Integrating Clinical and Epidemiologic Data on Allergic Diseases Across Birth Cohorts: A Harmonization Study in the Mechanisms of the Development of Allergy Project

Integrating clinical and epidemiological data on allergic diseases across birth cohorts: a MeDALL harmonization study

Marta Benet,[1,2,3] Richard Albang,[4] Mariona Pinart,[1,2,3,5] Cynthia Hohmann,[6] Christina G Tischer,[1,2,3] Isabella Annesi-Maesano,[7] Nour Baïz,[7] Carsten Bindslev-Jensen,[8] Karin C Lødrup Carlsen,[9] Kai-Hakon Carlsen,[9] Lourdes Cirugeda,[1,2,3] Esben Eller,[8] Maria Pia Fantini,[10] Ulrike Gehring,[11] Beatrix Gerhard,[4] Davide Gori,[10] Eva Hallner,[12,13] Inger Kull,[14,15] Jacopo Lenzi,[10] Rosemary McEachan,[16] Eleonora Minina,[4] Isabelle Momas,[17,18] Petter Mowinkel,[9] Silvia Narduzzi,[19] Emily S Petherick,[20] Daniela Porta,[19] Fanny Rancière,[17] Marie Standl,[21] Maties Torrent,[3,22] Alet H Wijga,[23] John Wright,[16] Manolis Kogevinas,[1,2,3,5,24] Stefano Guerra,[1,2,3,25] Jordi Sunyer,[1,2,3,5] Thomas Keil,[6] Jean Bousquet,[26,27] Dieter Maier,[4] Josep M Anto,[1,2,3,5] Judith Garcia-Aymerich,[1,2,3]*

Affiliations:

1- ISGlobal, Barcelona, Spain

2- Universitat Pompeu Fabra (UPF), Barcelona, Spain

3- CIBER Epidemiología y Salud Pública (CIBERESP), Barcelona, Spain

4- Biomax Informatics AG, Planegg, Germany

5- IMIM (Hospital del Mar Research Institute), Barcelona, Spain

6- Institute for Social Medicine, Epidemiology and Health Economics, Charité - Universitätsmedizin Berlin, Berlin, Germany

7- Epidemiology of Allergic and Respiratory Diseases Department, i-PLESP, INSERM & UPMC, Medical School Saint-Antoine, Paris, France

8- Odense Research Center for Anaphylaxis (ORCA), Odense University Hospital, Dept. of Dermatology and Allergy Center, Odense, Denmark

9- Department of Paediatrics, Oslo University Hospital and University of Oslo, Oslo, Norway

10- Department of Biomedical and Neuromotor Sciences, Alma Mater Studiorum - University of Bologna, Bologna, Italy

11- Institute for Risk Assessment Sciences (IRAS), Utrecht University, Utrecht, The Netherlands

12- Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden

13- Centre of Occupational and Environmental Medicine, Stockholm County Council, Stockholm, Sweden

14- Sachs' Children and Youth Hospital, South General Hospital Stockholm, Stockholm, Sweden

15- Department of Clinical Science and Education, Karolinska Institutet, Stockholm, Sweden

16- Bradford Institute for Health Research, Bradford Teaching Hospitals NHS Foundation Trust, Bradford UK

17- Université Paris Descartes, Sorbonne Paris Cité, EA 4064 Epidémiologie environnementale, Paris, France

18- Mairie de Paris, Direction de l'Action Sociale de l'Enfance et de la Santé, Cellule Cohorte, Paris, France

19- Department of Epidemiology, Lazio Regional Health Service, Rome, Italy

20- School of Sport, Exercise & Health Sciences, Loughborough University, Loughborough UK

21- Institute of Epidemiology I, Helmholtz Zentrum München – German Research Center for Environmental Health Neuherberg, Germany

22- ib-salut, Area de Salut de Menorca, Spain

23- Centre for Nutrition, Prevention and Health Services, National Institute for Public Health and the Environment, Bilthoven, the Netherlands

24- National School of Public Health, Athens, Greece

25- Asthma and Airway Disease Research Center, University of Arizona, Tucson, AZ, USA

26- MACVIA-France, Contre les MAladies Chroniques pour un VIeillissement Actif en France European Innovation Partnership on Active and Healthy Ageing Reference Site, Montpellier, France

27- 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

Corresponding author: Judith Garcia Aymerich. Barcelona Institute for Global Health. Doctor Aiguader 88, 08003 Barcelona, Spain. Phone: +34932147307; Fax: +34932147302; e-mail: judith.garcia@isglobal.org

Text word count: 2694

ABSTRACT

BACKGROUND: International collaborations among birth cohorts to better understand asthma and allergies have increased in the last years. However, differences in definitions and assessment of relevant variables preclude direct pooling of original individual participant data. As part of the Mechanisms of the Development of Asthma and Allergies (MeDALL) project, we aimed to harmonize multiple birth cohort data allowing for pooled analysis of asthma, rhinitis, and eczema.

