American Journal of Epidemiology

Vol. 188, No. 2
DOI: 10.1093/aje/kwy242
Advance Access publication:
October 22, 2018

Practice of Epidemiology

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

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 Gehring, Beatrix Gerhard, Davide Gori, Eva Hallner, Inger Kull, Jacopo Lenzi, Rosemary McEachan, Eleonora Minina, Isabelle Momas, 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*

* Correspondence to Dr. Judith Garcia-Aymerich, ISGlobal (Barcelona Institute for Global Health), Doctor Aiguader 88, 08003 Barcelona, Spain (e-mail: judith.garcia@isglobal.org).

Initially submitted April 24, 2018; accepted for publication October 16, 2018.

The numbers of international collaborations among birth cohort studies designed to better understand asthma and allergies have increased in the last several years. However, differences in definitions and methods preclude direct pooling of original data on individual participants. As part of the Mechanisms of the Development of Allergy (MeDALL) Project, we harmonized data from 14 birth cohort studies (each with 3–20 follow-up periods) carried out in 9 European countries during 1990–1998 or 2003–2009. The harmonization process followed 6 steps: 1) organization of the harmonization panel; 2) identification of variables relevant to MeDALL objectives (candidate variables); 3) proposal of a definition for each candidate variable (reference definition); 4) assessment of the compatibility of each cohort variable with its reference definition (inferential equivalence) and classification of this inferential equivalence as *complete*, *partial*, or *impossible*; 5) convocation of a workshop to agree on the reference definitions and classifications of inferential equivalence; and 6) preparation and delivery of data through a knowledge management portal. We agreed on 137 reference definitions. The inferential equivalence of 3,551 cohort variables to their corresponding reference definitions was classified as *complete*, *partial*, and *impossible* for 70%, 15%, and 15% of the variables, respectively. A harmonized database was delivered to MeDALL investigators. In asthma and allergy birth cohorts, the harmonization of data for pooled analyses is feasible, and high inferential comparability may be achieved. The MeDALL harmonization approach can be used in other collaborative projects.

allergy; asthma; birth cohorts; data accuracy; data harmonization; data pooling; data sharing

Abbreviations: CHICOS, Developing a Child Cohort Research Strategy for Europe; DataSHaPER, Data Schema and Harmonization Platform for Epidemiologic Research; ENRIECO, Environmental Health Risks in European Birth Cohorts; FP6, Sixth Framework Programme for Research and Technological Development; FP7, Seventh Framework Programme for Research and Technological Development; GA^{2}LEN, Global Allergy and Asthma European Network; ISAAC, International Study of Asthma and Allergies in Childhood; MeDALL, Mechanisms of the Development of Allergy; PARIS, Pollution and Asthma Risk: An Infant Study; PIAMA, Prevention and Incidence of Asthma and Mite Allergy.

Over the past 30 years, more than 130 birth cohort studies with data on asthma and allergy have been initiated (1). The information gathered by these birth cohort studies has already significantly advanced our understanding of allergy and asthma, particularly during the first few years of life (2). However, these data are usually held in isolated, independent databases.

Although methods of assessing the data vary, the majority of the studies have followed rigorous methodology, and the resultant data are relatively readily available in electronic format.

Since 2004, the European Union's Sixth and Seventh Framework Programmes for Research and Technological Development (FP6 and FP7, respectively) have funded several projects designed to identify, compare, and evaluate pooling data from existing European birth cohort studies (Global Allergy and Asthma European Network (GA^2LEN) (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 cohort studies and have produced a large number of joint studies that have frequently used meta-analysis based on cohort original data (1, 12, 13). Though a few studies have integrated data from different birth cohorts in a single pooled analysis (7, 14), a formal reproducible approach for data harmonization has not been reported.

