

PROFESSOR MICHAEL KABESCH (Orcid ID : 0000-0003-0697-1871)

DR JON GENUNEIT (Orcid ID : 0000-0001-5764-1528)

PROFESSOR MARKUS EGE (Orcid ID : 0000-0001-6643-3923)

Article type : Original

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

Norbert Krautenbacher, MSc,<sup>1,2</sup> Michael Kabesch, MD,<sup>3,4,5</sup> Elisabeth Horak, MD,<sup>6</sup> Charlotte Braun-Fahrlander, MD,<sup>7,8</sup> Jon Genuneit, MD,<sup>9,10</sup> Andrzej Boznanski,<sup>11</sup> Erika von Mutius, MD,<sup>5,12,13</sup> Fabian Theis, PhD,<sup>1,2</sup> Christiane Fuchs, PhD,<sup>1,2,14\*</sup> Markus J. Ege, MD,<sup>5,12\*</sup> and the GABRIELA and PASTURE study groups

<sup>1</sup> Institute of Computational Biology, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany

<sup>2</sup> Technische Universität München, Center for Mathematics, Chair of Mathematical Modeling of Biological Systems, Garching, Germany

<sup>3</sup> University Children's Hospital Regensburg (KUNO), Regensburg, Germany

<sup>4</sup> Hannover Medical School, Clinic for Pediatric Pneumology and Neonatology, Hannover, Germany

<sup>5</sup> The German Center for Lung Research (DZL), Germany

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the [Version of Record](#). Please cite this article as [doi: 10.1111/PAI.13385](https://doi.org/10.1111/PAI.13385)

This article is protected by copyright. All rights reserved

<sup>6</sup> Innsbruck Medical University, Department of Pediatrics and Adolescents, Innsbruck, Austria

<sup>7</sup> Swiss Tropical and Public Health Institute Basel, Basel, Switzerland

<sup>8</sup> University of Basel, Basel, Switzerland

<sup>9</sup> Institute of Epidemiology and Medical Biometry, Ulm University, Ulm, Germany

<sup>10</sup> Pediatric Epidemiology, Department of Pediatrics, Medical Faculty, Leipzig University, Leipzig, Germany

<sup>11</sup> Wroclaw Medical University, Wroclaw, Poland

<sup>12</sup> Dr von Hauner Children's Hospital, LMU Munich, Munich, Germany

<sup>13</sup> Institute of Asthma and Allergy Prevention, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany

<sup>14</sup> Department of Business Administration and Economics, Bielefeld University, Bielefeld, Germany

\* equal contribution

***Address for correspondence***

Markus Ege, MD MPH

Dr. von Hauner Children's Hospital

Ludwig Maximilians University Munich

Lindwurm Str. 4

80337 Munich

Germany

Email: markus.ege@med.lmu.de

Word count: 2494

Number of tables: 2

Number of figures: 4

Electronic repository: Extended methods, 3 extended tables, 3 extended figures

## **Conflict of interest statement**

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 Ökosoiales 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, outside the submitted work.

## **Financial support**

The GABRIELA study was supported by the European Commission as part of GABRIEL (a multidisciplinary study to identify the genetic and environmental causes of asthma in the European Community) contract number 018996 under the Integrated Program LSH-2004-1.2.5-1.

The Protection Against Allergy—Study in Rural Environments (PASTURE) study was supported by the European Commission research grants QLK4-CT-2001-00250, FOOD-CT-2006-31708, and KBBE-2007-2-2-06.

The current statistical analysis was supported by the German Research Foundation within SFB 1243, Subproject A17 (CF, FJT) and by the Federal Ministry of Education and Research, Grant Number 01DH17024 (CF) and 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. 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.

### **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.

### **Key words**

Childhood asthma, farming, risk prediction, GWAS, environment, SNPs, statistical learning, machine learning, random forest, penalized regression

## 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.<sup>1</sup> This modest result conflicted with the strong hereditary background postulated from twin studies.<sup>2</sup> 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.<sup>3</sup> 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 to 0.7 million independent loci in the human genome separately.<sup>4</sup> 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 state-of-the-art statistical tools which consider predictor variables integratively and provide high statistical power.

## Methods

### *Population and questionnaires*

The cross-sectional GABRIEL Advanced Studies (GABRIELA) was designed to study gene-environment interactions.<sup>5</sup> From the Austrian, Swiss, and German arms of GABRIELA, 1,707 schoolchildren were randomly selected from all 34,491 children eligible for genotyping in a stratified design (Figure E1).<sup>5,6</sup> 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.<sup>7</sup> The questionnaires contained items on individual and family health, socioeconomic 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=1,224) 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.

### *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.<sup>1</sup> SNPs were imputed by Markov chain-based haplotyper<sup>8</sup> using the 1000 Genomes pilot 1 release.<sup>9</sup> SNPs were filtered for imputation quality ( $Rsq \geq 0.30$ ) and minor allele frequency ( $MAF \geq 0.05$ ) 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.<sup>10</sup>

Candidate SNPs (Table 1) were defined as SNPs included in the GWAS catalog for childhood onset asthma.<sup>11</sup>

### *Computational and Statistical Analysis*

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

Environmental variables (Supplemental Table 1) had less than 25% missing values. Missing values of variables were imputed by multiple imputation resulting in five imputation data sets,<sup>13</sup> 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,<sup>1,10</sup> 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).<sup>14,15</sup> Additionally random forests were built on 20,000 trees.<sup>16</sup>

Model selection and 5-fold cross-validation was performed on the 1,410 Swiss and German participants. The best models were then externally validated in the 297 Austrian participants (the smallest center) and additionally in PASTURE.

As a metric for model comparison, we applied the area under the receiver operating characteristics (ROC) curve (AUC) with a bootstrapped 95%-confidence interval.<sup>17</sup> 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.

## 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 E1).

When predicting asthma by groups of variables separately, i.e. 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).

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 E2) 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, center 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 E2) with an AUC of 0.86 [0.59-0.99] averaged over the prediction scores of random forest and IPF-LASSO (Figure E3). 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 E3) 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).

## 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 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.<sup>18,19</sup> 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.<sup>1</sup> 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, i.e. 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.<sup>19,20</sup>

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 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.<sup>10</sup> Prediction quality for asthma has already been shown to decrease beyond 100 SNPs.<sup>21</sup>

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.<sup>19,20</sup> On the other hand, 99.5% of the genome-wide SNPs did not significantly contribute to the prediction model and may have increased noise.<sup>22</sup> 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 is 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.<sup>23</sup> Likewise, a family history may represent conditions during pregnancy, e.g. epigenetic mechanisms or an inflammatory status of the mother shaping the fetal immune system and thus contributing to disease transmission.<sup>24</sup> 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 GWAS analyses.

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.<sup>25</sup> 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.<sup>26</sup> 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.<sup>25,27,28</sup>

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.<sup>25</sup> The alpha chain of the IL5 receptor is essential for differentiation and maturation of eosinophils, and inactivation of IL5 reduces airway eosinophilia.<sup>29</sup> 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.<sup>30</sup> 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,<sup>31</sup> 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.<sup>30,32</sup> 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.<sup>30</sup>

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**

Authors' contributions: Design and implementation of the GABRIELA study: EH, CBF, JG, AB, and EvM. Genetic analyses within GABRIELA: MK. Statistical analyses: NK. Supervision of statistical analyses: CF, FT, and MJE. Drafting the manuscript for important intellectual content: NK, CF, and MJE.

## 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. *Genetic epidemiology*. 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. *Paediatric and perinatal epidemiology*. 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. *Genetic epidemiology*. 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. *Genetic epidemiology*. 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: 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 in R. *Journal of Statistical Software*. 2011;45(3).
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 New York; 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. *Machine Learning*. 2001;45(1):5-32.
17. Fawcett T. An introduction to ROC analysis. *Pattern Recognition Letters*. 2006;27(8):861-874.
18. Ober C. Asthma Genetics in the Post-GWAS Era. *Annals of the American Thoracic Society*. 2016;13 Suppl 1:S85-90.
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*. Anderson Graduate School of Management, UCLA; Dec 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. *Nature genetics*. 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. *Human molecular genetics*. 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.

## Tables

**Table 1: SNPs reported for childhood asthma in the GWAS catalog**

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

\* p-values for associations with childhood onset asthma are taken from the GWAS catalog.<sup>11</sup>

\*\* Genes are reported by authors of original publications.<sup>11</sup> If no genes are reported, mapped genes are given in the comments column.

**Table 2: Potential determinants of asthma**

| Characteristic                                   | Cases (%)<br>n=850 | Controls (%)<br>n=857 | p-value* |
|--------------------------------------------------|--------------------|-----------------------|----------|
| female sex                                       | 39.70%             | 49.40%                | 0.002    |
| age <sup>§</sup>                                 | 8.32 (0.06)        | 8.19 (0.06)           | 0.150    |
| body mass index <sup>§</sup>                     | 17.11 (0.11)       | 16.99 (0.11)          | 0.375    |
| family history of atopy                          | 70.00%             | 49.40%                | < 0.001  |
| family history of asthma                         | 30.06%             | 12.40%                | < 0.001  |
| living on a farm                                 | 9.00%              | 13.60%                | < 0.001  |
| at least two siblings                            | 0.42 (0.02)        | 0.45 (0.02)           | 0.374    |
| high parental education                          | 27.30%             | 28.80%                | 0.633    |
| maternal smoking during pregnancy                | 12.40%             | 8.50%                 | 0.037    |
| consumption of farm milk during past 12 months   | 13.40%             | 19.40%                | < 0.001  |
| consumption of farm milk in first year of life   | 6.20%              | 11.80%                | < 0.001  |
| consumption of farm milk (pregnancy to age 3yrs) | 20.70%             | 27.60%                | < 0.001  |
| contact with cows (past 12 months)               | 12.90%             | 16.60%                | 0.020    |
| contact with cows (pregnancy to age 3yrs)        | 14.60%             | 20.30%                | < 0.001  |
| contact with straw (past 12 months)              | 15.70%             | 21.10%                | 0.009    |
| contact with straw (pregnancy to age 3yrs)       | 12.40%             | 16.20%                | 0.009    |
| contact with hay (past 12 months)                | 29.70%             | 33.50%                | 0.145    |

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

<sup>§</sup> mean and standard error of mean

Only variables are shown that appeared as most relevant in subsequent analyses. A complete list of environmental variables is provided in Table E1 in the Online Repository.