METHODS: We included 17 birth cohorts from ten European countries. The harmonization process consisted in: (1) organization of the harmonization panel, (2) identification of candidate variables relevant to MeDALL research objectives, (3) proposal of reference definitions for each candidate variable, (4) classification of the inferential equivalence of each cohort variable to its reference definition as *complete*, *partial,* or *impossible*, (5) consensus agreement workshop to agree on the reference definitions and classifications of the inferential equivalence, and (6) data preparation and delivery for analyses through a knowledge management portal.

RESULTS: We agreed on 137 reference definitions and classified the inferential equivalence of 3551 cohort variables (17 cohorts with three to 20 follow-ups) to their corresponding reference definition as *complete* (70% of the variables), *partial* (15%). or *impossible* (15%). The agreement (Cohen's kappa) between classifications before and during the workshop ranged between 0.32 and 0.76. A harmonized database was delivered.

CONCLUSION: In birth cohorts of asthma and allergies, the harmonization of data for pooled analyses is complex but feasible and may achieve high inferential comparability. The MeDALL harmonization approach can be used in other collaborative projects.

Abstract word count: 250

Keywords: harmonization, data pooling, data quality, birth cohorts, asthma, allergy, data sharing, epidemiology

KEY MESSAGES

- The harmonization of individual participant data 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 expert representatives of all cohorts.

- We provide reference definitions and detailed pairing rules for the harmonization of variables about asthma and allergic symptoms, diseases, and risk factors in children.

- The MeDALL approach and reference definitions can be used in future collaborative studies of asthma and allergy.

INTRODUCTION

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. 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 for Research and Technological Development FP6-FP7 have funded projects to identify, compare, and evaluate pooling data from existing European birth cohorts (GA²LEN: Global Allergy and European Network, FP6 [2-6], ENRIECO: Environmental Health Risks in European Birth Cohorts, FP7 [7, 8], CHICOS: Developing a Child Cohort Research Strategy for Europe, FP7 [9], and MeDALL: Mechanisms of the Development of ALLergy, FP7 [9-12]). 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. Though few studies have integrated data from different birth cohorts in single pooled analysis, a formal reproducible approach for data harmonization has not been reported.

An approach to harmonize data from different cohorts has been recently proposed by the DataSHaPER project [13] and the Maelstrom Research guidelines [14]. These studies have provided guidelines aiming to facilitate rigorous, transparent, and effective harmonization. However, only few studies have adopted a formal harmonization approach [15-17].

Therefore, we report the strategy, process, and results of the harmonization developed during the MeDALL (Mechanisms of the Development of ALLergy) FP7 project [10-12]. We adapted the DataSHaPER approach and capitalized on the experience in previous

harmonization efforts by the partners mentioned above [3, 7, 9] and the technological support provided by a knowledge management portal for systems medicine [18].

METHODS

Birth Cohorts

The harmonization included questionnaire information from 17 population-based birth cohorts that recruited pregnant women and mothers with new born babies in ten European countries [19] (details on cohorts are provided in the supplementary material). Eight of them (from now on referred to as older cohorts) recruited children between 1990 and 1998: AMICS-Menorca, Spain [20], BAMSE, Sweden [21,22], DARC, Denmark [23], ECA, Norway [24], GINIplus, Germany [25], LISAplus, Germany [26], MAS, Germany [27], and PIAMA, Netherlands [28]. Remaining nine cohorts (younger cohorts) included children recruited between 2003 and 2009: BIB, United Kingdom [29], EDEN, France [30], INMA Guipuzkoa, Spain [18], INMA Sabadell, Spain [18], INMA Valencia, Spain [18], PARIS, France [31], RHEA, Greece [32], ROBBIC–Rome, Italy [33], and ROBBIC–Bologna, Italy [33]. 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 supplementary table S1. 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) [34].

