete”.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Before the age of 2 years “playing or when outdoors” are considered as “exercise” (question asked at follow-up age two years or earlier).</td>
</tr>
<tr>
<td>Partial</td>
<td>Has your child had trouble breathing in connection with exertion in the past 12 months?</td>
<td>The symptoms regarding breathing difficulties asked in this question were considered to be broader than the ones asked in the reference definition, which focused on wheezing.</td>
</tr>
<tr>
<td>---------</td>
<td>--------------------------------------------------------------------------------------</td>
<td>----------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Impossible</td>
<td>In the past 24 months, has your child's chest sounded wheezy during or after exercise?</td>
<td>The timeframe from this definition is broader than the one asked in the reference definition, 24 months vs...</td>
</tr>
</tbody>
</table>
Has your child ever sounded like that (wheezing and whistling) after exercise?  


Has your child ever sounded like that after exercise?

---

**a**Reference definition. In the past 12 months, has your child's chest sounded wheezy during or after exercise? (Yes/No)

**b**Variable name, Wheezing after exercise last 12 months;
Table 2: Distribution by Cohort of Variables Inferential Equivalence Classification Before and After the Consensus Workshop to Harmonize Asthma and Allergy Data in 14 Birth Cohorts from Nine European Countries

<table>
<thead>
<tr>
<th>Cohort (ordered by recruitment year)</th>
<th>Recruitment</th>
<th>n° definitions</th>
<th>Before Workshop</th>
<th>After the Workshop</th>
<th>Kappa</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>year</td>
<td></td>
<td>Complete</td>
<td>Partial</td>
<td>Impossible</td>
</tr>
<tr>
<td>MAS</td>
<td>1990</td>
<td>393</td>
<td>205 52</td>
<td>185 47</td>
<td>3 1</td>
</tr>
<tr>
<td>ECA</td>
<td>1992/1993</td>
<td>304</td>
<td>232 76</td>
<td>60 20</td>
<td>12 4</td>
</tr>
<tr>
<td>BAMSE</td>
<td>1994/1996</td>
<td>219</td>
<td>119 54</td>
<td>100 46</td>
<td>0 0</td>
</tr>
<tr>
<td>PIAMA</td>
<td>1996/1997</td>
<td>420</td>
<td>290 69</td>
<td>128 30</td>
<td>2 1</td>
</tr>
<tr>
<td>GINIplus</td>
<td>1996/1998</td>
<td>338</td>
<td>108 32</td>
<td>210 62</td>
<td>20 6</td>
</tr>
<tr>
<td>AMICS-Menorca</td>
<td>1997/1998</td>
<td>422</td>
<td>344 82</td>
<td>78 18</td>
<td>0 0</td>
</tr>
<tr>
<td>LISAplus</td>
<td>1997/1998</td>
<td>335</td>
<td>100 30</td>
<td>230 69</td>
<td>5 1</td>
</tr>
<tr>
<td>ROBBIC-Roma</td>
<td>2003/2004</td>
<td>114</td>
<td>65 57</td>
<td>21 18</td>
<td>28 25</td>
</tr>
<tr>
<td>EDEN</td>
<td>2003/2005</td>
<td>150</td>
<td>94 63</td>
<td>48 32</td>
<td>8 5</td>
</tr>
<tr>
<td>PARIS</td>
<td>2003/2006</td>
<td>401</td>
<td>349 87</td>
<td>38 9</td>
<td>14 4</td>
</tr>
<tr>
<td>ROBBIC-Bologna</td>
<td>2004/2005</td>
<td>72</td>
<td>61 85</td>
<td>11 15</td>
<td>0 0</td>
</tr>
<tr>
<td>INMA-Sabadell</td>
<td>2004/2007</td>
<td>114</td>
<td>60 53</td>
<td>53 46</td>
<td>1 1</td>
</tr>
<tr>
<td>RHEA</td>
<td>2007/2008</td>
<td>119</td>
<td>84 71</td>
<td>35 29</td>
<td>0 0</td>
</tr>
<tr>
<td>BIB</td>
<td>2007/2009</td>
<td>150</td>
<td>95 63</td>
<td>46 31</td>
<td>9 6</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>3,551</td>
<td>2,206 62</td>
<td>1,243 35</td>
<td>102 3</td>
</tr>
</tbody>
</table>