Several approaches for harmonizing data from different cohorts have been proposed (15–21). Among them, the Data Schema and Harmonization Platform for Epidemiological Research (DataSHaPER) Project (15) and the Maelstrom Research guidelines (16) have provided guidance aiming to facilitate rigorous, transparent, and effective data 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 FP7 MeDALL Project (2, 11, 12). We adapted the DataSHaPER approach and capitalized on the previous harmonization experience of 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 process included the use of questionnaire information collected on children in one of the 14 longitudinal population-based birth cohort studies. A total of 47,998 children of pregnant women or mothers with newborn babies recruited in 9 European countries were included (25). Seven of the studies recruited children between 1990 and 1998: the Asthma Multicentre Infant Cohort Study—Menorca (AMICS–Menorca), Spain (26); the Children Allergy, Milieu, Stockholm, Epidemiology (BAMSE) Study, Sweden (27, 28); the Environment and Childhood Asthma (ECA) Study in Oslo, Norway (29); the 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 (LISAplus) Study, Germany (31); the Multicenter Allergy Study (MAS), Germany (32); and Prevention and Incidence of Asthma and Mite Allergy (PIAMA), the Netherlands (33). The remaining 7 cohort studies included children recruited between 2003 and 2009: the Born in Bradford (BiB) Study, United Kingdom (34); the Study on the Pre- and Early Postnatal Determinants of Child Health and Development (EDEN), France (35); the Environment and Childhood Project—Sabadell (INMA–Sabadell), Spain (26); Pollution and Asthma Risk: An Infant Study (PARIS), France (36); the Rhea Mother-Child Cohort Study, Heraklion, Greece (37); Rome and Bologna Birth Italian Cohorts—Rome (ROBBIC–Roma) Study, Italy (38); and Rome and Bologna Birth Italian Cohorts—Bologna (ROBBIC–Bologna) Study, Italy (38). In all cohorts, parents gave written informed consent, and the studies were approved by local ethics review boards.

Variables

All birth cohort investigators collected information on participants for a minimum of 3 follow-up periods and a maximum of 20 follow-up periods (from pregnancy to 20 years of age) (see Web Table 1, available at https://academic.oup.com/aje). All of the researchers followed standardized protocols and included several validated questions regarding the outcome variables in their questionnaires, such as the one used in the International Study of Asthma and Allergies in Childhood (ISAAC) (39). Investigators followed strict quality control measures before, during, and after data collection to ensure the validity of the data collected.

Data harmonization process

The harmonization process was adapted from the DataSHaPER Project (15) and followed 6 steps (see Figure 1).

Step 1: organization of the harmonization panel. The data harmonization panel was formed by the harmonization coordinators and cohort experts. The harmonization coordinators were in charge of organizing the entire process, contacting investigators in each cohort study, and ensuring active participation of the cohort experts. These included, for each birth cohort, a principal investigator and a statistician or data manager who was 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 preselected those for which 1) an agreed-upon reference definition was likely to be found or produced by expert consensus and 2) enough data were available to provide sufficient statistical power for the envisioned analyses (i.e., at least 3 cohorts had data available for the variable). The candidate variables were then classified with regard to whether complex harmonization was needed or basic harmonization was needed (e.g., age, sex, and height). A total of 122 variables were preliminarily classified as "complex harmonization needed" and were allocated to one of 5 dimensions: 1) symptoms, 2) treatment, 3) environmental exposures, 4) sociodemographic factors, and 5) physical activity. (See a complete list of variables per dimension in Web Table 2.) A total of 28 variables were classified as "basic