Harmonization process

The harmonization process was adapted from the DataSHaPER project [13] and followed six steps (see figure 1).

Step 1**: Organization of the harmonization panel**, formed by the harmonization coordinators and harmonization experts. The harmonization coordinators were in charge of organizing all the process, contacting each cohort, and ensuring active participation of the harmonization 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 harmonization experts identified relevant variables for ongoing and future research objectives within MeDALL. From the identified variables, the harmonization coordinators pre-selected those for which (1) an agreed reference definition was likely to be found or produced by expert consensus, and (2) 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) harmonization needed, and (ii) harmonization not needed (e.g. age, gender, height). A total of 122 variables were classified as "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 supplementary table S2).

Step 3: **Proposal of a reference definition**. The harmonization coordinators proposed a reference definition for each variable based on the validated ISAAC questionnaire [34] and the MeDALL core questionnaires [35]. When a reference definition was not available in these sources, the harmonization experts were asked to propose one. All proposed reference definitions can be found in the supplementary table S2.

Step 4: **Inferential equivalence classification of cohort variables to reference definitions**. Each principal investigator assessed the compatibility (inferential equivalence) of their own variables to the corresponding reference definitions using three qualification categories (*complete*, *partial, and impossible)* adapted from the ones proposed in the DataSHaPER project [13, 17]. 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 no information was collected on a specific variable in a given cohort inferential equivalence classification was not possible. 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 harmonization experts to agree on: (1) reference definitions, (2) variables inferential equivalence classification, and (3) pairing rules for variables with a *partial* qualification. 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 supplementary table S2.

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 [18, 36, 37], provided dedicated technological support during all the steps. Biomax developed a knowledge management portal for the project (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 (e.g. stated gender was checked with available experimental data on chromosomal information).

Statistical Analysis

For each cohort we report numbers and percentages of all harmonized variables, including all available follow-up periods, by the different qualification categories (*complete*, *partial,* and *impossible*), before and 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 Supplementary table S2 for all proposed reference definitions together with modifications), and classified the inferential equivalence to the reference definition of 3551 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. 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, 2206 variables (62%) were qualified as *complete*, 1243 (35%) as *partial,* and 102 (3%) as *impossible.* After the workshop 2481 (70%) were qualified as *complete*, 550 (15%) *partial,* and 520 (15%) *impossible* (table 2). 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% for *complete* classifications, 13% for *partial,* and 14% for *impossible*. The proportion of *complete* was higher (79%) in the environmental exposures dimension and lower in the treatment (57%) and physical activity (40%) dimensions. More than 40% of variables in the physical activity dimension were classified as *impossible*. Final classifications for all included variables are available in supplementary figures SF1 to SF13. 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 the younger cohorts than for those from the older ones. 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 [13, 17] group. After six months of preparation and a four-day workshop we have agreed on 137 reference definitions and classified their inferential equivalence to 3551 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

To our knowledge, apart from the DataSHaPER [13, 14, 17] this is the first manuscript providing details on the harmonization procedure of data from a large consortium of different birth cohort across Europe on allergic disease. Of note, similar initiatives have now resulted

in evaluation [38] of harmonized outcome measures for atopic eczema (HOME) [39]. A special feature of our harmonization process is that it was not driven by a single or few specific research questions, but it rather integrated a broader spectrum of them to approach multiple explorative analysis [40, 41], with the potential for associations to omics results [42].

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 overall kappa coefficient of 0.49, in variable qualification before and after the workshop (moderate agreement), suggests that decisions on harmonization of relevant variables, of a given research question, would had been different if taken by individual experts as compared with a full group involved in a standardized harmonization exercise. Our approach overcomes both waste of time and reliability for 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 [43]. In general, no significant differences in results have been found between meta- and pooled analyses although pooled analysis exhibits higher precision of estimates [44, 45, 46]. 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

A strength of the present work is the use of a technological support (the MeDALL knowledge management portal) that includes all reference definitions, variables, and codification as well as all the knowledge used in order to take decisions. Existing long term collaboration of most birth cohorts starting with the GA^{2}LEN initiative [2, 3], and continued through the ENRIECO [7] and CHICOS [9] projects were fundamental to this commitment and to establish a birth-cohort alliance in the HELIX project [47] 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 [42]. The harmonization process involved a panel of multidisciplinary experts including medical, epidemiological, psychological, biostatistical, data management, and IT experts.