## Figure legends

### **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.

### **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.

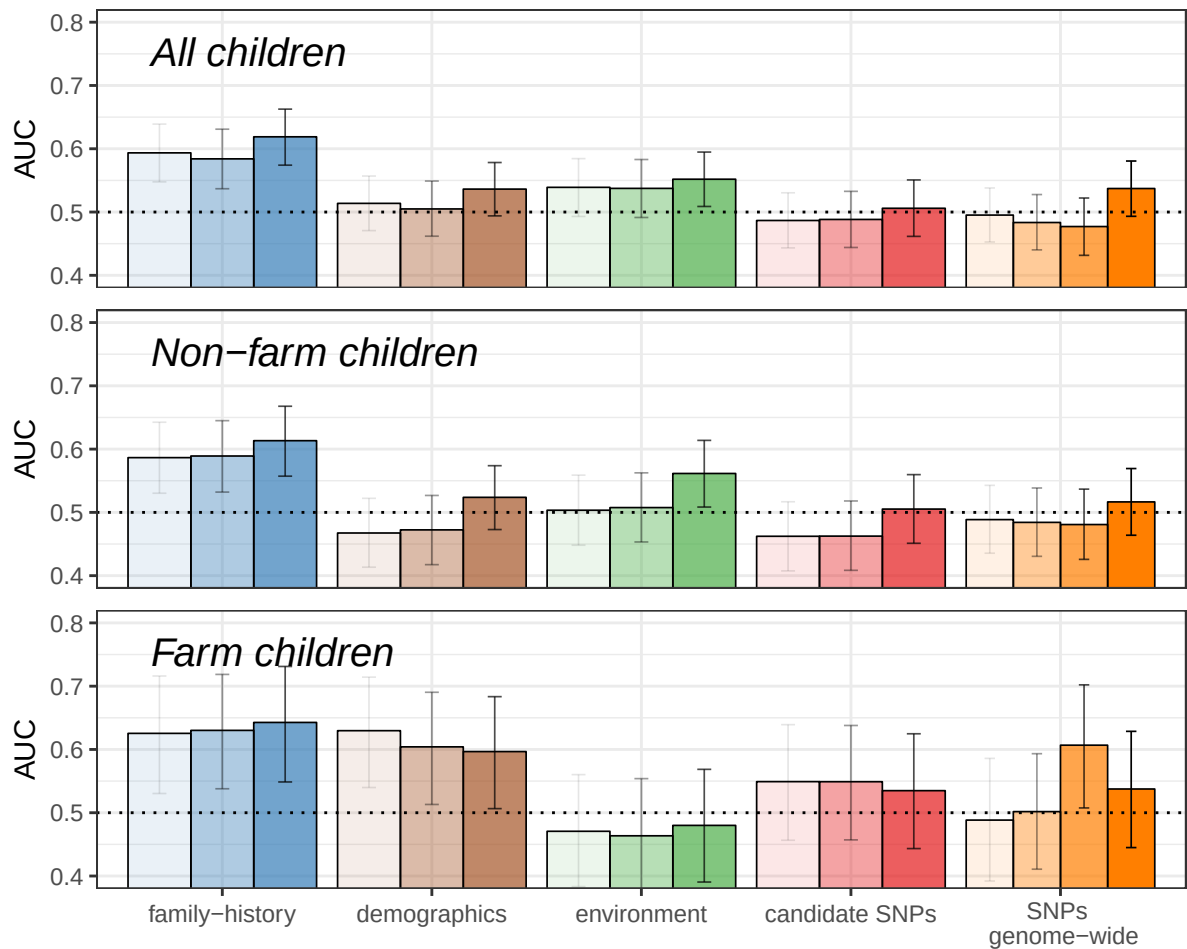
### Figure 3: Importance of variables contributing to the best prediction models for childhood asthma

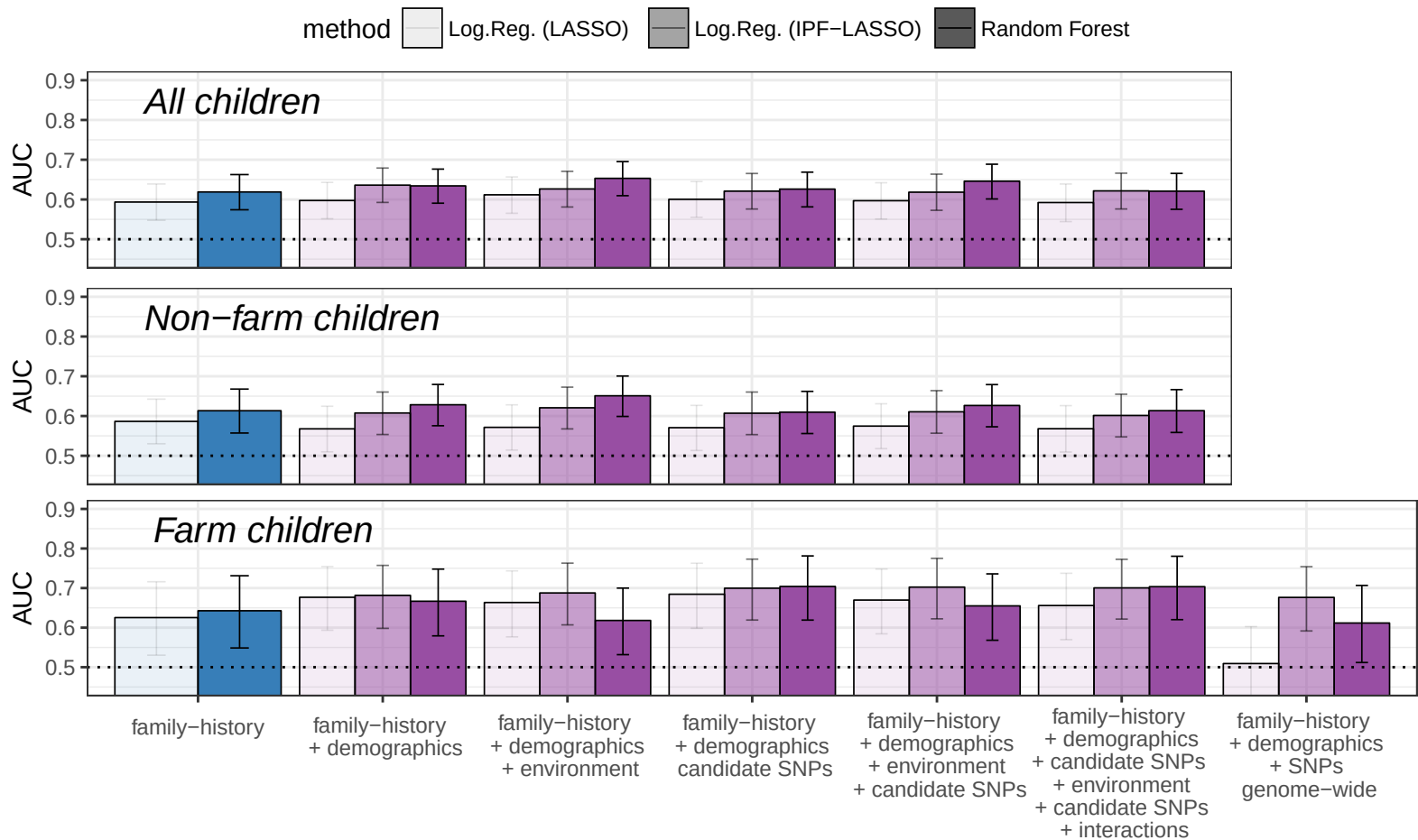
Variable importance determined by random forest models for all children and stratified for farm and non-farm children.