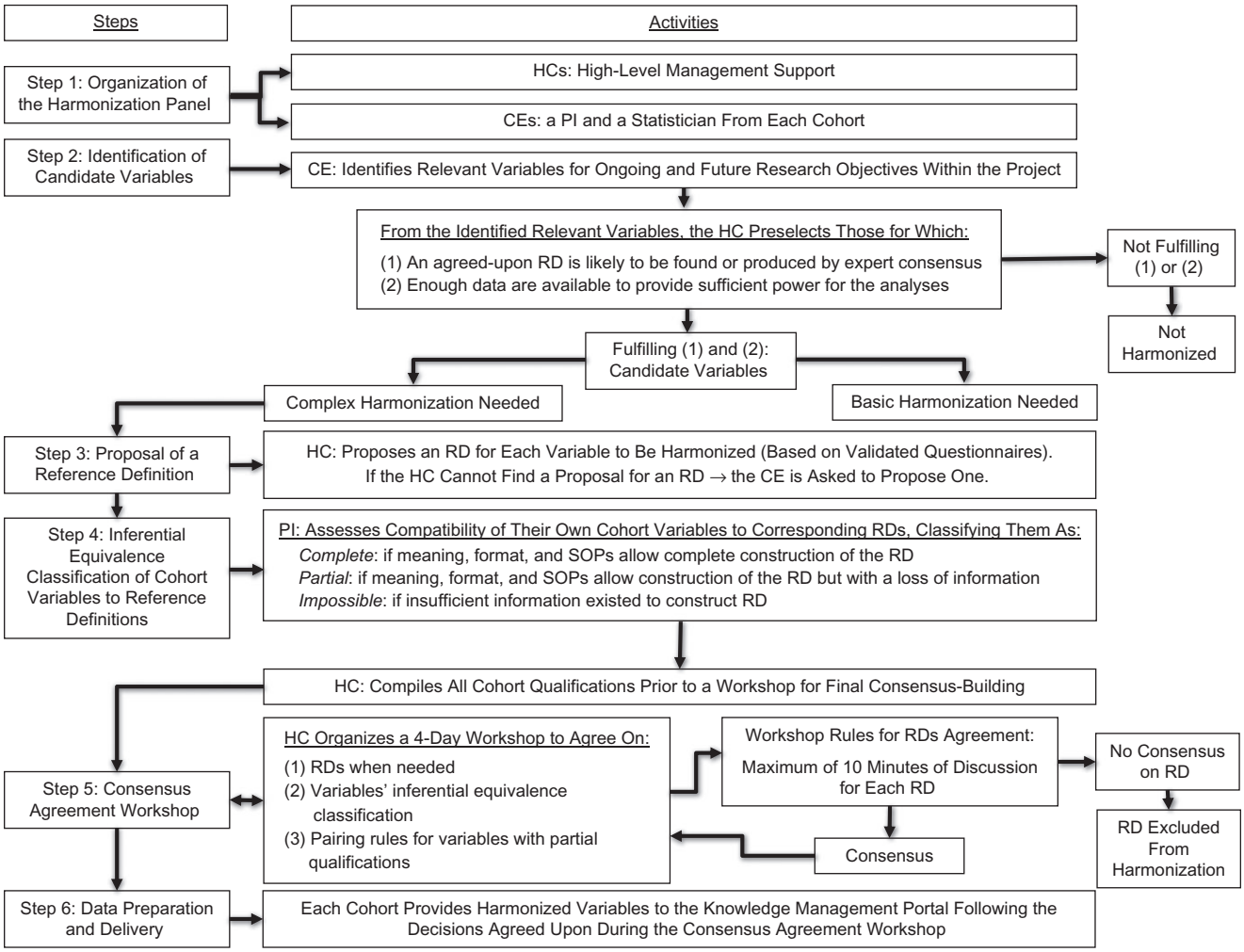

Figure 1. Process used for harmonization of data on asthma and allergy variables in 14 birth cohorts from 9 European countries, MeDALL Project, 2010–2015. CE, cohort expert; HC, harmonization coordinator; MeDALL, Mechanisms of the Development of Allergy; PI, principal investigator; RD, reference definition; SOPs, standard operating procedures.

harmonization needed." They covered basic demographic characteristics, early-life risk factors, and clinical information, and they 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 on the basis of 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 Web Table 2.

Step 4: Inferential equivalence classification of cohort variables to reference definitions. The principal investigator of each cohort study identified which question(s) matched each candidate variable in their cohort at the different follow-ups. Data for most questions (candidate variables) were collected during several follow-up periods (e.g., wheezing in the 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 with the corresponding reference definitions by assessing the meaning of, format of, and data collection procedure for 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 study, the inferential equivalence classification of that variable was considered 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) held for final consensus-building.

Step 5: Consensus agreement workshop. Harmonization coordinators organized a 4-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 to by the harmonization panel at the beginning of the workshop. For example, a maximum of 10 minutes was assigned for 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) was excluded from the final database. Notes were taken during the workshop by different participants and were checked by the harmonization coordinators for postworkshop quality control. The final reference definitions agreed upon can be found in Web Table 2.

Step 6: Data preparation and delivery. The investigators in each cohort study provided data on the harmonized variables to the knowledge management portal following the decisions agreed on during the workshop.

The MeDALL partner Biomax (Biomax Informatics AG, Planegg, Germany), a bioinformatics company with experience in systems medicine (24, 41, 42), provided dedicated technological support during all of above 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 of the data were integrated in the portal, where different algorithmic checks were performed to ensure data quality. Data-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?), a data type check (e.g., float, thesaurus), consistency checks of the collected data across follow-up periods (e.g., if a "yes" answer was provided for "ever receiving a physician's diagnosis of asthma," then responses to the same question in the following examinations were set to "yes"), and outlier detection. Units and codes were automatically converted and unified on the basis of unit ontology and code mapping. We also included logical checks where possible; for example, stated sex was confirmed with chromosomal information. Finally, data-quality descriptors were provided as summary statistics, including completeness, coverage (availability at different ages), and standard statistics for data distribution.

Statistical analysis

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

Cohen's κ coefficient was calculated to evaluate the agreement between the qualifications conducted by investigators in each cohort study before the workshop and the qualifications resulting from it. This coefficient was calculated overall, by cohort, by domain, and by variable.

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 2 or more definitions into 1), 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 on which data were collected across the multiple follow-ups of the 14 cohort studies.