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 [20-22,24-26, 28, 31, 33] while others focused on psychological factors [20-22, 25-27]. 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.

Funding: 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 agreement number 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

REFERENCES

[1]Bousquet J, Gern JE, Martinez FD, et al. Birth cohorts in asthma and allergic diseases: Report of a NIAID, NHLBI, MeDALL joint workshop. J Allergy Clin Immunol 2014;133(6):1535-46.

[2]Keil T, Kulig M, Simpson A, et al. European birth cohort studies on asthma and atopic diseases: I. Comparison of study designs -- a GALEN initiative. Allergy 2006;61(2):221-8.

[3]Keil T, Kulig M, Simpson A, et al. European birth cohort studies on asthma and atopic diseases: II. Comparison of outcomes and exposures--a GALEN initiative. Allergy 2006;61(9):1104-11.

[4]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):969-77.

[5]Eller E, Roll S, Chen CM, et al. Meta-analysis of determinants for pet ownership in 12 European birth cohorts on asthma and allergies: a GA2LEN initiative. Allergy 2008;63(11):1491-8.

[6]Lødrup-Carlsen K, Roll S, Carlsen K, et al. Does pet ownership in infancy lead to asthma or allergy at school age? Pooled analysis of individual participant data from 11 European birth cohorts. Plos One 2012;7(8):e43214.

[7]Tischer CG, Hohmann C, Thiering E, et al. Meta-analysis of mould and dampness exposure on asthma and allergy in eight European birth cohorts: an ENRIECO initiative. Allergy 2011;66(12):1570-9.

[8]Vrijheid M, Casas M, Bergstrom A, et al. European birth cohorts for environmental health research. Environ Health Perspect 2012;120(1):29-37.

[9]Bousquet J, Anto J, Sunyer J, et al. Pooling birth cohorts in allergy and asthma: European Union funded initiatives A MeDALL, CHICOS, ENRIECO, and GA²LEN joint paper. Int Arch Allergy Immunol 2013;161(1):1-10.

[10]Bousquet J, Anto J, Auffray C, et al. MeDALL (Mechanisms of the Development of ALLergy): an integrated approach from phenotypes to systems medicine. Allergy 2011;66(5):596-604.

[11]Bousquet J, Anto JM, Akdis M, et al. Paving the way of systems biology and precision medicine in allergic diseases: The MeDALL success story. Allergy 2016;71(11):1513-25.

[12]Anto JM, Bousquet J, Akdis M, et al. Mechanisms of the Development of Allergy (MeDALL): Introducing novel concepts in allergy phenotypes. J Allergy Clin Immunol 2017;139(2):388-99.

[13]Fortier I, Burton PR, Robson PJ, et al. Quality, quantity and harmony: the DataSHaPER approach to integrating data across bioclinical studies. Int J Epidemiol 2010;39(5):1383-93.

[14]Fortier I, Parminder R, Van den Heuvel ER, et al. Maelstrom Research guidelines for rigorous retrospective data harmonization. Int J Epidemiol 2017;46(1):103-5.

[15]Navis GJ, Blankestijn PJ, Deegens J, et al. The Biobank of Nephrological Diseases in the Netherlands cohort: the String of Pearls Initiative collaboration on chronic kidney disease in the university medical centers in the Netherlands. Nephrol Dial Transplant 2014;29(6):1145-50.

[16]Doiron D, Burton P, Marcon Y, et al. Data harmonization and federated analysis of population-based studies: the BioSHaRE project. Emerg Themes Epidemiol 2013;10:12.

[17]Fortier I, Doiron D, Little J, et al. Is rigorous retrospective harmonization possible? Application of the DataSHaPER approach across 53 large studies. Int J Epidemiol 2011;40(5):1314-28.

[18]Maier D, Kalus W, Wolff M, et al. Knowledge management for systems biology a general and visually driven framework applied to translational medicine. BMC Syst Biol 2011;5:38.

[19]Antó JM, Pinart M, Akdis M, et al. Understanding the complexity of IgE-related phenotypes from childhood to young adulthood: a Mechanisms of the Development of Allergy (MeDALL) seminar. J Allergy Clin Immunol 2012;129(4):943-54.e4.