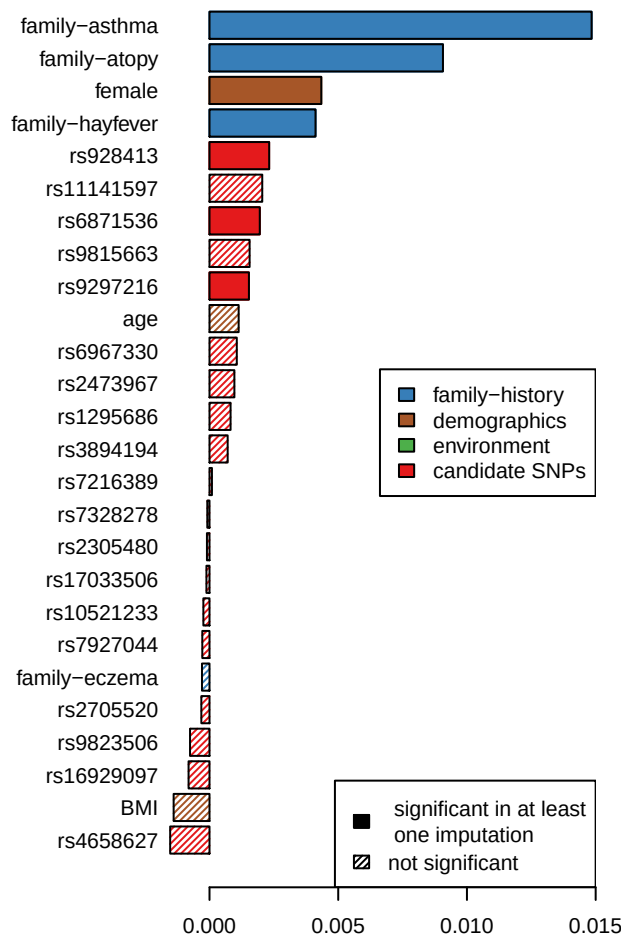
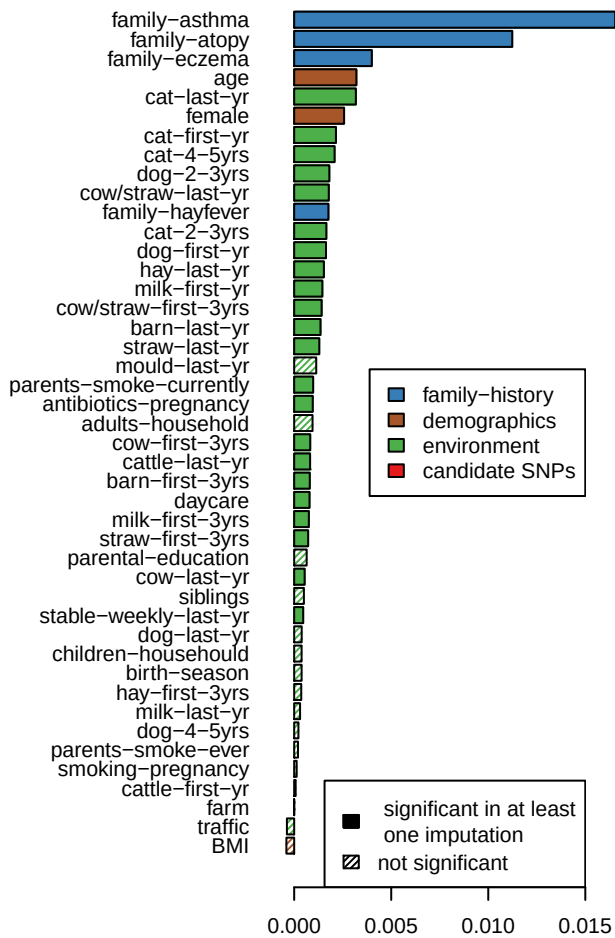
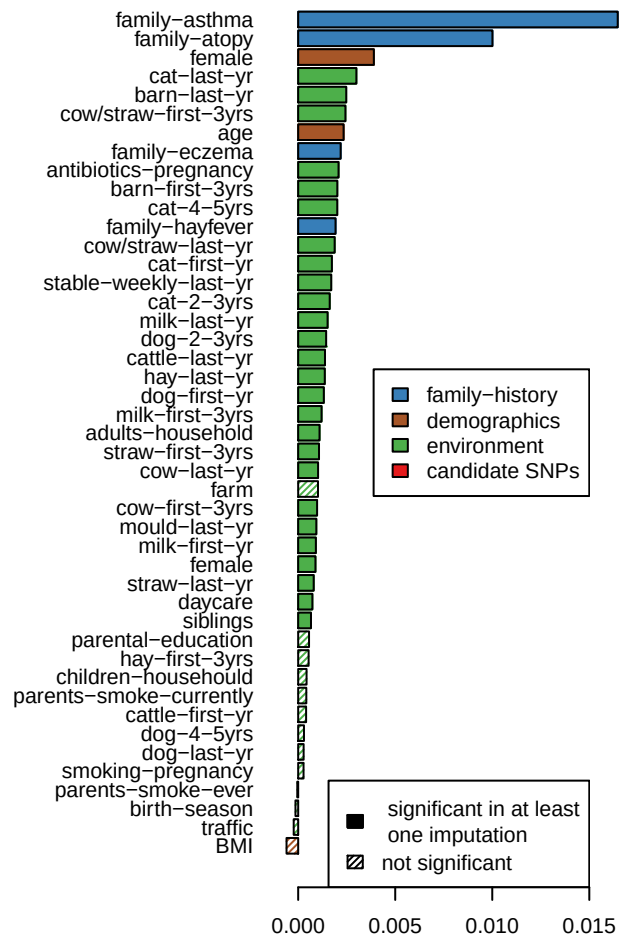
### Figure 4: External validation of prediction models for childhood asthma

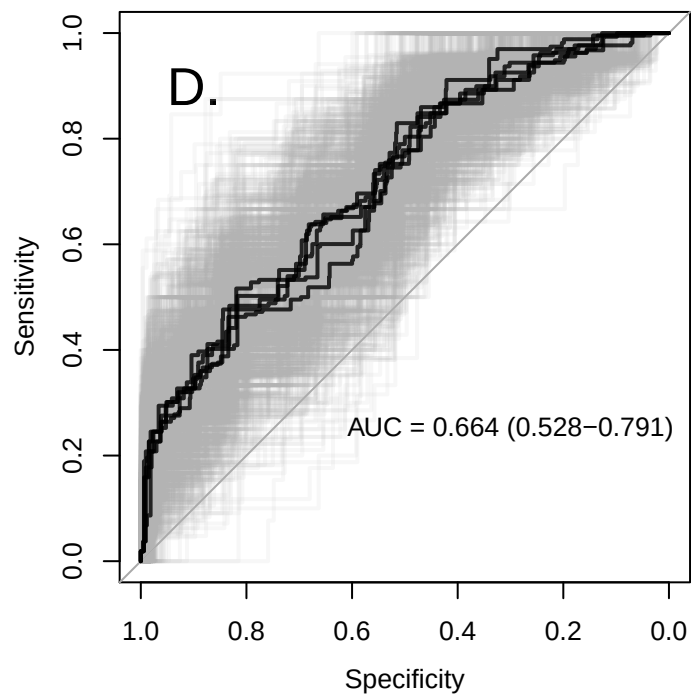
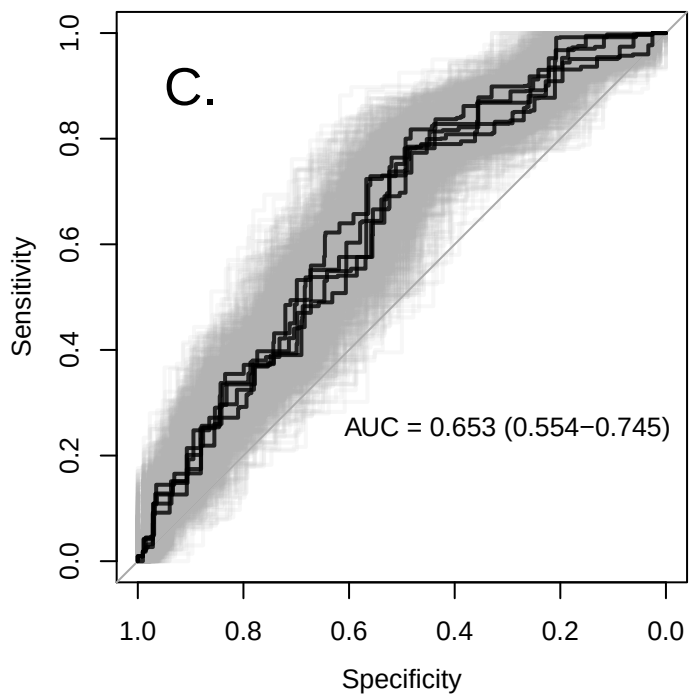
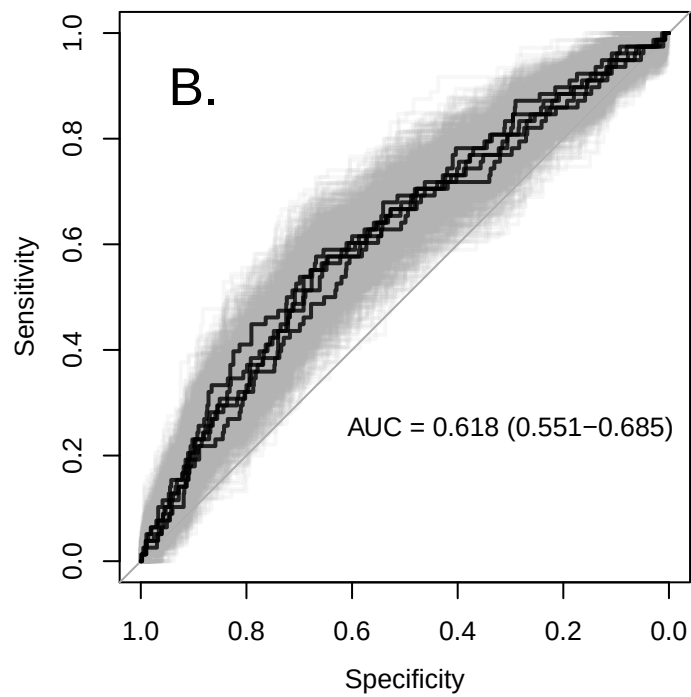
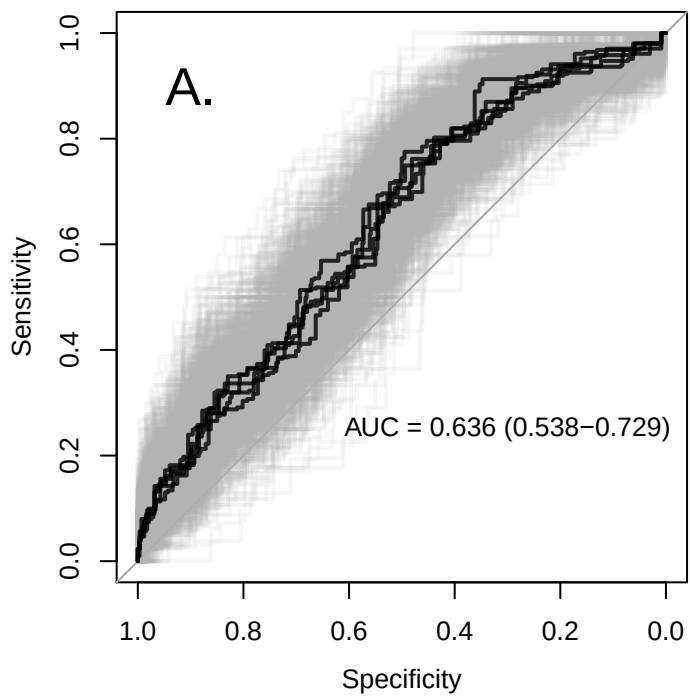
For external validation, individual ROC curves of the 5 imputations with averaged area under the curve (AUC) are shown. The internally validated random forest prediction model of a parent-reported doctor-diagnosis of asthma once or obstructive bronchitis twice based on *family-history, demographics, 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.

method    Log.Reg. (LASSO)    Log.Reg. (Elastic Net)    Random Forest    GWAS-Log.Reg.