Pairing rules

During the harmonization workshop, we agreed on the pairing rules for classifying the inferential equivalence of each variable to its reference definition. For example, a variable would result in a *complete* qualification if differences from the reference definition consisted of: 1) minor additional response categories (e.g., having the explicit missing-data option "don't know" or "don't wish to answer") or 2) equivalent methods of data generation (e.g., telephone interview vs. paper questionnaire). A *partial* qualification would result if: 1) minor language differences were found (e.g., a single synonym was not covered) or 2) a minor part of the definition was not asked about (e.g., "had an asthma attack" instead of "ever had an asthma attack"). Finally, an *impossible* qualification would result if: 1) questions asked about different time frames (e.g., "at least 2 weeks" instead of "at least 6 months"); 2) variables had strongly more restrictive definitions than the reference definition (e.g., asking about a specific allergic reaction instead of asking about an allergic reaction in general); or 3) different methods of data generation had been used (e.g., physical activity data 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 upon 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 (https://ssl.biomax.de/medall) and can be provided upon request.

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%) were classified as *partial*, and 102 (3%) were classified 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 for which data were not available (missing) in a given cohort or period were not included in the denominator, since their inferential equivalence (complete, partial, or impossible) remained 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 classifications according to the 5

Table 1. Example of a Reference Definition[a] and Paring Rules Used to Classify the Inferential Equivalence of Each Original Cohort Variable[b] to the Reference Definition, as Part of the Process of Harmonizing Data on Asthma and Allergy in 14 Birth Cohorts From 9 European Countries, MeDALL Project, 2010–2015

Inferential Equivalence Classification (Qualification)	Definition Provided by Cohort Study Investigators	Pairing Rules
Complete	"Has your child had wheezing or whistling in the chest during or after exercise in the last 12 months?"	Synonyms for "wheezing" were accepted because they were language- and culture-specific.
		The timing of wheezing relative to exercise could be either during exercise or after it.
	"Has your child ever had wheeziness when playing or when outdoors with/without having a cold?"	All questions not specifying "in the last 12 months" but where the 12-month period was respected because of the follow-up time frame were considered to provide "complete" data.
	"Has your child had wheeziness when playing or when outdoors with/without having a cold after the age of 1 year?"	Before the age of 2 years, "playing or when outdoors" was considered to represent "exercise" (question asked at follow-up at age 2 years or earlier).
	"In the past 12 months, has running around ever made your child wheezy?"	This question was asked at 3 and 4 years of age; the panel judged that "running around" at these ages was equivalent to exercise.
	"In the past 12 months, in which of the following situations has your child had a whistling, wheezy sound of breathing during or after exercise?"	Though in some cases the wording was different, all of these definitions were judged as equivalent.
	"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?"	
Partial	"Has your child had trouble breathing in connection with exertion in the past 12 months?"	The symptoms regarding breathing difficulties asked about in this question were considered broader than the ones asked about 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?"	The time frame from this definition was broader than the one used in the reference definition (24 months vs. 12 months).
	"Has your child ever sounded like that (wheezing and whistling) after exercise?"	The time frame from these definitions was broader than the one used in the reference definition (ever vs. 12 months).
	"Has your child ever sounded like that after exercise?"	

Abbreviation: MeDALL, Mechanisms of the Development of Allergy.
[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 in the last 12 months.

variable dimensions mentioned above. The "symptoms" dimension was the closest to the overall classification, with 73% of the variables in this dimension being 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 shown in Web Figures 1–14. All variables, and their inferential equivalence classifications, have been integrated into the final MeDALL database in order to provide researchers with additional information with which to conduct sensitivity analyses and test for misclassification.

Agreement between inferential equivalence classifications before and during the workshop

The overall agreement between the inferential equivalence classifications assigned to all variables before the workshop by the cohort principal investigators and the final qualifications agreed upon during the workshop was 0.49; agreement ranged from 0.32 in the PIAMA cohort to 0.76 in the PARIS cohort (Table 2). In general, agreement was higher for variables from studies that recruited children between 2003 and 2009 than for variables from studies that recruited children between 1990 and 1998. Fair-to-moderate agreement was obtained for all 5 dimensions (0.40–0.50); data on agreement by dimension and variable are available from the authors upon request.

DISCUSSION

Main findings

The present MeDALL harmonization study showed that harmonization of databases from different European asthma and allergy birth cohorts is feasible and successful when following and adapting the steps reported by the DataSHaPER (15, 21) group. After 6 months of preparation and a 4-day workshop, we 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% were classified as either *partial* or *impossible*.