[20]Guxens M, Ballester F, Espada M, et al. Cohort Profile: the INMA—Infancia y Medio Ambiente—(Environment and Childhood) Project. Int J Epidemiol 2012;41(4):930-40.

[21]Ballardini N, Kull I, Lind T, et al. Development and comorbidity of eczema, asthma and rhinitis to age 12: data from the BAMSE birth cohort. Allergy 2012;67(4):537-44.

[22]Wickman M, Kull I, Pershagen G, Nordvall SL. The BAMSE project: presentation of a prospective longitudinal birth cohort study. Pediatr Allergy Immunol 2002;13(Suppl 15):11–13.

[23]Kjaer HF, Eller E, Høst A, Andersen KE, Bindslev-Jensen C. The prevalence of allergic diseases in an unselected group of 6-year-old children. The DARC birth cohort study. Pediatr Allergy Immunol 2008;19(8):737-45.

[24]Lødrup Carlsen KC. The environment and childhood asthma (ECA) study in Oslo: ECA-1 and ECA-2. Pediatr Allergy Immunol 2002;13(Suppl 15):29-31.

[25]von Berg A, Krämer U, Link E, et al. Impact of early feeding on childhood eczema: development after nutritional intervention compared with the natural course - the GINIplus study up to the age of 6 years. Clin Exp Allergy 2010;40(4):627-36.

[26]Zutavern A, Brockow I, Schaaf B, et al. Timing of solid food introduction in relation to eczema, asthma, allergic rhinitis, and food and inhalant sensitization at the age of 6 years: results from the prospective birth cohort study LISA. Pediatrics 2008;121(1):e44-52.

[27]Bergmann RL, Bergmann KE, Lau-Schadensdorf S, et al. Atopic diseases in infancy. The German multicenter atopy study (MAS-90). Pediatr Allergy Immunol 1994;5(6 Suppl):19-25.

[28]Wijga AH, Kerkhof M, Gehring U, et al. Cohort profile: the prevention and incidence of asthma and mite allergy (PIAMA) birth cohort. Int J Epidemiol 2014;43(2):527-35.

[29]Wright J, Small N, Raynor P, et al. Cohort profile: the born in Bradford multi-ethnic family cohort study. Int J Epidemiol 2013;42(4):978-91.

[30]Drouillet P, Forhan A, De Lauzon-Guillain B, et al. Maternal fatty acid intake and fetal growth: evidence for an association in overweight women. The “EDEN mother-child”

cohort (study of pre- and early postnatal determinants of the child's development and health). Br J Nutr 2009; 101(4), 583–591.

[31]Clarisse B, Nikasinovic L, Poinsard R, Just J, Momas I. The Paris prospective birth cohort study: which design and who participates? Eur J Epidemiol 2007;22(3):203-10.

[32]Chatzi L, Leventakou V, Vafeiadi,M, et al. Cohort Profile: The Mother-Child Cohort in Crete, Greece (Rhea Study). Int J Epidemiol 2017. [Epub ahead of print].

[33]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;32:350–57.

[34]Asher MI, Keil U, Anderson HR, et al. International Study of Asthma and Allergies in Childhood (ISAAC): rationale and methods. Eur Respir J 1995;8(3):483-91.

[35]Hohmann C, Pinart M, Tischer C, et al. The development of the MeDALL Core Questionnaires for a harmonized follow-up assessment of eleven European birth cohorts on asthma and allergies. Int Arch Allergy Immunol 2014;163(3):215-24.

[36]Pison C, Magnan A, Botturi K, et al. Prediction of chronic lung allograft dysfunction: a systems medicine challenge. Eur Respir J 2014;43(3):689-93.

[37]Burrowes KS, De Backer J, Smallwood R, et al. Multi-scale computational models of the airways to unravel the pathophysiological mechanisms in asthma and chronic obstructive pulmonary disease (AirPROM). Interface Focus 2013;3(2):20120057.

[38]Gerbens LA, Prinsen CA, Chalmers JR, et al. Evaluation of the measurement properties of symptom measurement instruments for atopic eczema: a systematic review. Allergy 2017;72(1):146-63.

[39]Chalmers JR, Simpson E, Apfelbacher CJ, et al. Report from the fourth international consensus meeting to harmonize core outcome measures for atopic eczema/dermatitis clinical trials (HOME initiative). Br J Dermatol 2016;175(1):69-79.

