Received: 27 May 2020 | Revised: 22 September 2020 | Accepted: 23 September 2020

DOI: 10.1111/pai.13385

WILEY

ORIGINAL ARTICLE

Asthma in farm children is more determined by genetic polymorphisms and in non-farm children by environmental factors

Norbert Krautenbacher[1,2] | Michael Kabesch[3,4,5] | Elisabeth Horak[6] | Charlotte Braun-Fahrländer[7,8] | Jon Genuneit[9,10] | Andrzej Boznanski[11] | Erika von Mutius[5,12,13] | Fabian Theis[1,2] | Christiane Fuchs[1,2,14] | Markus J. Ege[5,12] | the GABRIELA, PASTURE study groups

[1]Helmholtz Zentrum München, German Research Center for Environmental Health, Institute of Computational Biology, Neuherberg, Germany

[2]Center for Mathematics, Chair of Mathematical Modeling of Biological Systems, Technische Universität München, Garching, Germany

[3]University Children's Hospital Regensburg (KUNO), Regensburg, Germany

[4]Clinic for Pediatric Pneumology and Neonatology, Hannover Medical School, Hannover, Germany

[5]The German Center for Lung Research (DZL), Germany

[6]Department of Pediatrics and Adolescents, Innsbruck Medical University, Innsbruck, Austria

[7]Swiss Tropical and Public Health Institute Basel, Basel, Switzerland

[8]University of Basel, Basel, Switzerland

[9]Institute of Epidemiology and Medical Biometry, Ulm University, Ulm, Germany

[10]Pediatric Epidemiology, Department of Pediatrics, Medical Faculty, Leipzig University, Leipzig, Germany

[11]Wroclaw Medical University, Wroclaw, Poland

[12]Dr von Hauner Children's Hospital, LMU Munich, Munich, Germany

[13]Helmholtz Zentrum München, German Research Center for Environmental Health, Institute of Asthma and Allergy Prevention, Neuherberg, Germany

[14]Department of Business Administration and Economics, Bielefeld University, Bielefeld, Germany

Correspondence
Markus J. Ege, Dr. von Hauner Children's Hospital, Ludwig Maximilians University Munich, Lindwurm Str. 4, 80337 Munich, Germany.
Email: markus.ege@med.lmu.de

Funding information
European Commission as part of GABRIEL (a multidisciplinary study to identify the genetic and environmental causes of asthma in the European Community); European Commission research, Grant/Award Number: QLK4-CT-2001-00250, FOOD-CT-2006-31708 and KBBE-2007-2-2-06; German Research Foundation; Federal Ministry of Education and Research, Grant/Award Number: 01DH17024; German Center for Lung Research (MJE)

Abstract

Background: The asthma syndrome is influenced by hereditary and environmental factors. With the example of farm exposure, we study whether genetic and environmental factors interact for asthma.

Methods: Statistical learning approaches based on penalized regression and decision trees were used to predict asthma in the GABRIELA study with 850 cases (9% farm children) and 857 controls (14% farm children). Single-nucleotide polymorphisms (SNPs) were selected from a genome-wide dataset based on a literature search or by statistical selection techniques. Prediction was assessed by receiver operating characteristics (ROC) curves and validated in the PASTURE cohort.

Results: Prediction by family history of asthma and atopy yielded an area under the ROC curve (AUC) of 0.62 [0.57-0.66] in the random forest machine learning approach.

Christiane Fuchs and Markus J. Ege equal contribution.

Editor: Ömer Kalaycı

By adding information on demographics (sex and age) and 26 environmental exposure variables, the quality of prediction significantly improved (AUC = 0.65 [0.61-0.70]). In farm children, however, environmental variables did not improve prediction quality. Rather SNPs related to IL33 and RAD50 contributed significantly to the prediction of asthma (AUC = 0.70 [0.62-0.78]).

Conclusions: Asthma in farm children is more likely predicted by other factors as compared to non-farm children though in both forms, family history may integrate environmental exposure, genotype and degree of penetrance.

KEYWORDS

childhood asthma, environment, farming, genome-wide association studies, machine learning, penalized regression, random forest, risk prediction, single-nucleotide polymorphisms, statistical learning

1 | INTRODUCTION

A meta-analysis of genome-wide association studies (GWAS) on childhood onset asthma yielded low predictive values with a rather low area under the receiver operating characteristic (ROC) curve of 0.58.[1] This modest result conflicted with the strong hereditary background postulated from twin studies.[2] Twins do not only share their genetic background but usually also their environment, and gene-environment interactions particularly with the farming environment seem to play an important role for childhood asthma.[3] This may suggest that prediction of asthma by genetic factors might be improved by considering environmental influences.

On the other hand, the overall statistical power of prediction is severely reduced by the multiplicity of the commonly applied univariate tests assessing 0.6-0.7 million independent loci in the human genome separately.[4] This is a conceptual limitation of the classical test theory and can hardly be overcome by increasing case numbers. In addition, univariate models ignore potentially important dependencies between loci.

The aim of the present study was to test whether prediction of childhood asthma by genetic determinants varies with the environmental setting, particularly the farm exposure. For this purpose, we applied the state-of-the-art statistical tools that consider predictor variables integratively and provide high statistical power.

Key Message

This study demonstrates that prediction of polygenic diseases such as childhood onset asthma by genome-wide data might be improved when considering environmental determinants. Populations with a homogeneous environmental exposure structure might be more suitable for predicting genetic risk in individuals. This is illustrated by farm children, in whom asthma was predominantly predicted by family history and genetic determinants, whereas in non-farm children environmental determinants and family history outperform genetic predictors.