Table 2. Distribution of Variables' Inferential Equivalence Classifications, by Cohort, Before and After a Consensus Workshop Held to Harmonize Asthma and Allergy Data on 14 Birth Cohorts From 9 European Countries, MeDALL Project, 2010–2015

Cohort Study (Ordered by Year of Recruitment)	Recruitment Year	No. of Definitions[a]	Before Workshop						After Workshop						κ
			Complete		Partial		Impossible		Complete		Partial		Impossible		
			No.[b]	%	No.	%	No.	%	No.	%	No.	%	No.	%	
MAS	1990	393	205	52	185	47	3	1	253	65	76	19	64	16	0.47
ECA Study in Oslo	1992/1993	304	232	76	60	20	12	4	225	74	28	9	51	17	0.53
BAMSE Study	1994/1996	219	119	54	100	46	0	0	127	58	44	20	48	22	0.43
PIAMA Study	1996/1997	420	290	69	128	30	2	1	335	80	32	8	53	12	0.32
GINIplus	1996/1998	338	108	32	210	62	20	6	172	51	92	27	74	22	0.43
AMICS–Menorca	1997/1998	422	344	82	78	18	0	0	349	83	21	5	52	12	0.44
LISAplus Study	1997/1998	335	100	30	230	69	5	1	182	54	109	33	44	13	0.37
ROBBIC–Roma Study	2003/2004	114	65	57	21	18	28	25	71	62	22	19	21	19	0.50
EDEN Study	2003/2005	150	94	63	48	32	8	5	100	67	11	7	39	26	0.55
PARIS	2003/2006	401	349	87	38	9	14	4	346	86	33	8	22	6	0.76
ROBBIC–Bologna Study	2004/2005	72	61	85	11	15	0	0	48	67	10	14	14	19	0.58
INMA–Sabadell	2004/2007	114	60	53	53	46	1	1	68	60	25	22	21	19	0.35
Rhea Mother-Child Cohort Study	2007/2008	119	84	71	35	29	0	0	91	77	18	15	10	8	0.56
BiB Study	2007/2009	150	95	63	46	31	9	6	114	76	29	19	7	5	0.69
Total		3,551	2,206	62	1,243	35	102	3	2,481	70	550	15	520	15	0.49

Abbreviations: AMICS–Menorca, Asthma Multicentre Infant Cohort Study—Menorca; BAMSE, Children Allergy, Milieu, Stockholm, Epidemiology; BiB, Born in Bradford; ECA, Environment and Childhood Asthma; EDEN, Study on the Pre- and Early Postnatal Determinants 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, Environment 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; MeDALL, Mechanisms of the Development of Allergy; PARIS, Pollution and Asthma Risk: An Infant Study; PIAMA, Prevention and Incidence of Asthma and Mite Allergy; ROBBIC–Bologna, Rome and Bologna Birth Italian Cohorts—Bologna; ROBBIC–Roma, Rome and Bologna Birth Italian Cohorts—Rome.

[a] From a total of 122 requested variable definitions, the number of definitions per cohort depended on the number of follow-up periods for which data on each variable were available.

[b] Number of variables.

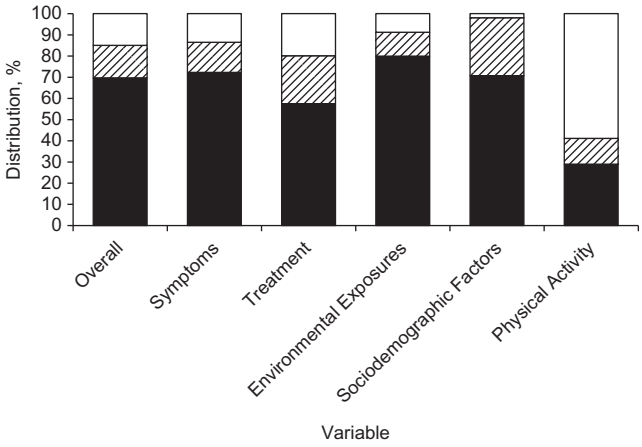

Figure 2. Distribution of the inferential equivalence classifications of cohort variables to reference definitions, overall and by variable dimension, in 14 birth cohorts from 9 European countries, MeDALL Project, 2010–2015. The figure shows the percentage of variables classified as complete (black), partial (lined), and impossible (white) among the total that required complex data harmonization. Web Figures 1–14 show the distribution of inferential equivalence classifications for each variable, as follows: symptoms—asthma and wheezing (Web Figure 1), rhinitis (Web Figure 2), eczema (Web Figure 3), other allergy-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), physician consultations for allergic diseases in the last 12 months (Web Figure 8), triggers of allergic diseases in the last 12 months (Web Figure 9), and absence from school or outdoor activities because of allergic diseases in the last 12 months (Web Figure 10); environmental exposures—indoor exposures (gas cooking, dampness, mold, pets) (Web Figure 11) and smoking exposures (Web Figure 12); sociodemographic factors—siblings and other children (parity at birth and number of other children in the home at the time of each examination) (Web Figure 13); and physical activity—type, intensity, and periods of physical activity (Web Figure 14). MeDALL, Mechanisms of the Development of Allergy.