## Online supplement

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

Norbert Krautenbacher, MSc,<sup>1,2</sup> Michael Kabesch,<sup>3,4,5</sup> Elisabeth Horak, MD,<sup>6</sup> Charlotte Braun-Fahrlander, MD,<sup>7,8</sup> Jon Genuneit, MD,<sup>9,10</sup> Andrzej Boznanski,<sup>11</sup> Erika von Mutius, MD,<sup>5,12,13</sup> Fabian Theis, PhD,<sup>1,2</sup> Christiane Fuchs, PhD,<sup>1,2,14\*</sup> Markus J. Ege,<sup>5,12\*\*</sup> and the GABRIELA and PASTURE study groups

<sup>1</sup> Institute of Computational Biology, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany

<sup>2</sup> Technische Universität München, Center for Mathematics, Chair of Mathematical Modeling of Biological Systems, Garching, Germany

<sup>3</sup> University Children's Hospital Regensburg (KUNO), Regensburg, Germany

<sup>4</sup> Hannover Medical School, Clinic for Paediatric Pneumology and Neonatology, Hannover, Germany

<sup>5</sup> The German Center for Lung Research (DZL), Germany

<sup>6</sup> Innsbruck Medical University, Innsbruck, Austria

<sup>7</sup> Swiss Tropical and Public Health Institute Basel, Basel, Switzerland

<sup>8</sup> University of Basel, Basel, Switzerland

<sup>9</sup> Institute of Epidemiology and Medical Biometry, Ulm University, Ulm, Germany

<sup>10</sup> Pediatric Epidemiology, Department of Pediatrics, Medical Faculty, Leipzig University, Leipzig, Germany

25 <sup>11</sup> Wroclaw Medical University, Wroclaw, Poland  
26 <sup>12</sup> Dr von Hauner Children's Hospital, LMU Munich, Munich, Germany  
27 <sup>13</sup> Institute of Asthma and Allergy Prevention, Helmholtz Zentrum München, German Research  
28 Center for Environmental Health, Neuherberg, Germany  
29 <sup>14</sup> Faculty of Business Administration and Economics, Bielefeld University, Bielefeld, Germany  
30 \* equal contribution  
31  
32  
33 Address for correspondence:  
34 Markus Ege, MD MPH  
35 Dr. von Hauner Children's Hospital  
36 Ludwig Maximilians University Munich  
37 Lindwurm Str. 4  
38 80337 Munich  
39 Germany  
40 Email: [markus.ege@med.lmu.de](mailto:markus.ege@med.lmu.de)

## Methods:

### Computational and Statistical Analysis

#### Multiple Imputation

The environmental variables contained missing values up to 6 % per variable except for exposure to dog and cats at age 4/5 years (15%), at age 2/3 years (20%), and during their first year (24%). In order to retain variability, multiple imputation with the R package *MICE* was used to impute the missing values by chain equations [1]. Continuous variables were imputed by predictive mean matching (number of closest values selected from:  $k=5$ ), binary variables by logistic regression, categorical variables by a multinomial logit model. We used five imputations and kept the required multiple imputation steps through all our analyses according to Rubin [2].

#### Integration of several modalities

We classified the potential determinants of asthma in five categories, which we termed modalities. The modality environmental determinants comprised 38 environmental variables; family history (FH) of asthma or atopy contained 4 variables; demographics included sex, age, and BMI; candidate SNPs included 19 SNPs and genome-wide SNPs covered 744,908 GWAS SNPs. Each prediction model was applied to these modalities separately. Furthermore the modalities were entered subsequently in the model until all modalities were included. Finally, we explored interaction terms between the 19 candidate SNPs and 7 environmental variables, i.e. number of siblings, exposure to farm environment, interaction between sex and farm environment, FH of atopy, FH of asthma, FH of hay fever, and FH of eczema.

#### Response

Apart from the main response variable *doctor-diagnosed asthma*, we investigated the asthma phenotype in a more unambiguous way. To that end, we left out subjects who reported current

wheeze or used asthma sprays but were not diagnosed with asthma (*sensitivity analysis*). By this definition, 166 of the 1707 subjects dropped out.

In a further investigation, we defined children to have asthma irrespectively of bronchitis. Children without doctor-diagnosed asthma were treated as controls. By this definition 4 of the 1707 subjects dropped out due to missing data on asthma diagnosis.

### **Stratified sample selection**

For economic reasons, the GWAS sample of the GABRIEL study, i.e. the current analysis population, was enriched for informative observations. This was achieved by oversampling of children with farm exposure and simultaneous oversampling of asthma cases by a stratified random selection. Adequate statistical analyses were applied to avoid sampling bias [3]. Essentially, these techniques weight the selected sample ( $n=1707$ ) back to the original unbiased study population ( $n=34,4391$ ) within the predefined strata of farm exposure and asthma and atopy definitions. As weight the inverse of the selection probability within the stratum is used. The following sections also explain how this correction was applied when establishing and validating prediction models.

### **Statistical learning models**

In order to predict the probability of a binary response such as asthma, logistic regression is a standard method for building a prediction model and thus a risk score [4]. Let  $n$  be the sample size,  $p$  the number of variables in the data. Then let  $y_i$  denote the response, i.e. the disease status ( $y_i = 0$  for no disease or 1 for disease), for subject  $i$ . Further let  $x$  contain the variables used for prediction (i.e.  $x$  is a matrix with  $n$  rows and  $p$  columns),  $x_i$  be the variable values for subject  $i$  and  $x_{ij}$  be the value for subject  $i$  and variable  $j$ .

Then in the multivariable logistic regression model,  $y_i|x_i$  is assumed to be Bernoulli distributed with success probability  $P(y_i = 1|x_i) = (1 + \exp(-(\beta_0 + \sum_{j=1}^p x_{ij}\beta_j)))^{-1}$  where  $\beta_0$  and  $\beta = (\beta_1, \dots, \beta_p)'$  are unknown parameters, which represent effects of the variables  $x_i$  on  $y_i$ .

During maximum likelihood estimation, the log-likelihood  $l(\beta_0, \beta) = \sum_i^n (y_i(\beta_0 + x_i'\beta) - \log(1 + \exp(\beta_0 + x_i'\beta)))$  with respect to  $\beta_0$  and  $\beta$  is maximized.

To correct the point estimates for the effects for sample selection bias, the observations are weighted by inverse-probability weights  $w_i$  [3] which are given after the stratified random selection process [5]. Then the modified log-likelihood is given by

$$l_{mod}(\beta_0, \beta) = \sum_{i=1}^n w_i (y_i(\beta_0 + x_i'\beta) - \log(1 + \exp(\beta_0 + x_i'\beta))).$$

Since this is a modification of maximum likelihood estimation in the common sense, this function is called pseudo-maximum-likelihood [5].

The use of multivariable logistic regression is suitable for classification problems, however, can be improved in prediction by using regularization [6]. Further, if  $n < p$ , logistic regression without regularization is analytically not feasible, which is the case for our analyses using all SNPs.

In regularized logistic regression,  $\hat{\beta} = \arg \min_{\beta} \{ \sum_{i=1}^n -l(\beta_0, \beta) + \lambda \cdot pen(\beta) \}$  with  $\lambda$  as tuning parameter and  $pen$  a penalization function. The latter can be chosen to be based on the  $L_1$ -norm, i.e.  $\sum_{j=1}^p |\beta_j|$  (LASSO-penalty), or the  $L_2$ -norm, i.e.  $\sum_{j=1}^p \beta_j^2$  (ridge-penalty), or include both  $\sum_{j=1}^p (\alpha \beta_j^2 + (1 - \alpha)|\beta_j|)$  (elastic-net penalty). Our goal was to select important variables from many candidate variables. Thus, we used those penalty terms that penalize non-influential coefficients to zero. These were the LASSO and the elastic net penalty. For the elastic net we chose  $\alpha=0.5$  in order to leave important coefficients non-zero when variables have moderate effects but still allow for feature selection so that non-important variables do not enter the final model as in the LASSO case. In both penalized regression model versions, the LASSO and the

elastic net, the optimal  $\lambda$  was obtained via 10-fold cross-validation. The multiple  $\lambda_m$  for IPF-LASSO was obtained via (5 times repeated) 5-fold cross-validation.