2 | METHODS

2.1 | Population and questionnaires

The cross-sectional GABRIEL Advanced Studies (GABRIELA) was designed to study gene-environment interactions.[5] From the Austrian, Swiss, and German arms of GABRIELA, 1707 schoolchildren were randomly selected from all 34 491 children eligible for genotyping in a stratified design (Figure S1).[5,6] This approach was chosen to perform a case-control study with simultaneously enriching farm exposure. The outcome childhood asthma was defined as a physician diagnosis of asthma at least once or of asthmatic bronchitis at least twice.[7] The questionnaires contained items on individual and family health, socio-economic background and farm-related exposures. If a child lived on a farm run by the family, the child was termed "farm child" (n = 483) and "non-farm child" (n = 1224) when not living on a farm. Other farm-related exposures were related to raw milk consumption or contact with animals or animal feed. Those variables were included either as exposure in the first years of life or as exposure during the past 12 months.

The final model was externally validated in 928 children of the prospective PASTURE birth cohort. Both studies were approved by the respective local ethics committees. Written informed consent was obtained from parents or guardians.

2.2 | Genotyping

Genotyping was performed with the Illumina Human610 quad array (Illumina Inc, San Diego, Calif, http://www.illumina.com), and

quality was assessed as described previously.[1] SNPs were imputed by Markov chain-based haplotyper[8] using the 1000 Genomes pilot 1 release.[9] SNPs were filtered for imputation quality (Rsq ≥ 0.30) and minor allele frequency (MAF ≥ 0.05) and pruned for linkage disequilibrium by removing SNPs within a $5 \cdot 10^5$ SNP window that had $r^2 > 0.95$ resulting in 744 908 SNPs.[10]

Candidate SNPs (Table 1) were defined as SNPs included in the GWAS catalog for childhood onset asthma.[11]

2.3 | Computational and statistical analysis

All statistical analyses were performed with R software.[12] Details are provided in this article's Online Repository. R code is available at https://github.com/fuchslab/gabriela.

Environmental variables (Table S1) had <25% missing values. Missing values of variables were imputed by multiple imputation resulting in five imputation data sets,[13] which means that subsequent analyses were performed 5-fold and averaged.

Prediction was performed in the entire dataset and additionally in the two strata of farm and non-farm children. In addition to classical GWAS performing an association test univariately for each single SNP,[1,10] we incorporated all variables at once in multivariable statistical learning models with the following regularization methods: the least absolute shrinkage and selection operator (LASSO), elastic net, and the integrative L1-penalized regression with penalty factors (IPF-LASSO).[14,15] Additionally, random forests were built on 20 000 trees.[16]

Model selection and 5-fold cross-validation were performed on the 1410 Swiss and German participants. The best models were then externally validated in the 297 Austrian participants (the smallest centre) and additionally in PASTURE.

As a metric for model comparison, we applied the area under the receiver operating characteristics (ROC) curve (AUC) with a

TABLE 1 SNPs reported for childhood asthma in the GWAS catalog

SNP	Region	*P*-value*	Gene**	Comments
rs4658627	1q44	6×10^{-6}	C1orf100	chromosome 1 open reading frame 100
rs9815663	3p26.2	2×10^{-8}	IL5RA	interleukin 5 receptor subunit alpha
rs2705520	3q13.2	2×10^{-6}	ATG3	Autophagy Related 3
rs17033506	3p22.3	4×10^{-7}	(intergenic)	ARPP21: cAMP Regulated Phosphoprotein 21
rs9823506	3q12.2	6×10^{-8}	ABI3BP	ABI Family Member 3 Binding Protein
rs6871536	5q31.1	8×10^{-7}	RAD50	RAD50 Double Strand Break Repair Protein
rs1295686	5q31.1	2×10^{-6}	IL13	Interleukin 13
rs2473967	6q21	2×10^{-6}	(intergenic)	LOC105377956, LOC105377953
rs6967330	7q22.3	3×10^{-14}	CDHR3	Cadherin Related Family Member 3
rs9297216	8p12	1×10^{-6}	(intergenic)	LOC105379365
rs16929097	9p23	8×10^{-9}	(intergenic)	TYRP1 (Tyrosinase Related Protein 1)
rs11141597	9q21.33	2×10^{-6}	(intergenic)	LOC105376124/ GAS1
rs928413	9p24.1	9×10^{-13}	IL33	Interleukin 33
rs7927044	11q24.2	7×10^{-9}	(intergenic)	LOC107984373, LOC387820
rs7328278	13q13.3	3×10^{-6}	(not reported)	DCLK1 (Doublecortin Like Kinase 1)
rs10521233	17p12	3×10^{-6}	(intergenic)	LOC105371544, LOC107985014
rs2305480	17q21.1	6×10^{-23}	GSDMB	Gasdermin B
rs3894194	17q21.1	3×10^{-21}	GSDMA	Gasdermin A
rs7216389	17q21.1	9×10^{-11}	ORMDL3	ORMDL sphingolipid biosynthesis regulator 3

**P*-values for associations with childhood onset asthma are taken from the GWAS catalog.[11]

***Genes are reported by authors of original publications.[11] If no genes are reported, mapped genes are given in the comments column.

bootstrapped 95%—confidence interval.[17] The ROC curve plots sensitivity against 1—specificity; the AUC thus integrates measures of prediction quality. An AUC of 1.0 means perfect prediction, whereas an AUC of 0.5 reflects no prediction at all. If not indicated otherwise, AUC values refer to random forest models.

3 | RESULTS

The n = 850 cases and n = 857 controls included in the present analyses differed with respect to sex, family history of asthma and atopy, and various farm-related exposures (Table 2, Table S1).

When predicting asthma by groups of variables separately, that is family history, demographics (sex, age and BMI), environment and genetics, the explored multivariable learning approaches did not differ with respect to prediction quality (Figure 1, upper panel). Family history was the best predictor of childhood asthma with an AUC value of 0.62 [0.57-0.66] in the random forest model. All other groups of variables did not predict better than by chance except for environmental variables in the random forest model (AUC = 0.55 [0.51-0.59]). Findings were similar when restricting the model to non-farm children (Figure 1, middle panel). For farm children, however, a different prediction model emerged: instead of environmental variables, demographics and genome-wide SNPs (AUC = 0.61 [0.51-0.70]) predicted significantly (Figure 1, lower panel).