[40]Uphoff EP, Bird PK, Antó JM, et al. Variations in the prevalence of childhood asthma and wheeze in MeDALL cohorts in Europe. ERJ Open Res 2017;3(3):00150-2016.

[41]Gehring U, Wijga AH, Hoek G, et al. Exposure to air pollution and development of asthma and rhinoconjunctivitis throughout childhood and adolescence: a population-based birth cohort study. Lancet Respir Med 2015;3(12):933-42.

[42]Guerra S, Melén E, Sunyer J, et al. Genetic and epigenetic regulation of YKL-40 in childhood. J Allergy Clin Immunol 2017; [Epub ahead of print].

[43]Keller T, Hohmann C, Standl M, et al. The sex-shift in single disease and multimorbid asthma and rhinitis during puberty – a study by MeDALL.Allergy 2017; [Epub ahead of print].

[44]Hohmann C, Govarts E, Bergström A, et al. Joint Data Analyses of European Birth Cohorts: Two Different Approaches. WebmedCentral EPIDEMIOLOGY 2012;3(12):WMC003869.

[45]Yoshida K, Radner H, Kavanaugh A, et al. Use of data from multiple registries in studying biologic discontinuation: challenges and opportunities. Clin Exp Rheumatol 2013;31(4 Suppl 78):S28-32.

[46]Gehring U, Wijga AH, Hoek G, et al. Exposure to air pollution and development of asthma and rhinoconjunctivitis throughout childhood and adolescence: a population-based birth cohort study. Lancet Respir Med 2015;3(12):933-42.

[47]Vrijheid M, Slama R, Robinson O, et al. The human early-life exposome (HELIX): project rationale and design. Environ Health Perspect 2014;122(6):535-44.

Table 1: Example of a reference definition and paring rules to classify the inferential equivalence of each original cohort variable to this reference definition, as part of the harmonization process of asthma and allergy data in European birth cohorts

Variable Name: wheezing after exercise last 12 months **Reference Definition**: In the past 12 months, has your child's chest sounded wheezy during or after exercise? (Yes/No)		
Inferential equivalence classification (qualification)	**Definition provided by birth cohorts**	**Pairing rules**
Complete	Has your child had wheezing or **whistling** in the chest during or after exercise in the last 12 months Has your child ever had wheeziness when playing or when outdoors with/without having a cold? Has your child had wheeziness when playing or when outdoors with/without having a cold after the age of one year? In the past 12 months, has running around ever made your child's wheezy? - In the past 12 months, in which of the following situations	- Synonyms for "wheezing" are accepted as they are language and cultural specific - The timing of wheezing relative to exercise can be either during or after it. - 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". - Before the age of 2 years "playing or when outdoors" are considered as "exercise" (question asked at follow-up age two years or earlier). - This question is asked at three and four years of age, it was judged by the panel that "running around" at these ages is equivalent to exercise. - Though in some cases the wording is different, all

	your child has had whistling, wheezy sound of breathing during or after exercise? - Has your child´s breathing ever sounded wheezy during exertion during the past 12 months? - Has your child had wheezing or raspy breathing in conjunction with physical exertion in the last 12 months? - Did exercise impair wheezing in the last 12 months?	these definitions are judged as equivalent.
Partial	Has your child had trouble breathing in connection with exertion in the past 12 months?	- 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.
Impossible	In the past 24 months, has your child's chest sounded wheezy during or after exercise? - Has your child ever sounded like that (wheezing and whistling) after exercise? - Has your child ever sounded like that after exercise?	- The timeframe from this definition is broader than the one asked in the reference definition, 24 months *vs* 12 months respectively. - The timeframe from these definitions is broader than the one asked in the reference definition, ever *vs* 12 months respectively.