Comparison with similar initiatives

This work supports and extends previous and ongoing initiatives on data harmonization (15–21, 43–46). Two novel features of our harmonization process were: 1) the consensus workshop as a key step that allowed discussion of and agreement on all reference definitions and inferential equivalences and 2) the broad spectrum of harmonized exposures and outcomes, which were not driven by a single specific research question but were integrated to eventually answer multiple research questions (47, 48), including questions on -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 for the same variables on multiple occasions for different analyses, 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 κ coefficient of 0.49) may have resulted from numerous issues that an individual expert would consider differently when thinking alone than when participating in group discussions. These aspects may include the 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 κ values should be interpreted not as a measure of the quality of the first inferential equivalence classification but as a marker of the complexity of data harmonization and the necessity of the harmonization process.

Our approach ensures a more efficient use of time and financial resources, improves the reliability of results of pooled analysis within the MeDALL Project, and allows performing meta-analyses with other projects' data with a clear framework for 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 (48, 51, 52). Since a big limitation of pooling data is heterogeneity, a harmonization process, such as the one reported here, will also facilitate strategies for pooling in the future.

Strengths and limitations

Strengths of the present work included the very rigorous process applied, which allows others to reproduce the approach; the use of technological support (the MeDALL knowledge management portal) that included all reference definitions, variables, and codification; and the expert knowledge used in order to make decisions. Existing long-term collaboration between most of the birth cohort studies, starting with the GA^{2}LEN initiative (3, 4) and continuing through the ENRIECO (8) and CHICOS (10) projects, was fundamental to this commitment and to the establishment of a birth-cohort alliance in the Human Early-Life Exposome (HELIX) Project (53), which links information on all of the environmental hazards mothers and children are exposed to with data on the health, growth, and development of children. Harmonized data from these cohort studies increase the range of exposures, increase the sample size and thus the statistical power of the studies, and allow for more detailed stratification. Therefore, in a collaborative project, the use of harmonized data (when performing either pooled analysis or meta-analysis) will increase the reproducibility, reliability, and validity of its results (49).

The harmonization process involved a panel of multidisciplinary experts, including medical, epidemiologic, psychological, biostatistical, data management, and information technology experts. Following harmonization, all MeDALL partners agreed to keep the data management portal up to date and active so that several research studies could be conducted and their results published (50, 54, 55). Finally, it was possible to harmonize data from 3 additional MeDALL birth cohort studies that did not participate in the first harmonization process thanks to the detailed harmonization reports available in the knowledge management portal (details available from Biomax upon request), 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 into 3 words: Giemen, Pfeifen, and Brummen). Second, the cohorts were heterogeneous regarding the spectrum and assessment methods of environmental and psychosocial exposures. For instance, some of the cohort studies had more detailed

questions on indoor environment than others (26–31, 33, 36, 38), while others focused on psychological factors (26–28, 30–32). Notably, data on some exposures and diseases could not be harmonized because of a large amount of heterogeneity or lack of data. Thus, the new common database created after performance of the MeDALL harmonization work does not yet include data on all variables, but it does include data on a large set of 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 that used the same variables. This is an area deserving of attention in future research. Finally, we did not consider between-country differences in intellectual property rights or ethical rules and regulations. Such considerations fall beyond the scope of a data harmonization exercise.

Conclusions

We have shown that data harmonization from different birth cohort studies and periods with cross-cultural differences is feasible and may achieve high comparability by using a predefined strategy, technological support, and commitments from all involved researchers. We encourage investigators in 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 regarding the effects of the data harmonization process on the validity of study results is needed.