TABLE 2 Potential determinants of asthma

Characteristic	Cases (%) n = 850	Controls (%) n = 857	*P*-value*
Female sex	39.70%	49.40%	.002
age[a]	8.32 (0.06)	8.19 (0.06)	.150
Body mass index[a]	17.11 (0.11)	16.99 (0.11)	.375
Family history of atopy	70.00%	49.40%	<.001
Family history of asthma	30.06%	12.40%	<.001
Living on a farm	9.00%	13.60%	<.001
At least two siblings	0.42 (0.02)	0.45 (0.02)	.374
High parental education	27.30%	28.80%	.633
Maternal smoking during pregnancy	12.40%	8.50%	.037
Consumption of farm milk during past 12 mo	13.40%	19.40%	<.001
Consumption of farm milk in first year of life	6.20%	11.80%	<.001
Consumption of farm milk (pregnancy to age 3 y)	20.70%	27.60%	<.001
Contact with cows (past 12 mo)	12.90%	16.60%	.020
Contact with cows (pregnancy to age 3 y)	14.60%	20.30%	<.001
Contact with straw (past 12 mo)	15.70%	21.10%	.009
Contact with straw (pregnancy to age 3 y)	12.40%	16.20%	.009
Contact with hay (past 12 mo)	29.70%	33.50%	.145

Only variables are shown that appeared as most relevant in subsequent analyses. A complete list of environmental variables is provided in Table S1 in the Online Repository.

[a]Mean and standard error of mean.

**P*-values based on Fisher's exact test or, in case of continuous variables, Wilcoxon tests.

When complementing prediction models of asthma by family history with the other groups of variables, random forest and IPF-LASSO performed much better than simple LASSO (Figure 2, upper panel) and the other techniques (data not shown). Prediction by family history was significantly (Table S2) improved by demographics and environmental variables (AUC = 0.65 [0.61-0.70]) or, in case of farm children, by demographics and candidate SNPs (AUC = 0.70 [0.62-0.78]), whereas GWAS SNPs and interaction terms did not further improve prediction quality (Figure 2, lower panel).

Besides family history of asthma and atopy, age and sex, 26 environmental exposure variables such as contact to cats, dogs, cows, straw and hay importantly contributed to the random forest prediction model for all children (AUC = 0.64 [0.54-0.73], Figure 3, left panel) and non-farm children (AUC = 0.63 [0.53-0.72], Figure 3, centre panel). For farm children (Figure 3, right panel), we found, beyond family history and sex, three candidate SNPs, one of them intergenic. The two other SNPs are known to be related to IL33 and RAD50 (Table 1). Sensitivity analyses using IPF-LASSO confirmed the IL33 SNPs from the random forest prediction model (Figure S2) with an AUC of 0.86 [0.59-0.99] averaged over the prediction scores of random forest and IPF-LASSO (Figure S3). A sensitivity analysis revealed AUCs of 0.57 [0.51-0.64] and 0.55 [0.51-0.58] for prediction by candidate SNPs and demographics in all children with and without a family history of asthma, respectively.

External validation in the Austrian GABRIELA arm (Figure 4A) and the PASTURE birth cohort (Figure 4B, Table S3) confirmed the AUC values from the previously cross-validated random forest prediction model of asthma based on family history, demographics and environment. Sensitivity analyses yielded a better prediction quality for a model excluding individuals with current wheeze or asthma medication from the reference group (Figure 4C) and a model assigning children with recurrent obstructive bronchitis but without an established asthma diagnosis to the control group (Figure 4D).

4 | DISCUSSION

With the use of advanced statistical methods from the area of machine learning, which allow for multivariable consideration of predictors without susceptibility to multiple testing issues, performance of prediction was improved noticeably beyond the classical logistic regression approach. In combined models, prediction of asthma was driven by various environmental variables in addition to family history and sex, whereas candidate and genome-wide SNPs did not improve prediction. Only in farm children, genetic information