Table 2: Distribution by cohort of variables inferential equivalence classification before and after the harmonization workshop

		Before Workshop			After the Workshop			Kappa
	nº definitions*	Complete	Partial	Impossible	Complete	Partial	Impossible	
Older birth cohorts								
Amics- Menorca	422	344 (82)	78 (19)	0 (0)	349 (83)	21 (5)	52 (12)	0.44
BAMSE	219	119 (54)	100 (46)	0 (0)	127 (58)	44 (20)	48 (22)	0.43
ECA	304	232 (76)	60 (20)	12 (4)	225 (74)	28 (9)	51 (17)	0.53
GINIplus	338	108 (32)	210 (62)	20 (6)	172 (51)	92 (27)	74 (22)	0.43
LISAplus	335	100 (30)	230 (69)	5 (2)	182 (54)	1009 (33)	44 (13)	0.37
MAS	393	205 (52)	185 (47)	3 (1)	253 (64)	76 (19)	64 (16)	0.47
PIAMA	420	290 (69)	128 (31)	2 (1)	335 (80)	32 (8)	53 (13)	0.32
Total for older birth cohorts	2431	1398 (58)	991 (41)	42 (2)	1643 (68)	402 (17)	386 (9)	0.46
Younger birth cohorts								
BIB	150	95 (63)	46 (31)	9 (6)	114 (76)	29 (19)	7 (5)	0.69
EDEN	150	94 (63)	48 (32)	8 (5)	100 (67)	11 (7)	39 (26)	0.55
INMA-Sabadell	114	60 (53)	53 (47)	1 (1)	68 (60)	25 (22)	21 (18)	0.35
PARIS	401	349 (87)	38 (10)	14 (4)	346 (86)	33 (8)	22 (6)	0.76
RHEA	119	84 (71)	35 (29)	0 (0)	91 (77)	18 (15)	10 (8)	0.56
ROBBIC-Bologna	72	61 (85)	11 (15)	0 (0)	48 (67)	10 (14)	14 (19)	0.58
ROBBIC-Roma	114	65 (57)	21 (18)	28 (25)	71 (62)	22 (19)	21 (18)	0.50
Total for younger birth cohorts	1120	808 (72)	252 (23)	60 (5)	838 (75)	148 (13)	134 (12)	0.56
Total	3551	2206 (62)	1243 (35)	102 (3)	2481 (70)	550 (16)	520 (15)	0.49

*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.

FIGURE LEGENDS AND FOOTNOTES

Figure 1: Flow chart of the harmonization process of asthma and allergy variables in 17 European birth cohorts

Footnote

*ISAAC and MeDALL core questionnaires in the current study; others depending on the scientific research question.

†The duration of the workshop depends on the number of proposed reference definitions, involved cohorts, and follow-up periods.

Figure 2: Distribution of inferential equivalence classification of cohort variables to reference definitions, overall and by variables dimensions

Footnote

Figures SF1 to SF14 in the supplementary material include the distribution of inferential equivalence classification for each variable, as follows: Symptoms: asthma and wheezing (figure SF1), rhinitis (figure SF2), eczema (figure SF3), other allergic related variables (figure SF4), family history of allergic diseases (figure SF5), and puberty (figure SF6); Treatment: treatments for allergic diseases in the last 12 months (figure SF7), doctor consultations for allergic diseases in the last 12 months (figure SF8), triggers of allergic diseases in the last 12 months (figure SF9), school or outdoor activities absenteeism due to allergic diseases in the last 12 months (figure SF10); Environmental exposures: indoor (figure SF11), and smoking (figure SF12) exposures; Sociodemographic: siblings and other children at home (figure SF13); and Physical Activity: type, intensity, and period of physical activity (figure SF14).