ACKNOWLEDGMENTS

Author affiliations: ISGlobal (Barcelona Institute for Global Health), Barcelona, Spain (Marta Benet, Mariona Pinart, Christina G. Tischer, Lourdes Cirugeda, Manolis Kogevinas, Stefano Guerra, Jordi Sunyer, Josep M. Anto, Judith Garcia-Aymerich); Department of Experimental and Health Sciences, Universitat Pompeu Fabra, Barcelona, Spain (Marta Benet, Mariona Pinart, Christina G. Tischer, Lourdes Cirugeda, Manolis Kogevinas, Stefano Guerra, Jordi Sunyer, Josep M. Anto, Judith Garcia-Aymerich); Consorcio Centro de Investigación Biomédica en Red Epidemiología y Salud Pública, Barcelona, Spain (Marta Benet, Mariona Pinart, Christina G. Tischer, Lourdes Cirugeda, Maties Torrent, Manolis Kogevinas, Stefano Guerra, Jordi Sunyer, Josep M. Anto, Judith Garcia-Aymerich); Biomax Informatics AG, Planegg, Germany (Richard Albang, Beatrix Gerhard, Eleonora Minina, Dieter Maier); Hospital del Mar Research Institute, Barcelona, Spain (Mariona Pinart, Manolis Kogevinas, Jordi Sunyer, Josep M. Anto); Institute for Social Medicine, Epidemiology and Health Economics, Charité-Universitätsmedizin Berlin, Berlin, Germany (Cynthia Hohmann, Thomas Keil); Epidemiology of Allergic and Respiratory Diseases Department, Institut Pierre Louis d'Epidémiologie et de Santé Publique, Institut National de la Santé et de la Recherche Médicale, Paris, France (Isabella Annesi-Maesano, Nour Baïz); Saint-Antoine Medical School, Université Pierre et Marie Curie, Paris, France (Isabella Annesi-Maesano, Nour Baïz); Odense Research Center for Anaphylaxis, Department of Dermatology and Allergy Center, Odense University Hospital, Odense, Denmark (Carsten Bindslev-Jensen, Esben Eller); Department of Paediatric Allergy and Pulmonology, Division of Paediatric and Adolescent Medicine, Faculty of Medicine, University of Oslo and Oslo University Hospital, Oslo, Norway (Karin C. Lødrup Carlsen, Kai-Hakon Carlsen); Department of Biomedical and Neuromotor Sciences, Alma Mater Studiorum–University of Bologna, Bologna, Italy (Maria Pia Fantini, Davide Gori, Jacopo Lenzi); Institute for Risk Assessment Sciences, Utrecht University, Utrecht, the Netherlands (Ulrike Gehring); Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden (Eva Hallner); Centre for 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, United Kingdom (Rosemary McEachan, John Wright); Université Paris Descartes, Sorbonne Paris Cité, EA 4064 Epidémiologie Environnementale, Paris, France (Isabelle Momas, 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, Daniela Porta); School of Sport, Exercise and Health Sciences, Loughborough University, Loughborough, United Kingdom (Emily S. Petherick); Institute of Epidemiology I, Helmholtz Zentrum München–German Research Center for Environmental Health, Neuherberg, Germany (Marie Standl); Servei de Salut de les Illes Balears, 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, Arizona (Stefano Guerra); 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 Institut National de la Santé et de la Recherche Médicale, Unité Mixte de Recherche 1168 (Aging and Chronic Diseases: Epidemiological and Public Health Approaches), Villejuif, France (Jean Bousquet).

This work was supported by the Mechanisms of the Development of Allergy (MeDALL) Project, a collaborative study conducted within the European Union under the Health Cooperation Work Programme of the Seventh Framework Programme for Research and Technological Development (grant 261357).

The Barcelona Institute for Global Health (ISGlobal) is part of the Centres de Recerca de Catalunya (CERCA) Programme, Generalitat de Catalunya (Catalonia, Spain).

The sponsor of this study had no role in study design, data collection, data analysis, data interpretation, or the writing of the manuscript. The corresponding author (J.G.-A.) had full access to all study data and had final responsibility for the decision to submit the article for publication.

R.A., B.G., E.M., and D.M. are employed by Biomax Informatics AG (Planegg, Germany).