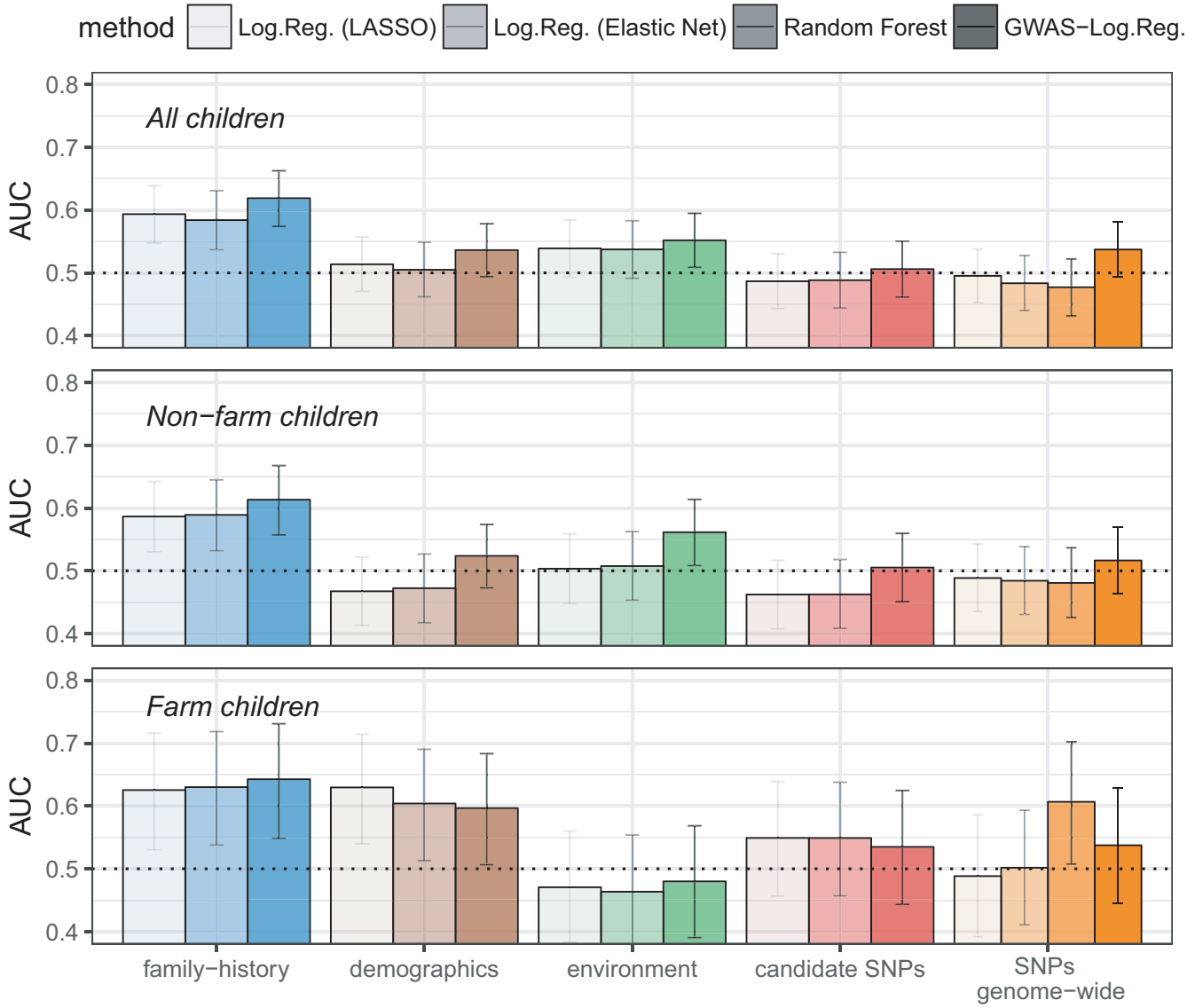

FIGURE 1 Comparison of prediction performance for different modalities and statistical methods for the prediction of childhood asthma. Prediction performance of the variable groups *family history*, *demographics*, *environment*, *candidate SNPs* (Table 1) and *genome-wide SNPs* on a stand-alone basis. As statistical methods, we used multivariable logistic regression with LASSO penalty, multivariable logistic regression with elastic net penalty, the random forest and, for genome-wide SNPs, multivariable logistic regression models. The AUC is calculated as mean over 5 imputation data sets with 95% confidence intervals constructed by bootstrap using selection probabilities. The dotted line at 0.5 corresponds to the AUC value where a prediction model classifies cases and controls not better than at random

contributed significantly to the prediction model, while environmental exposure did not add to prediction models in this group of children.

The GWAS of the last two decades were definitively a success when considering the discovery of new loci and the confirmation or invalidation of candidate genes.[18,19] However, prediction of polygenic disease such as asthma remains difficult on an individual level. Moffatt and colleagues already reported a low AUC of 0.58 for the seven top SNPs identified by their meta-analysis for childhood asthma.[1] However, the prediction model was fitted on the entire dataset leaving no independent sample for validation, which may have resulted in a too optimistic AUC. In our population, such an approach would have resulted in an AUC of 0.60 for GWAS SNPs, instead of 0.54 as reported in Figure 1. Otherwise, Moffatt and colleagues integrated only the top seven SNPs, that is the ones reaching genome-wide significance. Thereby, they disregarded information conveyed by additional SNPs and thus did not fully exploit the predictive power of a genome-wide approach.[19,20]

Therefore, we integrated all available genetic information by multivariable modelling and complemented that with questionnaire data on familial predisposition and strong environmental determinants. In addition, we applied random forest and various forms of penalized multivariable logistic regression such as LASSO, elastic net and IPF-LASSO. These models find an optimal trade-off between model