AMICS-Menorca, Asthma Multicenter Infants Cohort Study; BAMSE study, Children Allergy, Milieu, Stockholm, Epidemiology; BIB, Born in Bradford; ECA, Environment and Childhood Asthma study in Oslo; EDEN, Study of the determinants pre and postnatal of child health and development; GINIplus, German Infant Study on the influence of Nutrition Intervention PLUS environmental and genetic influences on allergy development; INMA Sabadell, Environmental and childhood Project – Sabadell; LISAplus, Influence of life-style factors on the development of the immune system and allergies in East and West Germany Plus the influence of traffic emissions and genetics; MAS, Multicenter Allergy Study; PARIS, Pollution and Asthma Risk: An Infant Study; PIAMA, Prevention and Incidence of Asthma and Mite Allergy; RHEA, Mother-child cohort in Crete; ROBBIC–Roma, Roma and Bologna Birth Italian Cohorts – Roma; ROBBIC–Bologna, Roma and Bologna Birth Italian Cohorts – Bologna. *From a total of 122 requested variable definitions, the number of definitions per cohort depends on the number of follow-up periods where each variable was available. 

Kappa values indicate the level of agreement between the classifications before and after the workshop.
Figure 1: Flow Chart of the Harmonization Process of Asthma and Allergy Variables in 14 Birth Cohorts from Nine European Countries. Abbreviations: HC, harmonization coordinators; CE, cohort experts; PI, principal investigator; RD, reference definition; SOP, standard operating procedures.

Figure 2: Distribution of Inferential Equivalence Classification of Cohort Variables to Reference Definitions, Overall and by Variables Dimensions. The Figure presents the percentage of variables classified as complete (black), partial (lined), and impossible (white) among the total that required complex harmonization. Web Figures 1 to 14 include the distribution of inferential equivalence classification for each variable, as follows: Symptoms: asthma and wheezing (Web Figure 1), rhinitis (Web Figure 2), eczema (Web Figure 3), other allergic related variables (Web Figure 4), family history of allergic diseases (Web Figure 5), and puberty (Web Figure 6); Treatment: treatments for allergic diseases in the last 12 months (Web Figure 7), doctor consultations for allergic diseases in the last 12 months (Web Figure 8), triggers of allergic diseases in the last 12 months (Web Figure 9), school or outdoor activities absenteeism due to allergic diseases in the last 12 months (Web Figure 10); Environmental exposures: indoor (Web Figure 11), and smoking (Web Figure 12) exposures; Sociodemographic: siblings and other children at home (Web Figure 13); and Physical Activity: type, intensity, and period of physical activity (Web Figure 14).
From the identified relevant variables the HC pre-selects those for which:

1. An agreed reference definition (RD) is likely to be found or produced by expert consensus
2. Enough data is available to provide sufficient power for the analyses

From the identified relevant variables the HC pre-selects those for which:

- Complex harmonization needed
- Basic harmonization needed

HC: proposes a RD for each variable to be harmonized (based on validated questionnaires).
If HC cannot find a proposal for a RD → CE is asked to propose one

PIs: assesses compatibility of their own cohort variables to the corresponding RDs, classifying them as:
- Complete: if meaning, format, and standard operating procedures (SOPs) allow complete construction of the RD
- Partial: if meaning, format, and SOPs allow construction of RD but with a loss of information
- Impossible: insufficient information existed to construct RD

HC: compiles all cohort qualifications prior to a workshop for final consensus building

HC organizes a four-day workshop to agree on:
1. RDs when needed
2. Variables inferential equivalence classification
3. Pairing rules for variables with partial qualifications

Workshop rules for RDs agreement:
Maximum of 10 minutes discussion for each RD

Consensus

No Consensus on RD
RD Excluded From Harmonization

Each cohort provided harmonized variables following the decisions agreed during the consensus agreement workshop to the knowledge management portal