A further issue in the data was the situation of several modalities. By *modalities* we refer to blocks of variables which biologically belong together, in our case variables containing information about the environment, demographics, family history or genetics. If these differ substantially in their dimensions, practical problems can occur, i.e. variables of low-dimensional modalities can get put at a disadvantage [7]. Therefore we implemented a modified version of the LASSO, namely the IPF-LASSO (integrative L1-penalized regression with penalty factors), assigning additionally different penalty factors to different data modalities [7]. Their coefficients are obtained by minimizing

$$\sum_{i=1}^n \left( y_i \left( \beta_0 + \sum_{m=1}^M \sum_{j=1}^{p_m} x_{ij}^{(m)} \beta_j^{(m)} \right) - \log \left( 1 + \exp \left( \beta_0 + \sum_{m=1}^M \sum_{j=1}^{p_m} x_{ij}^{(m)} \beta_j^{(m)} \right) \right) \right) + \sum_{m=1}^M \lambda_m |\beta^m|$$

with  $\lambda_m > 0$  the penalty applied to the variables from modality  $m$ .

Since all presented regression methods are likelihood-based, inverse-probability weighting could simply be applied in order to correct for sample selection bias.

As a further statistical learning approach we used the random forest [8]. The method is based on combining many classification trees to forests. Classification trees themselves are built by partitioning the variable space into regions. Random forests are a powerful tool in machine learning and often outperform classical statistical methods in terms of prediction accuracy. They can be affected by the so-called imbalanced data classification problem [9]. In general, state-of-the-art correction methods for sample selection bias would cause such imbalanced data in the present study. Indeed, most correction methods decrease performance of the random forest if

the analysis population is selected by stratified random sampling [3]. Consequently, correcting for sample selection bias was skipped in the learning process of the random forest approach. However, it was applied in the validation step.

#### **Integrating genetic data**

We compared the above mentioned approaches to a state-of-the-art strategy of selecting SNPs univariately for prediction models from a genome-wide pool of SNPs. Following Wu et al. [10], the best prediction strategy in terms of AUC values first estimates univariate associations by simple logistic regression models for each SNP  $j$ , i.e.  $P(y_i = 1|x_{ij}) = (1 + \exp(-(\beta_0 + \beta_j x_{ij})))^{-1}$ . In a second step, a score  $S_i$  for subject  $i$  is calculated as the sum of the univariately estimated coefficients of the 100 SNPs with the lowest p-values:

$$S_i = \sum_{j=1}^{100} x_{ij} \hat{\beta}_j.$$

In this prediction model, we employed inverse-probability weights as described above. For  $p$ -value calculation, the standard errors of  $\hat{\beta}$  were adjusted by design-based standard errors with approximation via Taylor series with the R-package *survey* [11].

#### **Model selection and validation**

The prediction accuracy was evaluated on the basis of a receiver-operator-characteristics (ROC) curve, which is a measure widely used for evaluating prediction models, when the disease response variable is binary [12]. Our prediction model, however, returned the probability of a subject having the disease, which is expressed as continuous values between 0 and 1. This probability can be dichotomized at a threshold, say  $c$ . If the respective probability value exceeds the threshold, the subject is labeled as being tested positive, otherwise as being tested negative. If  $pr$  denotes the predicted risk, there are two commonly used measures of correct prediction:

the sensitivity (or true positive rate) defined by  $\text{Sensitivity}(c) = P(pr > c | \text{Diseased})$  and the specificity (or true negative rate) defined by  $\text{Specificity}(c) = P(pr \leq c | \text{Not Diseased})$ . Displaying  $\text{Sensitivity}(c)$  against  $1 - \text{Specificity}(c)$  for all possible choices of  $c$  yields the ROC curve. Because of its common acceptance, we used the area under the ROC curve (AUC) as measure for model comparison and evaluation.

#### Creating confidence intervals for AUC values

Sample selection bias does not only affect the training of a model but also its evaluation. In addition, confidence intervals for the AUC would be biased without correcting for sample selection bias. Hence, both the evaluation and the confidence intervals should be corrected, e.g. by weighting [13]. However, the established evaluation weighting approach, which works well for certain loss functions, cannot be used for AUC values [13]. Therefore, we chose a bootstrap approach [14]. For the predicted risk  $\hat{y} = \{\hat{y}_1, \dots, \hat{y}_n\}$  and the corresponding true response  $y = \{y_1, \dots, y_n\}$ , the corrected AUC is then given by

$$\frac{1}{B} \sum_{b=1}^B L_{AUC}(\hat{y}^b, y^b)$$

where the pair  $(\hat{y}^b, y^b)$  corresponds to the  $b$ -th bootstrap sample,  $b \in \{1, \dots, B\}$ , which is built by resampling  $n$  elements with replacement from  $(\hat{y}, y)$ , using selection probabilities proportional to  $w_i$  for observation  $i$ .  $L_{AUC}$  denotes the loss function, which here corresponds to the AUC. We construct a percentile-confidence interval

$$[L_{AUC,2.5\%}(\hat{y}^b, y^b), L_{AUC,97.5\%}(\hat{y}^b, y^b)]$$

with  $L_{AUC,q}(\hat{y}^b, y^b)$  denoting the empirical  $q$ -quantile of the  $B$  bootstrap values  $L_{AUC}(\hat{y}^b, y^b)_{b=1, \dots, B}$ . For all our analyses, we chose  $B=10,000$ .

## Pairwise comparison of AUC values

In analogy to the corrected confidence intervals introduced above, we implemented a selection probability-based test for the pairwise comparison of two AUC values, when validated on the same data. Let  $\hat{y}^{(1)} = \{\hat{y}_1^{(1)}, \dots, \hat{y}_n^{(1)}\}$  be the predicted risk by a first classifier,  $\hat{y}^{(2)} = \{\hat{y}_1^{(2)}, \dots, \hat{y}_n^{(2)}\}$  the predicted risk by a second classifier and  $y = \{y_1, \dots, y_n\}$  the corresponding true response. We regard the corrected difference of AUCs

$$\frac{1}{B} \sum_{b=1}^B D_{AUC}(\hat{y}^{(1),b}, \hat{y}^{(2),b}, y^b)$$

with

$$D_{AUC}(\hat{y}^{(1),b}, \hat{y}^{(2),b}, y^b) := L_{AUC}(\hat{y}^{(1),b}, y^b) - L_{AUC}(\hat{y}^{(2),b}, y^b),$$

where the pair  $(\hat{y}^{(k),b}, y^b)$  for classifier  $k \in \{1, 2\}$  corresponds to the  $b$ -th bootstrap sample,  $b \in \{1, \dots, B\}$ , again, taken by using selection probabilities proportional to  $w_i$ . We construct a percentile-confidence interval for this difference by

$$[D_{AUC, 2.5\%}(\hat{y}^{(1),b}, \hat{y}^{(2),b}, y^b), D_{AUC, 97.5\%}(\hat{y}^{(1),b}, \hat{y}^{(2),b}, y^b)]$$

with  $D_{AUC, q}(\hat{y}^{(1),b}, \hat{y}^{(2),b}, y^b)$  denoting the empirical  $q$ -quantile of the  $B$  terms  $D_{AUC}(\hat{y}^{(1),b}, \hat{y}^{(2),b}, y^b)_{b=1, \dots, B}$ , again obtained via bootstrap with selection probabilities proportional to  $w_i$ . As before, we chose  $B=10,000$ . We consider two classifiers to perform significantly different – i.e. we reject the null hypothesis  $H_0$ : „AUC for classifier 1 is equal to AUC for classifier 2“ – if the confidence interval does not overlap with zero. Analogously, one can test  $H_0$ : „AUC for classifier 1 is less than or equal to AUC for classifier 2“, if AUC for classifier 1 is expected to be at least as good as classifier 2. The corresponding one-sided percentile-confidence-interval is  $[D_{AUC, 5\%}(\hat{y}^{(1),b}, \hat{y}^{(2),b}, y^b), 1]$ .

## Variable importance

In order to identify the most important variables for prediction, we considered those variables that were selected by the most successful prediction model. In most of the analyses the random forest turned out to perform best, however, for the best model on farm children also IPF-LASSO nearly performed as well as the random forest. In the latter analyses we investigated both models.

For random forests, we generally determined the most important variables by a standard measure of variable importance, the permutation importance. For  $j \in \{1, \dots, p\}$ , it is given by

$$VI_j = \frac{1}{ntree} \sum_{t \in \{1, \dots, ntree\}} \frac{1}{|OOB_t|} \sum_{i \in OOB_t} \{I(y_i \neq \hat{y}_{it}^{(-j)}) - I(y_i \neq \hat{y}_{it})\},$$

where *ntree* denotes the number of trees in the forest,  $I$  denotes the indicator function,  $OOB_t$  the set of indices for observations not selected for building tree  $t$  (out-of-the-bag observations),  $\hat{y}_{it}$  the predictions by the  $t$ -th tree before, and  $\hat{y}_{it}^{(-j)}$  after permuting the  $j$ -th variable's values. Further details can be found in Janitzka et al. 2016 [15].

In order to obtain a selection of predictive variables, we applied a non-parametric version of the permutation-based test proposed by Altman et al. 2010 [16]. First,  $VI_j$  for the  $j$ -th variable is calculated. Then, a distribution of null importance values is built using the following three steps: (i) permuting the values of the response  $y$ , (ii) fitting a new random forest using the permuted response and (iii) computing  $VI_j$  again. The latter procedure is repeated  $s$  times. The p-value is computed as the fraction of null importance values that exceed the originally computed importance over the  $s$  iterations.

For variable importance in genome-wide applications we used a further version based on cross-validation instead of OOB (out-of-the-bag) observations, proposed by [15].

For the IPF-LASSO we determined the most important variables as follows. Any version of LASSO per definition only selects variables contributing to good prediction. Hence, we interpreted the

selected variables as the important ones. We determined the degree of importance of these variables as the size of the highest penalty  $\lambda$  for which the corresponding coefficient was not shrunk to zero. This is plausible since the more predictive a variable, the stronger one can penalize it without its coefficient being set to zero by the LASSO-procedure.