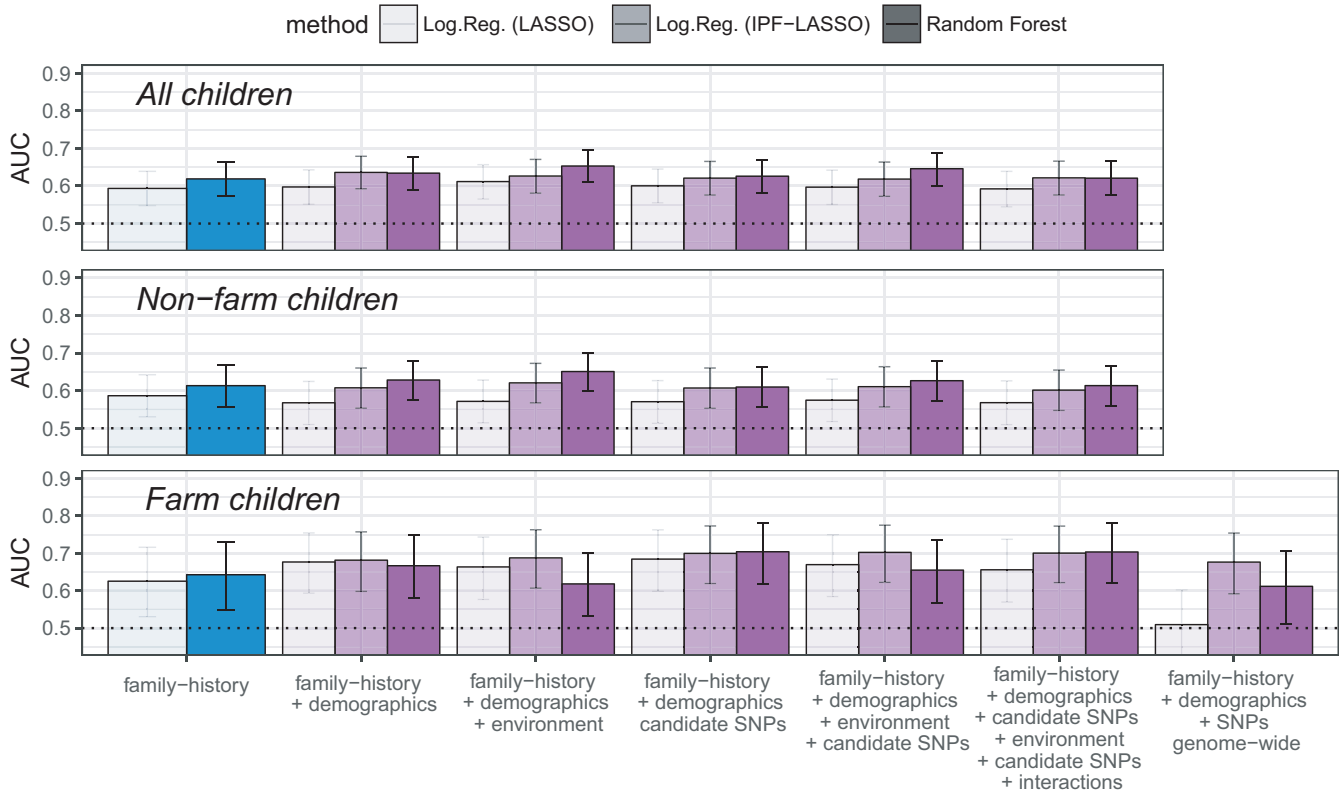

FIGURE 2 Prediction performance in predicting childhood asthma by family history and additional predictors for different statistical methods. Prediction performance for models based on *family history* alone and successively combined with *demographics*, *environment* or *candidate SNPs*, *environment* and *candidate SNPs*, *environment* interacting with *candidate SNPs*, and *demographics* plus *genome-wide SNPs*. Genome-wide SNPs were only assessed in farm children, since they were not predictive in non-farm children and the entire population (see Figure 1). As statistical methods, we used multivariable logistic regression with LASSO penalty, IPF-LASSO, and the random forest. The AUC is calculated as mean over 5 imputation data sets with 95% confidence intervals constructed by bootstrap using selection probabilities. The dotted line at 0.5 corresponds to the AUC value where a prediction model classifies cases and controls not better than at random

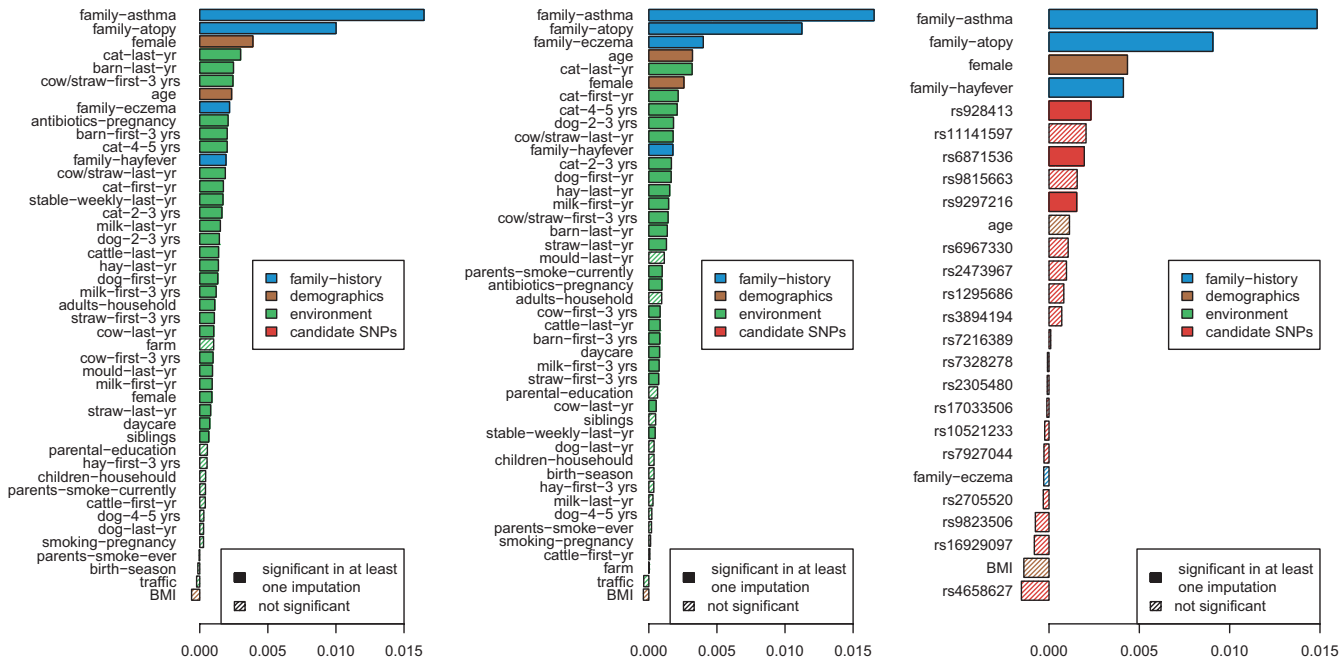

FIGURE 3 Importance of variables contributing to the best prediction models for childhood asthma. Variable importance determined by random forest models for all children and stratified for farm and non-farm children