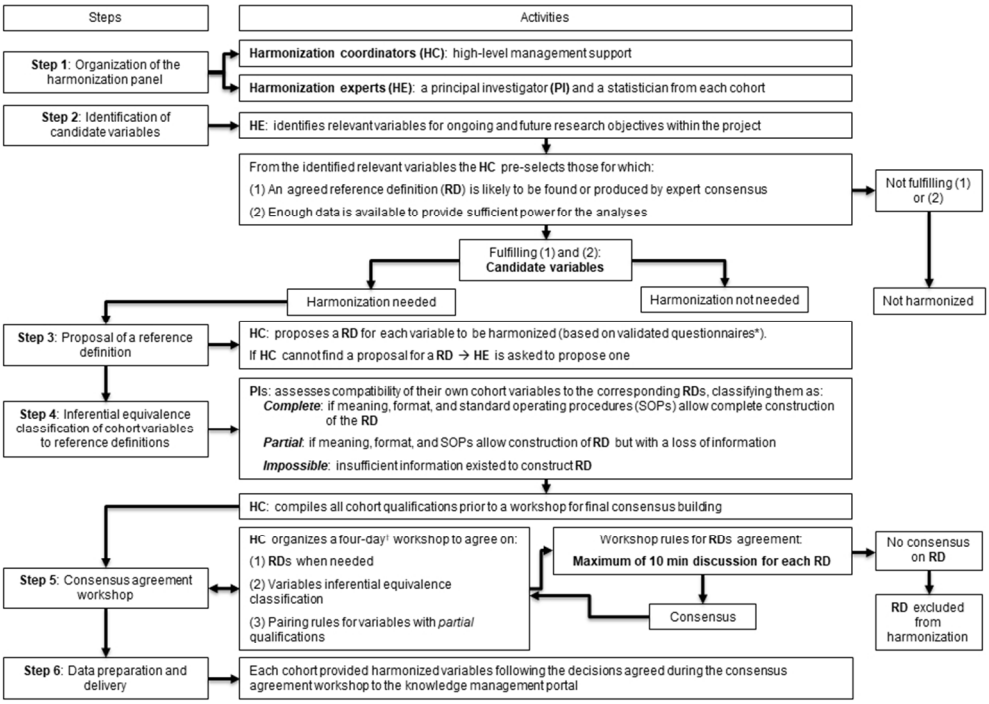

Figure 1: Flow chart of the harmonization process of asthma and allergy variables in 17 European birth cohorts

Footnote

*ISAAC and MeDALL core questionnaires in the current study; others depending on the scientific research question.

†The duration of the workshop depends on the number of proposed reference definitions, involved cohorts, and follow-up periods.

254x190mm (96 x 96 DPI)

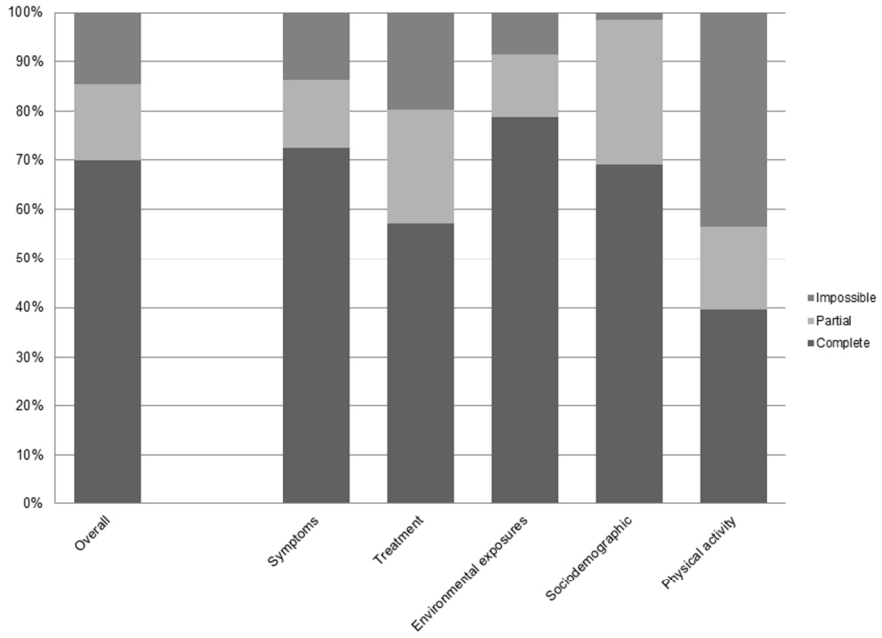

Figure 2: Distribution of inferential equivalence classification of cohort variables to reference definitions, overall and by variables dimensions

Footnote

Figures SF1 to SF14 in the supplementary material include the distribution of inferential equivalence classification for each variable, as follows: Symptoms: asthma and wheezing (figure SF1), rhinitis (figure SF2), eczema (figure SF3), other allergic related variables (figure SF4), family history of allergic diseases (figure SF5), and puberty (figure SF6); Treatment: treatments for allergic diseases in the last 12 months (figure SF7), doctor consultations for allergic diseases in the last 12 months (figure SF8), triggers of allergic diseases in the last 12 months (figure SF9), school or outdoor activities absenteeism due to allergic diseases in the last 12 months (figure SF10); Environmental exposures: indoor (figure SF11), and smoking (figure SF12) exposures; Sociodemographic: siblings and other children at home (figure SF13); and Physical Activity: type, intensity, and period of physical activity (figure SF14).

254x190mm (96 x 96 DPI)