### **Estimating the number of true/false positives**

In parts of our analyses we were rather interested in estimating *how many* SNPs were significant rather than figuring out *which*. Approaches like FDR or Bonferroni correction would have retained *which* SNPs were significant but would likely also have yielded false negatives: They would have marked SNPs non-significant which actually were significant. Hence, to estimate the number of significant SNPs, we proceeded as follows.

In a pessimistic scenario in which there are no truly influential SNPs, we expect 5% of these SNPs to be (wrongly) labelled significant when testing at a level of significance of 5%. This would yield 37,246.2 SNPs ( $744,924 * 0.05$ ) false positives in our application. In our study, we detected 40,000 significant SNPs (positives). This number exceeds the above estimate of the number of false positives by 3,700. We hence assume that there are 3,700 true positive findings. The estimate is conservative with respect to the general assumption of having 744,924 non-influential SNPs. It, however, ignores the possibility of strongly correlated SNPs, which may bias the calculation in both directions.

### **R Code**

In the spirit of open and reproducible research, R code is provided in the public online repository <https://github.com/fuchslab/gabriela>. For legal reasons, data cannot be provided for public download.

248 **Tables**

249 **Table E1: Distribution of environmental determinants and family history of atopy in GABRIELA**

| characteristic                                                   | cases                                                   | controls                                                | p-value |
|------------------------------------------------------------------|---------------------------------------------------------|---------------------------------------------------------|---------|
| center (study center (ref.: Innsbruck))                          | 16.4%<br>(Basel)<br>27.8%<br>(Munich)<br>37.7%<br>(Ulm) | 15.6%<br>(Basel)<br>32.6%<br>(Munich)<br>31.9%<br>(Ulm) | 0.123   |
| female (female gender)**                                         | 39.70%                                                  | 49.40%                                                  | 0.002   |
| age (age in years at 2007-01-01)**                               | 8.32<br>(se=0.06)                                       | 8.19<br>(se=0.06)                                       | 0.15    |
| BMI (body mass index)**                                          | 17.11<br>(se=0.11)                                      | 16.99<br>(se=0.11)                                      | 0.375   |
| farm (farming status)                                            | 9%                                                      | 13.60%                                                  | < 0.001 |
| siblings (>1 siblings)                                           | 41.80%                                                  | 44.60%                                                  | 0.374   |
| parental-education (high parental education)                     | 27.30%                                                  | 28.80%                                                  | 0.633   |
| smoking-pregnancy (maternal smoking in pregnancy)                | 12.40%                                                  | 8.50%                                                   | 0.037   |
| milk-last-yr (consumption of farm milk past 12 months)           | 13.40%                                                  | 19.40%                                                  | < 0.001 |
| milk-first-yr (consumption of farm milk in first year of life)   | 6.20%                                                   | 11.80%                                                  | < 0.001 |
| milk-first-3yrs (consumption of farm milk pregnancy to age 3yrs) | 20.70%                                                  | 27.60%                                                  | < 0.001 |
| cow-last-yr (contact with cows past 12 months)                   | 12.90%                                                  | 16.60%                                                  | 0.02    |
| cow-first-3yrs (contact with cows pregnancy to age 3yrs)         | 14.60%                                                  | 20.30%                                                  | 0.001   |
| straw-last-yr (contact with straw past 12 months)                | 15.70%                                                  | 21.10%                                                  | 0.009   |
| straw-first-3yrs (contact with straw pregnancy to age 3yrs)      | 12.40%                                                  | 16.20%                                                  | 0.009   |
| hay-last-yr (contact with hay past 12 months)                    | 29.70%                                                  | 33.50%                                                  | 0.145   |
| hay-first-3yrs (contact with hay pregnancy to age 3yrs)          | 21.60%                                                  | 26.10%                                                  | 0.028   |

|                                                                             |                                                                 |                                                                  |         |
|-----------------------------------------------------------------------------|-----------------------------------------------------------------|------------------------------------------------------------------|---------|
| cow/straw-first-3yrs (contact with cows and/or straw pregnancy to age 3yrs) | 5.1% (straw only) 7.5% (cow only) 7.2% (both)                   | 4.8% (straw only) 8.6% (cow only) 11.4% (both)                   | 0.024   |
| cow/straw-last-yr (contact with cows and/or straw past 12 months)           | 8.7% (straw only) 5.9% (cow only) 7.1% (both)                   | 11.9% (straw only) 7.1% (cow only) 9.4% (both)                   | 0.029   |
| barn-first-3yrs (stay in barn pregnancy to age 3yrs)                        | 14.60%                                                          | 18.50%                                                           | 0.015   |
| barn-last-yr (stay in barn past 12 months)                                  | 16.50%                                                          | 20.70%                                                           | 0.021   |
| stable-weekly-last-yr (stay in cattle stable once/week)                     | 8.70%                                                           | 12%                                                              | 0.016   |
| traffic (how often do trucks or busses drive by)                            | 35.9% (rarely) 27.6% (often during day) 9.5% (almost whole day) | 39.3% (rarely) 26.9% (often during day) 10.1% (almost whole day) | 0.558   |
| mold-last-yr (rooms with visible mold)                                      | 18%                                                             | 11.60%                                                           | 0.004   |
| dog-first-yr (dog allowed to stay in the room in first year of life)        | 5.30%                                                           | 6.50%                                                            | 0.456   |
| cat-first-yr (cat allowed to stay in the room in first year of life)        | 8.50%                                                           | 15.50%                                                           | 0.001   |
| dog-2-3yrs (dog allowed to stay in the room in 2. and 3. year of life)      | 6.90%                                                           | 7.20%                                                            | 0.838   |
| cat-2-3yrs (dog allowed to stay in the room in 2. and 3. year of life)      | 11.50%                                                          | 21.40%                                                           | < 0.001 |
| dog-4-5yrs (dog allowed to stay in the room in 4. and 5. year of life)      | 6.30%                                                           | 6.90%                                                            | 0.736   |
| cat-4-5yrs (dog allowed to stay in the room in 4. and 5. year of life)      | 15%                                                             | 24.60%                                                           | < 0.001 |
| dog-last-yr (dog allowed to stay in the room past 12 months)                | 7.70%                                                           | 8.90%                                                            | 0.506   |

|                                                                                 |                                                             |                                                             |         |
|---------------------------------------------------------------------------------|-------------------------------------------------------------|-------------------------------------------------------------|---------|
| cat-last-yr (cat allowed to stay in the room past 12 months)                    | 19.80%                                                      | 30.10%                                                      | < 0.001 |
| children-household (people between 0-18 years live currently in your household) | 2.39<br>(se=0.04)                                           | 2.37<br>(se=0.04)                                           | 0.921   |
| adults-household (people 18 years and older live currently in your household)   | 2.13<br>(se=0.06)                                           | 2.18<br>(se=0.03)                                           | 0.005   |
| parents-smoke-ever (have parents ever smoked after birth)                       | 61.60%                                                      | 55.60%                                                      | 0.052   |
| parents-smoke-currently (do parents currently smoke)                            | 33.20%                                                      | 26.80%                                                      | 0.024   |
| daycare (regular attendance of facilities with children until school enrolment) | 92%                                                         | 92%                                                         | 0.981   |
| antibiotics-pregnancy (mother takes antibiotics during pregnancy)               | 11.80%                                                      | 8.30%                                                       | 0.074   |
| cattle-last-yr (stay in cattle stable past 12 months)                           | 11.10%                                                      | 14.20%                                                      | 0.043   |
| cattle-first-yr (stay in cattle stable in first year of life)                   | 6.10%                                                       | 10.20%                                                      | < 0.001 |
| birth-season (season of birth (ref.: Summer))                                   | 25.6%<br>(Spring)<br>27.2%<br>(Autumn)<br>25.4%<br>(Winter) | 22.6%<br>(Spring)<br>23.5%<br>(Autumn)<br>25.6%<br>(Winter) | 0.11    |
| family-atopy (parental atopy or sibling atopy)                                  | 70%                                                         | 49.70%                                                      | < 0.001 |
| family-asthma (parental asthma or sibling asthma)                               | 30.60%                                                      | 12.40%                                                      | < 0.001 |
| family-hayfev (parental hay fever or sibling hay fever)                         | 48.30%                                                      | 36.60%                                                      | < 0.001 |
| family-fheczema (parental eczema or sibling eczema)                             | 36.50%                                                      | 25.80%                                                      | < 0.001 |

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

\*\* variables were subsumed under the group of demographic variables in specific models

se = standard error of mean, yr = year

**Table E2: : Confidence intervals for bootstrap test of the difference between two AUCs for combined prediction model of childhood asthma against a single determinant prediction model**

| Subgroup          | Best stand-alone model<br>(Figure 2, random forest) | Best synergy model<br>(Figure 3, random forest)      | lower bound of<br>95%-confidence interval<br>for AUC difference |
|-------------------|-----------------------------------------------------|------------------------------------------------------|-----------------------------------------------------------------|
| all children      | family history                                      | family history<br>+ demographics<br>+ environment    | 0.0051                                                          |
| non-farm children | family history                                      | family history<br>+ demographics<br>+ environment    | 0.0078                                                          |
| farm children     | family history                                      | family history<br>+ demographics<br>+ candidate SNPs | 0.0019                                                          |

For comparison of the AUC of the best prediction model combining several determinants (Figure 3) to best the prediction model for one determinant (always family-history, Figure 1), a one-sided bootstrap test for the difference between two AUCs was calculated with a level of significance of 0.05. The resulting lower bounds of the 95%-confidence interval of the differences in the AUCs are shown. For each subgroup (all children, non-farm children, farm children) the confidence interval for the difference to the AUCs does not overlap with the null – the AUC of the synergy model is significantly higher than the AUC for family history.