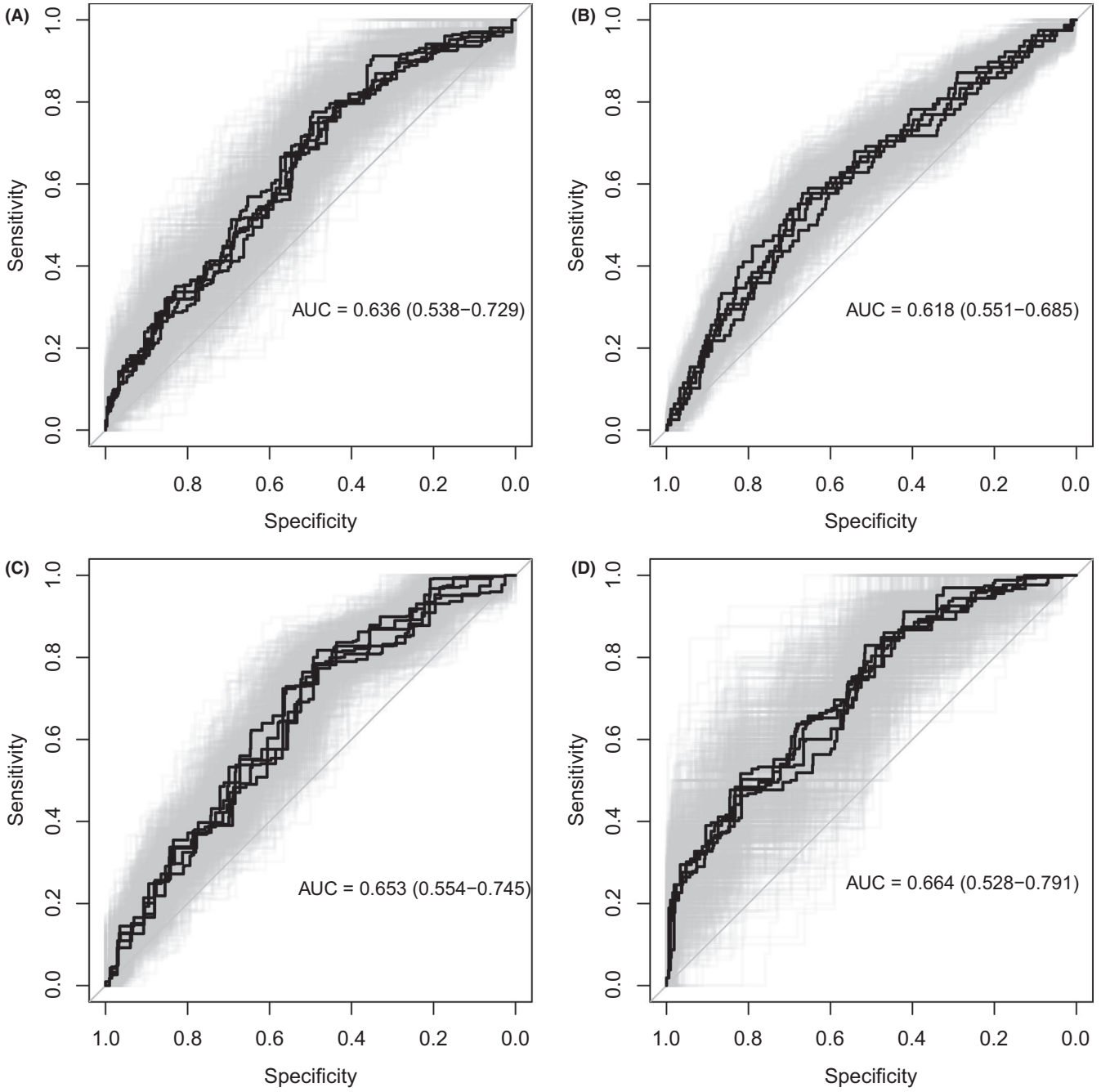

FIGURE 4 External validation of prediction models for childhood asthma. For external validation, individual ROC curves of the 5 imputations with averaged area under the curve (AUC) are shown. The internally validated random forest prediction model of a parent-reported doctor diagnosis of asthma once or obstructive bronchitis twice based on *family history*, *demographics* and *environment* yielded a AUC of 0.65 [0.61-0.70]. A, External validation in the Austrian arm of GABRIELA. B, External validation in the PASTURE birth cohort. C, Similarly to A, but excluding control individuals with current wheeze or use of asthma-specific drugs. D, Similarly to A, but for a parent-reported doctor-diagnosis of asthma once irrespectively of any obstructive bronchitis

complexity and the risk of overfitting; the latter might negatively impact external validity and thus predictive power. For comparison, we also applied a two-step approach creating a prediction score based on the 100 top hits of a previous simple logistic regression.[10] Prediction quality for asthma has already been shown to decrease beyond 100 SNPs.[21]

Given the rather weak genetic effects in polygenic diseases, the random forest prediction model by genome-wide SNPs with its AUC of 0.61 [0.51-0.70] in farm children is particularly remarkable (Figure 1, lower panel). It may reflect improved prediction by inclusion of SNPs beyond the genome-wide significance threshold. These non-significant SNPs might still be relevant for polygenic diseases and finally may help explaining the missing heritability.[19,20] On the other hand, 99.5% of the genome-wide SNPs did not significantly contribute to the prediction model and may have increased noise.[22] This may also apply to the candidate SNPs

from the GWAS catalogue, as some of them missed genome-wide significance (see Table 1).

When establishing combined prediction models based on several groups of variables, the genome-wide SNPs were replaced by candidate SNPs not among the top genome-wide SNPs and family history of asthma or other atopic diseases, which might be better proxies for predictive hereditary factors than the vast majority of genome-wide SNPs. Though genome-wide data include a lot of noise, family history and its effect on the index child are not free from misclassification and likewise affected by noise.

The insight that asthma runs in families is not trivial. Family history integrates a wealth of hereditary information though at much lower resolution as compared to genome-wide SNPs. Obviously, a family history may reflect shared environments such as the microbiome, which is clearly passed from mother to child.[23] Likewise, a family history may represent conditions during pregnancy, for example epigenetic mechanisms or an inflammatory status of the mother shaping the foetal immune system and thus contributing to disease transmission.[24] In conclusion, the simple question on a family history of asthma and atopy just integrates multifaceted information on several known environmental and genetic predictors and complements it with all the complexity of family life, which is captured neither by questionnaire records nor by genome-wide data.

The prediction models varied completely between farm and non-farm children with respect to genetics. Farm exposures may prevent many cases of asthma so that farm children might be affected mainly by genetically determined forms of asthma, which renders them an interesting population for genetic research. In more general terms, this notion may challenge the usefulness of populations with heterogeneous environmental exposures for analyses of GWAS.