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):1535–1546.
2. 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):388–399.
3. Keil T, Kulig M, Simpson A, et al. European birth cohort studies on asthma and atopic diseases: I. Comparison of study designs—a GA^{2}LEN initiative. *Allergy*. 2006;61(2):221–228.
4. Keil T, Kulig M, Simpson A, et al. European birth cohort studies on asthma and atopic diseases: II. Comparison of outcomes and exposures—a GA^{2}LEN initiative. *Allergy*. 2006; 61(9):1104–1111.
5. Bousquet J, Burney PG, Zuberbier T, et al. GA^{2}LEN (Global Allergy and Asthma European Network) addresses the allergy and asthma 'epidemic'. *Allergy*. 2009;64(7):969–977.
6. 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 GA^{2}LEN initiative. *Allergy*. 2008;63(11): 1491–1498.
7. Lødrup-Carlsen KC, Roll S, Carlsen KH, 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.
8. 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):1570–1579.
9. Vrijheid M, Casas M, Bergström A, et al. European birth cohorts for environmental health research. *Environ Health Perspect*. 2012;120(1):29–37.
10. Bousquet J, Anto J, Sunyer J, et al. Pooling birth cohorts in allergy and asthma: European Union-funded initiatives—a MeDALL, CHICOS, ENRIECO, and GA^{2}LEN joint paper. *Int Arch Allergy Immunol*. 2013;161(1):1–10.
11. 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.
12. 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):1513–1525.
13. Bousquet J, Hellings PW, Agache I, et al. ARIA 2016: care pathways implementing emerging technologies for predictive medicine in rhinitis and asthma across the life cycle. *Clin Transl Allergy*. 2016;6:Article 47.
14. Neuman A, Hohmann C, Orsini N, et al. Maternal smoking in pregnancy and asthma in preschool children: a pooled analysis of eight birth cohorts. *Am J Respir Crit Care Med*. 2012; 186(10):1037–1043.
15. 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):1383–1393.
16. Fortier I, Raina P, Van den Heuvel ER, et al. Maelstrom Research guidelines for rigorous retrospective data harmonization. *Int J Epidemiol*. 2017;46(1):103–105.
17. Rolland B, Reid S, Stelling D, et al. Toward rigorous data harmonization in cancer epidemiology research: one approach. *Am J Epidemiol*. 2015;182(12):1033–1038.
18. Boffetta P, Bobak M, Borsch-Supan A, et al. The Consortium on Health and Ageing: Network of Cohorts in Europe and the United States (CHANCES) project—design, population and data harmonization of a large-scale, international study. *Eur J Epidemiol*. 2014;29(12):929–936.
19. 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):1145–1150.
20. 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:Article 12.
21. 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):1314–1328.
22. Lesko CR, Jacobson LP, Althoff KN, et al. Collaborative, pooled and harmonized study designs for epidemiologic research: challenges and opportunities. *Int J Epidemiol*. 2018; 47(2):654–668.
23. Hamilton CM, Strader LC, Pratt JG, et al. The PhenX Toolkit: get the most from your measures. *Am J Epidemiol*. 2011;174(3) 253–260.
24. 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:Article 38.
25. 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):943–954.e4.
26. 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):930–940.
27. 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):537–544.
28. Wickman M, Kull I, Pershagen G, et al. The BAMSE project: presentation of a prospective longitudinal birth cohort study. *Pediatr Allergy Immunol*. 2002;13(suppl 15):11–13.
29. 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.
30. Berg AV, 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):627–636.
31. 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–e52.
32. 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.

33. 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):527–535.
34. 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):978–991.
35. 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.
36. Clarisse B, Nikasinovic L, Poinsard R, et al. The Paris prospective birth cohort study: which design and who participates? *Eur J Epidemiol*. 2007;22(3):203–210.
37. Chatzi L, Leventakou V, Vafeiadi M, et al. Cohort profile: the Mother-Child Cohort in Crete, Greece (Rhea Study). *Int J Epidemiol*. 2017;46(5):1392–1393k.
38. 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(6):350–357.
39. 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):483–491.
40. 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):215–224.
41. Pison C, Magnan A, Botturi K, et al. Prediction of chronic lung allograft dysfunction: a systems medicine challenge. *Eur Respir J*. 2014;43(3):689–693.
42. 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.
43. 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):146–163.
44. 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.
45. Brody JA, Morrison AC, Bis JC, et al. Analysis commons, a team approach to discovery in a big-data environment for genetic epidemiology. *Nat Genet*. 2017;49(11):1560–1563.
46. Oelsner EC, Balte PP, Cassano PA, et al. Harmonization of respiratory data from 9 US population-based cohorts: the NHLBI Pooled Cohorts Study. *Am J Epidemiol*. 2018;187(11):2265–2278.
47. 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):pii: 00150-2016.
48. 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):933–942.
49. Guerra S, Melén E, Sunyer J, et al. Genetic and epigenetic regulation of YKL-40 in childhood. *J Allergy Clin Immunol*. 2018;141(3):1105–1114.
50. 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*. 2018;73(3):602–614.
51. Hohmann C, Govarts E, Bergström A, et al. Joint data analyses of European birth cohorts: two different approaches. *WebmedCentral*. 2012;3(12):pii: WMC003869.
52. 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–S32.
53. Vrijheid M, Slama R, Robinson O, et al. The Human Early-Life Exposome (HELIX): project rationale and design. *Environ Health Perspect*. 2014;122(6):535–544.
54. Thacher JD, Gehring U, Gruzieva O, et al. Maternal smoking during pregnancy and early childhood and development of asthma and rhinoconjunctivitis—a MeDALL project. *Environ Health Perspect*. 2018;126(4):047005.
55. Xu CJ, Söderhäll C, Bustamante M, et al. DNA methylation in childhood asthma: an epigenome-wide meta-analysis. *Lancet Respir Med*. 2018;6(5):379–388.