268 **Table E3: Distribution of environmental determinants and family history of atopy in PASTURE**

| characteristic                                                              | cases                                          | controls                                       | p-value |
|-----------------------------------------------------------------------------|------------------------------------------------|------------------------------------------------|---------|
| female (female gender)                                                      | 33.30%                                         | 50.10%                                         | 0.005   |
| age (age in years at 2007-01-01)                                            | 6.12<br>(se=0.03)                              | 6.12 (se=0.01)                                 | 0.823   |
| BMI (body mass index)                                                       | 15.73<br>(se=0.17)                             | 15.78<br>(se=0.07)                             | 0.501   |
| farm (farming status)                                                       | 37.20%                                         | 49.30%                                         | 0.041   |
| siblings (>1 siblings)                                                      | 33.30%                                         | 34%                                            | 0.905   |
| parental-education (high parental education)                                | 89.70%                                         | 82.90%                                         | 0.121   |
| smoking-pregnancy (maternal smoking in pregnancy)                           | 14.50%                                         | 12.10%                                         | 0.555   |
| milk-last-yr (consumption of farm milk past 12 months)                      | 30.80%                                         | 43.20%                                         | 0.034   |
| milk-first-yr (consumption of farm milk in first year of life)              | 21.80%                                         | 33.50%                                         | 0.035   |
| milk-first-3yrs (consumption of farm milk pregnancy to age 3yrs)            | 41.10%                                         | 53.80%                                         | 0.038   |
| cow-last-yr (contact with cows past 12 months)                              | 23.70%                                         | 40.50%                                         | 0.004   |
| cow-first-3yrs (contact with cows pregnancy to age 3yrs)                    | 31.40%                                         | 44.50%                                         | 0.035   |
| straw-last-yr (contact with straw past 12 months)                           | 19.70%                                         | 40%                                            | < 0.001 |
| straw-first-3yrs (contact with straw pregnancy to age 3yrs)                 | 26.90%                                         | 43.80%                                         | 0.007   |
| hay-last-yr (contact with hay past 12 months)                               | 25.60%                                         | 42.80%                                         | 0.003   |
| hay-first-3yrs (contact with hay pregnancy to age 3yrs)                     | 38.20%                                         | 46.70%                                         | 0.181   |
| cow/straw-first-3yrs (contact with cows and/or straw pregnancy to age 3yrs) | 0% (straw only) 3% (cow only) 27.3% (both)     | 2.1% (straw only) 3.2% (cow only) 41.7% (both) | 0.053   |
| cow/straw-last-yr (contact with cows and/or straw past 12 months)           | 1.4% (straw only) 6.8% (cow only) 17.6% (both) | 5.1% (straw only) 6.7% (cow only) 34% (both)   | 0.006   |

|                                                                                 |                |                |         |
|---------------------------------------------------------------------------------|----------------|----------------|---------|
| barn-first-3yrs (stay in barn pregnancy to age 3yrs)                            | 84%            | 77.60%         | 0.457   |
| barn-last-yr (stay in barn past 12 months)                                      | 20.50%         | 42.80%         | < 0.001 |
| stable-weekly-last-yr (stay in cattle stable once/week)                         | 29.50%         | 52.30%         | < 0.001 |
| mold-last-yr (rooms with visible mold)                                          | 15.40%         | 16%            | 0.891   |
| dog-first-yr (dog allowed to stay in the room in first year of life)            | 15.40%         | 18.70%         | 0.472   |
| cat-first-yr (cat allowed to stay in the room in first year of life)            | 26.90%         | 29.30%         | 0.658   |
| dog-2-3yrs (dog allowed to stay in the room in 2. and 3. year of life)          | 18.70%         | 17.10%         | 0.726   |
| cat-2-3yrs (dog allowed to stay in the room in 2. and 3. year of life)          | 25.70%         | 30.70%         | 0.367   |
| dog-4-5yrs (dog allowed to stay in the room in 4. and 5. year of life)          | 25.30%         | 33.10%         | 0.168   |
| cat-4-5yrs (dog allowed to stay in the room in 4. and 5. year of life)          | 37.80%         | 54.90%         | 0.005   |
| dog-last-yr (dog allowed to stay in the room past 12 months)                    | 16.70%         | 19.10%         | 0.602   |
| cat-last-yr (cat allowed to stay in the room past 12 months)                    | 34.60%         | 40.60%         | 0.299   |
| children-household (people between 0-18 years live currently in your household) | 1.26 (se=0.13) | 1.16 (se=0.04) | 0.395   |
| adults-household (people 18 years and older live currently in your household)   | 2.05 (se=0.03) | 2.18 (se=0.02) | 0.106   |
| parents-smoke-ever (have parents ever smoked after birth)                       | 68.90%         | 61.30%         | 0.198   |
| parents-smoke-currently (do parents currently smoke)                            | 29.70%         | 27.10%         | 0.621   |
| daycare (regular attendance of facilities with children until school enrolment) | 93.80%         | 92.30%         | 0.682   |
| antibiotics-pregnancy (mother takes antibiotics during pregnancy)               | 29.90%         | 25.80%         | 0.441   |
| cattle-last-yr (stay in cattle stable past 12 months)                           | 35.30%         | 45.40%         | 0.109   |
| cattle-first-yr (stay in cattle stable in first year of life)                   | 35.10%         | 47.20%         | 0.042   |

|                                                         |                                                             |                                                        |         |
|---------------------------------------------------------|-------------------------------------------------------------|--------------------------------------------------------|---------|
| birth-season (season of birth (ref.: Summer))           | 25.6%<br>(Spring)<br>23.1%<br>(Autumn)<br>32.1%<br>(Winter) | 27% (Spring)<br>22.8%<br>(Autumn)<br>26.2%<br>(Winter) | 0.642   |
| family-atopy (parental atopy or sibling atopy)          | 83.10%                                                      | 61.10%                                                 | < 0.001 |
| family-asthma (parental asthma or sibling asthma)       | 35.10%                                                      | 14.20%                                                 | < 0.001 |
| family-hayfev (parental hay fever or sibling hay fever) | 68.90%                                                      | 44.20%                                                 | < 0.001 |
| family-fheczema (parental eczema or sibling eczema)     | 50%                                                         | 30.20%                                                 | < 0.001 |

\* p-values based on Fisher's exact test or, in case of continuous variables, Wilcoxon tests  
se = standard error of mean, yr = year