Two of the SNPs contained in the random forest prediction model for farm children are related to the genes IL33 (rs928413) and RAD50 (rs6871536), thereby representing two major asthma risk loci.[25] IL33 has been implied in allergies and autoimmune disorders, and a role in exuberant immune responses related to reduced numbers of regulatory T cells is discussed.[26] The other SNP is situated in an intron of RAD50 in the TH2 cytokine locus on chromosome 5 and has been reported to be associated with asthma, atopic eczema and total IgE levels.[25,27,28]

Though marginally missing statistical significance, two further candidate SNPs (rs9815663, rs6967330) may also be of interest as they are related to CDHR3 and IL5RA. Like other members of the cadherin family of transmembrane proteins, CDHR3 is associated with asthma-related traits and a function in epithelial polarity, cell-cell interaction and differentiation has thus been suggested.[25] The alpha chain of the IL5 receptor is essential for differentiation and maturation of eosinophils, and inactivation of IL5 reduces airway eosinophilia.[29] Taken together, the detected genes are mainly related to the allergic aspects of asthma; allergic asthma, in turn, is related specifically to lung function impairment and need for inhaled corticosteroids.[30] In contrast, the SNPs of the asthma risk locus on chromosome 17q21 did not relevantly contribute to the prediction models in farm children. This locus has been suggested to encode susceptibility to environmental signals,[31] which might not be relevant for the prediction of the sort of asthma farm children suffer from.

Six environmental variables, which all related to pet exposure in childhood, were affected by more than 6% missing values and were imputed. Although multiple imputation is designed to reduce systematic imputation bias, the corresponding results should be interpreted with caution. The most relevant predictor among pet exposure was contact to a cat during the last year, which had only 6% missing values. Consequently, relevant contribution to asthma prediction by pet exposure seems possible.

Essentially, asthma is an umbrella term for various disease entities manifesting with similar symptoms.[30,32] Children whose parents are not aware of an asthma diagnosis might be classified as controls even if they are treated with asthma drugs or experience current asthma symptoms. When excluding these children from the reference group (Figure 4), the prediction performed significantly better thereby implying true cases of asthma covered by this grey zone. The performance of the prediction also improved when asthma was defined irrespectively of recurrent diagnoses of obstructive bronchitis, which may include less severe asthma forms.[30]

Technically, we have exploited instruments of predicting childhood asthma with modern machine learning methods. Consistently, the highest prediction quality was achieved by random forest. In contrast to regression models, it is based on decision trees and can efficiently handle high-dimensional data. Random forest is unaffected by highly correlated variables and thus inherently robust. During the tree building process, random forest essentially stratifies for variables, thereby automatically considering interactions between predictor variables. One interpretation of the relatively good prediction quality in farm children might be found in gene-gene interactions, which are disregarded by all other methods. Taken together, we have applied computationally efficient, stable and robust methods, which run a low risk of model overfitting and can handle a high number of variables simultaneously and hence more appropriately. These properties render them ideal tools for prediction though they may be computationally demanding and require a powerful computing infrastructure.

In conclusion, asthma in farm children seems to be distinct from asthma in non-farm children, at least with respect to genetic and environmental predictors. The common denominator is family history, which may integrate genotype and degree of penetrance conditional on the environmental setting. Retrospectively, the potential of genome-wide data for the prediction of polygenic diseases might have been overrated, whereas the power of the environment merits a second look.

ACKNOWLEDGMENTS

Open access funding enabled and organized by ProjektDEAL.

CONFLICT OF INTEREST

Dr. Kabesch reports grants from European Commission during the conduct of the study. Dr. von Mutius reports grants from European

Commission during the conduct of the study; personal fees from Pharma Ventures, personal fees from Peptinnovate Ltd., personal fees from OM Pharma SA, personal fees from Boehringer Ingelheim, personal fees from HAL Allergie GmbH, personal fees from Ökosoziales Forum Oberösterreich, personal fees from Mundipharma Deutschland GmbH & Co. KG, personal fees from European Commission/European Research Council Executive Agency, personal fees from The Chinese University of Hong Kong, personal fees from University of Tampere/Tampereen Yliopisto, personal fees from University of Utrecht/Utrecht Universiteit, personal fees from University of Turku/Turun Yliopisto, personal fees from University of Helsinki/Helsingin yliopisto and outside the submitted work.

AUTHOR CONTRIBUTIONS

Norbert Krautenbacher: Conceptualization (equal); Data curation (equal); Formal analysis (lead); Methodology (equal); Software (lead); Validation (lead); Visualization (lead); Writing-original draft (supporting); Writing-review & editing (equal). **Michael Kabesch:** Investigation (equal); Methodology (equal); Resources (equal); Writing-review & editing (equal). **Elisabeth Horak:** Conceptualization (equal); Investigation (equal); Project administration (equal); Resources (equal); Writing-review & editing (equal). **Charlotte Braun-Fahrländer:** Conceptualization (equal); Funding acquisition (equal); Investigation (equal); Methodology (equal); Project administration (equal); Writing-review & editing (equal). **Jon Genuneit:** Conceptualization (equal); Data curation (equal); Methodology (equal); Project administration (equal); Resources (equal); Writing-review & editing (equal). **Andrzej Boznański:** Funding acquisition (equal); Investigation (equal); Methodology (equal); Project administration (equal); Resources (equal); Writing-review & editing (equal). **Erika von Mutius:** Conceptualization (equal); Funding acquisition (lead); Investigation (equal); Methodology (equal); Project administration (lead); Resources (equal); Writing-review & editing (equal). **Fabian Theis:** Methodology (equal); Resources (equal); Supervision (equal); Writing-review & editing (equal). **Christiane Fuchs:** Conceptualization (lead); Formal analysis (lead); Funding acquisition (equal); Investigation (equal); Methodology (lead); Software (lead); Supervision (lead); Validation (lead); Visualization (equal); Writing-original draft (equal); Writing-review & editing (equal). **Markus Ege:** Conceptualization (lead); Formal analysis (equal); Investigation (equal); Supervision (lead); Validation (equal); Writing-original draft (lead); Writing-review & editing (lead).

ORCID

Michael Kabesch https://orcid.org/0000-0003-0697-1871
Jon Genuneit https://orcid.org/0000-0001-5764-1528
Markus J. Ege https://orcid.org/0000-0001-6643-3923

REFERENCES

1. Moffatt MF, Gut IG, Demenais F, et al. A large-scale, consortium-based genomewide association study of asthma. *N Engl J Med*. 2010;363(13):1211-1221.
2. Duffy DL, Martin NG, Battistutta D, Hopper JL, Mathews JD. Genetics of asthma and hay fever in Australian twins. *Am Rev Respir Dis*. 1990;142(6 Pt 1):1351-1358.
3. Loss GJ, Depner M, Hose AJ, et al. The early development of wheeze. Environmental determinants and genetic susceptibility at 17q21. *Am J Respir Crit Care Med*. 2016;193(8):889-897.
4. Dudbridge F, Gusnanto A. Estimation of significance thresholds for genomewide association scans. *Genet Epidemiol*. 2008;32(3):227-234.
5. Ege MJ, Strachan DP, Cookson WO, et al. Gene-environment interaction for childhood asthma and exposure to farming in Central Europe. *J Allergy Clin Immunol*. 2011;127(1):138-144, 144 e131–134.
6. Genuneit J, Buchele G, Waser M, et al. The GABRIEL advanced surveys: study design, participation and evaluation of bias. *Paediatr Perin Epidemiol*. 2011;25(5):436-447.
7. Weiland SK, von Mutius E, Hirsch T, et al. Prevalence of respiratory and atopic disorders among children in the East and West of Germany five years after unification. *Eur Respir J*. 1999;14(4):862-870.
8. Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet Epidemiol*. 2010;34(8):816-834.
9. Genomes Project C, Abecasis GR, Auton A, et al. An integrated map of genetic variation from 1,092 human genomes. *Nature*. 2012;491(7422):56-65.
10. Wu J, Pfeiffer RM, Gail MH. Strategies for developing prediction models from genome-wide association studies. *Genet Epidemiol*. 2013;37(8):768-777.
11. MacArthur J, Bowler E, Cerezo M, et al. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res*. 2017;45(D1):D896-D901.
12. R Core Team. *R: A Language and Environment for Statistical Computing [Computer Program]*. Vienna, Austria: R Foundation for Statistical Computing; 2016.
13. Buuren SV, Groothuis-Oudshoorn K. mice: multivariate imputation by chained equations inR. *J Stat Softw*. 2011;45(3):1–67.
14. Hastie T, Tibshirani R, Friedman J. *Overview of Supervised Learning. The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York, NY: Springer; 2009:9-41.
15. Boulesteix AL, De Bin R, Jiang X, Fuchs M. IPF-LASSO: integrative L1-penalized regression with penalty factors for prediction based on multi-omics data. *Comput Math Methods Med*. 2017;2017:7691937.
16. Breiman L. Random forests. *Mach Learn*. 2001;45(1):5-32.
17. Fawcett T. An introduction to ROC analysis. *Pattern Recognit Lett*. 2006;27(8):861-874.
18. Ober C. Asthma genetics in the post-GWAS Era. *Ann Am Thor Soc*. 2016;13(Suppl 1):S85-S90.
19. Pasaniuc B, Price AL. Dissecting the genetics of complex traits using summary association statistics. *Nat Rev Genet*. 2017;18(2):117-127.
20. Boyle EA, Li YI, Pritchard JK. An expanded view of complex traits: from polygenic to omnigenic. *Cell*. 2017;169(7):1177-1186.
21. Spycher BD, Henderson J, Granell R, et al. Genome-wide prediction of childhood asthma and related phenotypes in a longitudinal birth cohort. *J Allergy Clin Immunol*. 2012;130(2):503-509 e507.
22. Torous W, Valkanov R. *Boundaries of Predictability: Noisy Predictive Regressions*. Los Angeles, CA: Anderson Graduate School of Management, UCLA; 2000.
23. Backhed F, Roswall J, Peng Y, et al. Dynamics and stabilization of the human gut microbiome during the first year of life. *Cell Host Microbe*. 2015;17(5):690–703.
24. Ege MJ. Asthma and prenatal inflammation. *Am J Respir Crit Care Med*. 2017;195(5):546-548.
25. Bonnelykke K, Sleiman P, Nielsen K, et al. A genome-wide association study identifies CDHR3 as a susceptibility locus for early childhood asthma with severe exacerbations. *Nat Genet*. 2014;46(1):51-55.

26. Schroder PC, Casaca VI, Illi S, et al. IL-33 polymorphisms are associated with increased risk of hay fever and reduced regulatory T cells in a birth cohort. *Pediatr Allergy Immunol*. 2016;27(7):687-695.
27. Li X, Howard TD, Zheng SL, et al. Genome-wide association study of asthma identifies RAD50-IL13 and HLA-DR/DQ regions. *J Allergy Clin Immunol*. 2010;125(2):328-335 e311.
28. Weidinger S, Willis-Owen SA, Kamatani Y, et al. A genome-wide association study of atopic dermatitis identifies loci with overlapping effects on asthma and psoriasis. *Hum Mole Genet*. 2013;22(23):4841-4856.
29. Sehmi R, Smith SG, Kjarsgaard M, et al. Role of local eosinophilopoietic processes in the development of airway eosinophilia in prednisone-dependent severe asthma. *Clin Exp Allergy*. 2016;46(6):793-802.
30. Depner M, Fuchs O, Genuneit J, et al. Clinical and epidemiologic phenotypes of childhood asthma. *Am J Respir Crit Care Med*. 2014;189(2):129-138.
31. Caliskan M, Bochkov YA, Kreiner-Moller E, et al. Rhinovirus wheezing illness and genetic risk of childhood-onset asthma. *N Engl J Med*. 2013;368(15):1398-1407.
32. Eder W, Ege MJ, von Mutius E. The asthma epidemic. *N Engl J Med*. 2006;355(21):2226-2235.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Krautenbacher N, Kabesch M, Horak E, et al; the GABRIELA, PASTURE study groups. Asthma in farm children is more determined by genetic polymorphisms and in non-farm children by environmental factors. *Pediatr Allergy Immunol*. 2020;00:1–10. https://doi.org/10.1111/pai.13385