Figures

Figure E1: Participant flow in the GABRIELA study

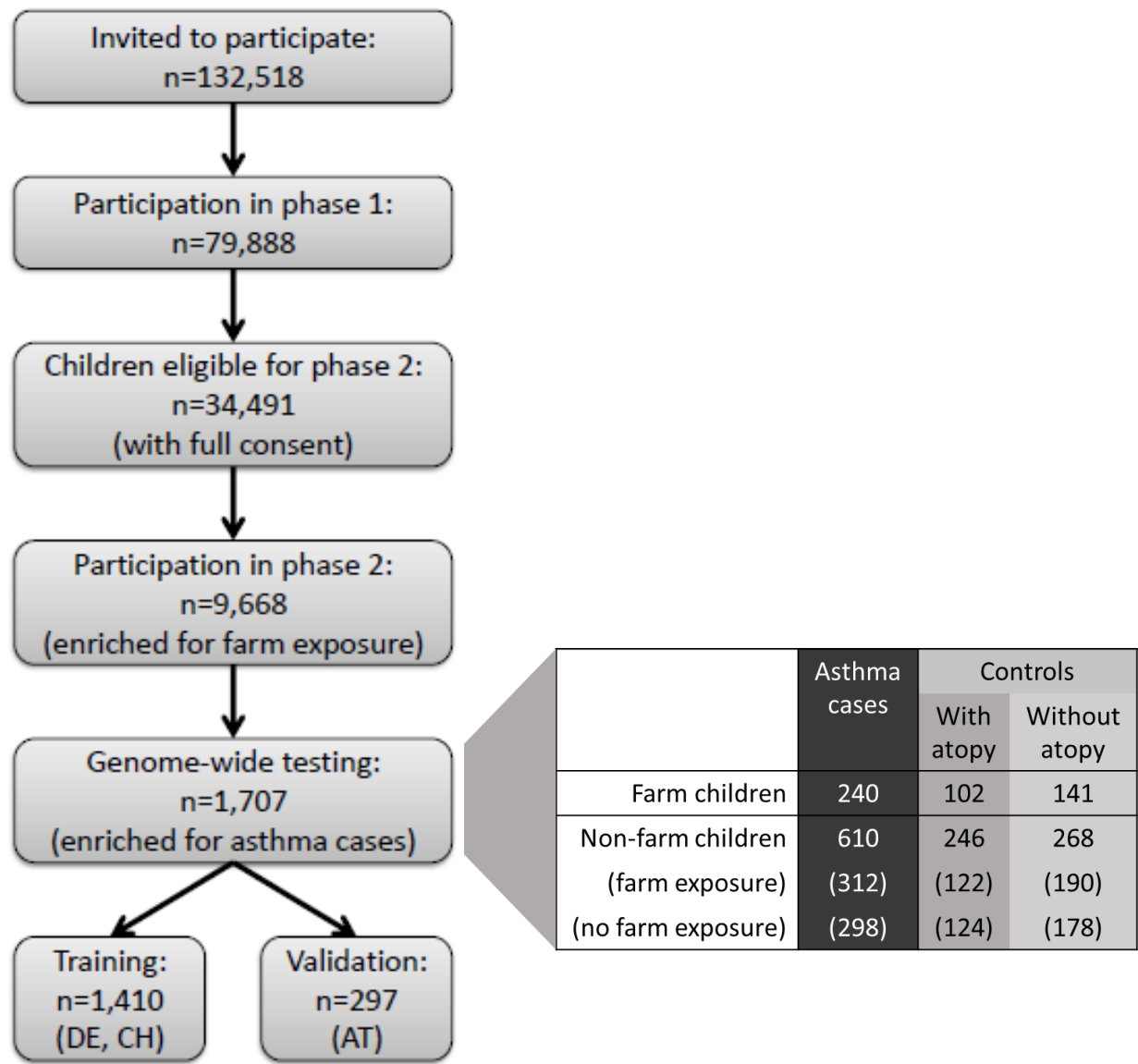
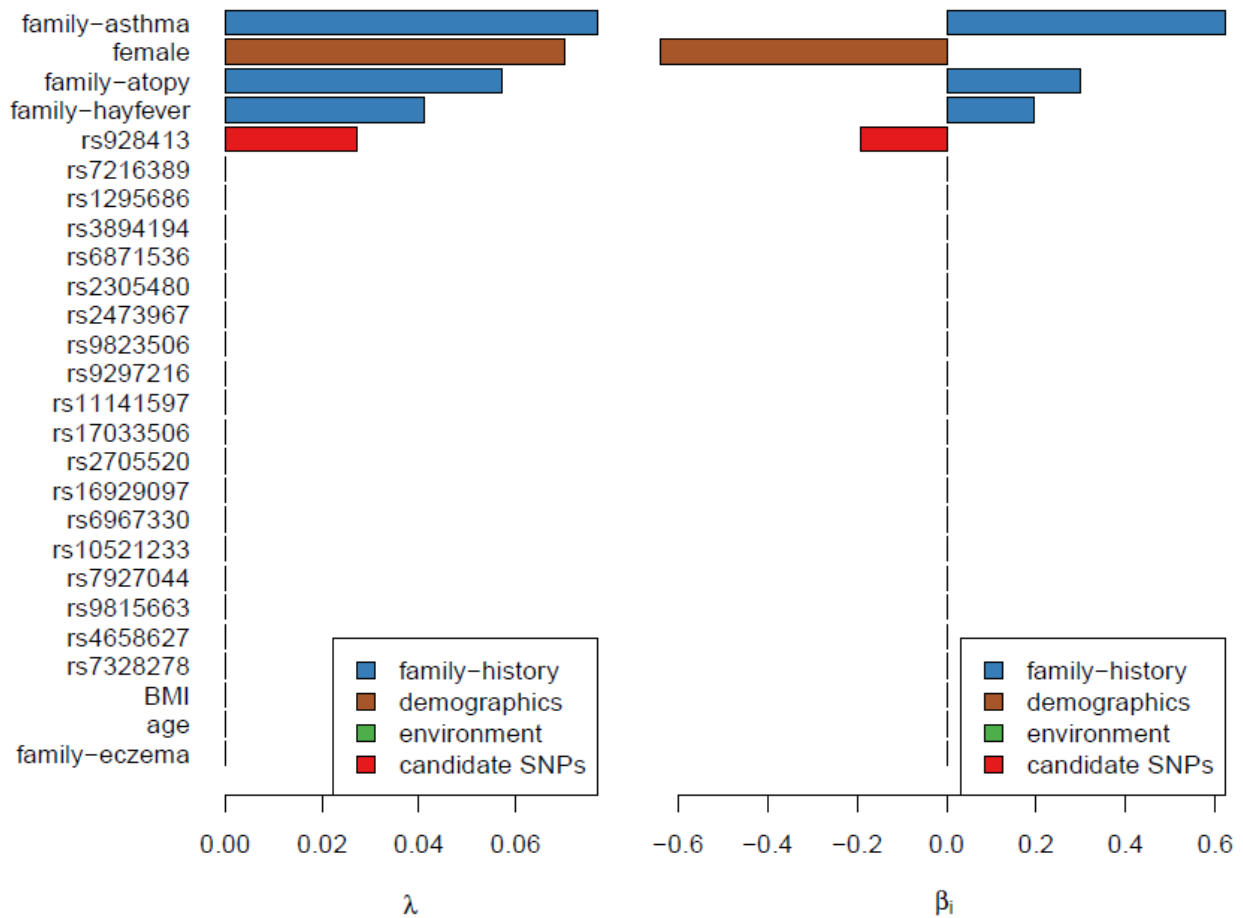
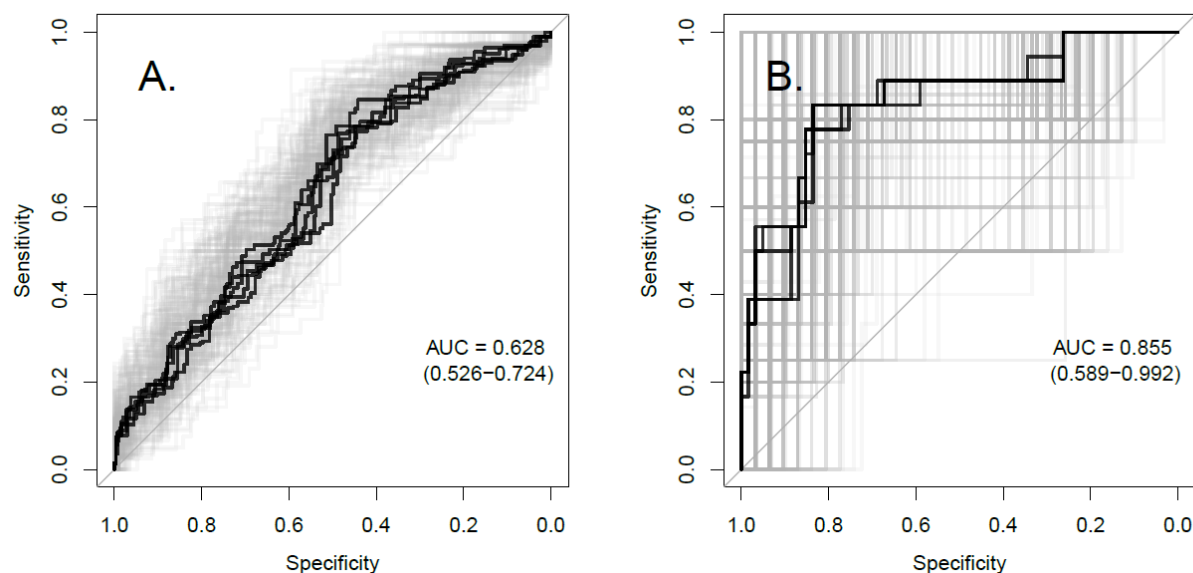


**Figure E2: Variable importance by IPF-LASSO for the best model in farm children**



Determinants of the prediction model IPF-LASSO in farm children which performed similarly to the random forest in the internal validation (Figure 3), sorted by importance ( $\lambda$ , left panel) with respective effect sizes ( $\beta_i$ , right panel). Positive  $\beta_i$  values represent risk factors; negative values represent protective factors;  $\lambda$  and  $\beta_i$  values are averaged over the 5 imputation datasets.

297 **Figure E3: ROC-curves for best models for non-farm and farm children validated on the Austrian**  
 298 **arm of GABRIELA**



299  
 300 Mean weighted ROC-curves of the 5 imputations (black) with 1000 weighted ROC curves obtained  
 301 via bootstrap (gray) for predictions on the Austrian GABRIELA arm. A. Best prediction model for  
 302 non-farm children (random forest) validated on the Austrian GABRIELA arm. B. Best prediction  
 303 model, resulting from model averaging over prediction scores of random forest and IPF-LASSO,  
 304 for farm children. Here, for guaranteeing robustness of AUC confidence intervals, high class  
 305 imbalance (occurring for farm children on the Austrian GABRIELA arm) was avoided by forcing at  
 306 least 10% asthma cases into each bootstrap sample.

## References

1. van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software* 2011; 45: 1-67.
2. Rubin DB. Multiple Imputation for Nonresponse in Surveys. Wiley; 1987.
3. Krautenbacher N, Theis FJ, Fuchs C. Correcting classifiers for sample selection bias in two-phase case-control studies. *Computational and Mathematical Methods in Medicine* 2017.
4. Fahrmeir L, Kneib T, Lang S. Regression. Springer Berlin Heidelberg; 2009.
5. Skinner C. J HD, Smith TMF. Analysis of complex surveys. 1989.
6. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning. New York, NY, USA: Springer New York Inc.; 2001.
7. 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. *Computational and Mathematical Methods in Medicine* 2017; 2017: 7691937.
8. Breiman L. Random forests. *Machine Learning* 2001; 45: 5-32.
9. Chen C, Liaw A, Breiman L. Using random forest to learn imbalanced data. *University of California, Berkeley* 2004: 1-12.
10. Wu J, Pfeiffer RM, Gail MH. Strategies for developing prediction models from genome-wide association studies. *Genetic epidemiology* 2013; 37: 768-777.
11. Lumley T. Analysis of complex survey samples. *Journal of Statistical Software* 2004; 9: 1-19.
12. Fawcett T. An introduction to ROC analysis. *Pattern Recogn Lett* 2006; 27: 861-874.
13. Cortes C, Mohri M, Riley M, Rostamizadeh A. Sample Selection Bias Correction Theory. 2008: 16-16.
14. Efron B. Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics* 1979; 7: 1-26.

- 331 15. Janitza S, Celik E, Boulesteix A-L. A computationally fast variable importance test for random  
332 forests for high-dimensional data. Springer Berlin Heidelberg; 2016.
- 333 16. Altmann A, Toloşi L, Sander O, Lengauer T. Permutation importance: a corrected feature  
334 importance measure. *Bioinformatics* 2010; 26: 1340-1347.
- 335