# Online supplement

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

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

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

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

3 University Children’s Hospital Regensburg (KUNO), Regensburg, Germany

4 Hannover Medical School, Clinic for Paediatric Pneumology and Neonatology, Hannover, Germany

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

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

14 Faculty 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](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](#_ENREF_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](#_ENREF_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](#_ENREF_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](#_ENREF_4)]. Let be the sample size, the number of variables in the data. Then let denote the response, i.e. the disease status ( for no diesease or 1 for disease), for subject . Further let contain the variables used for prediction (i.e. is a matrix with rows and columns), be the variable values for subject and be the value for subject and variable

Then in the multivariable logistic regression model, is assumed to be Bernoulli distributed with success probability where and are unknown parameters, which represent effects of the variables on .

During maximum likelihood estimation, the log-likelihood with respect to and is maximized.

To correct the point estimates for the effects for sample selection bias, the observations are weighted by inverse-probability weights [[3](#_ENREF_3)] which are given after the stratified random selection process [[5](#_ENREF_5)]. Then the modified log-likelihood is given by

Since this is a modification of maximum likelihood estimation in the common sense, this function is called pseudo-maximum-likelihood [[5](#_ENREF_5)].

The use of multivariable logistic regression is suitable for classification problems, however, can be improved in prediction by using regularization [[6](#_ENREF_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, with as tuning parameter and a penalization function. The latter can be chosen to be based on the -norm, i.e. (LASSO-penalty), or the -norm, i.e. (ridge-penalty), or include both (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 =T0.5 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 was obtained via 10-fold cross-validation. The multiple 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](#_ENREF_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](#_ENREF_7)]. Their coefficients are obtained by minimizing

with the penalty applied to the variables from modality

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](#_ENREF_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](#_ENREF_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](#_ENREF_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](#_ENREF_10)], the best prediction strategy in terms of AUC values first estimates univariate associations by simple logistic regression models for each SNP , i.e. exp(. In a second step, a score for subject is calculated as the sum of the univariately estimated coefficients of the SNPs with the lowest p-values:

In this prediction model, we employed inverse-probability weights as described above. For -value calculation, the standard errors of were adjusted by design-based standard errors with approximation via Taylor series with the R-package *survey* [[11](#_ENREF_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](#_ENREF_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 denotes the predicted risk, there are two commonly used measures of correct prediction: the sensitivity (or true positive rate) defined by Sensitivity() Diseased) and the specificity (or true negative rate) defined by Specificity() Not Diseased). Displaying Sensitivity() against 1- Specificity() for all possible choices of 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](#_ENREF_13)]. However, the established evaluation weighting approach, which works well for certain loss functions, cannot be used for AUC values [[13](#_ENREF_13)]. Therefore, we chose a bootstrap approach [[14](#_ENREF_14)]. For the predicted risk and the corresponding true response , the corrected AUC is then given by

where the pair corresponds to the *b-*th bootstrap sample, , which is built by resampling elements with replacement from , using selection probabilities proportional to for observation denotes the loss function, which here corresponds to the AUC. We construct a percentile-confidence interval

with denoting the empirical q-quantile of the B bootstrap values . 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 be the predicted risk by a first classifier, the predicted risk by a second classifier and the corresponding true response. We regard the corrected difference of AUCs

with

where the pair for classifier corresponds to the *b-*th bootstrap sample, , again, taken by using selection probabilities proportional to . We construct a percentile-confidence interval for this difference by

with denoting the empirical q-quantile of the B terms , again obtained via bootstrap with selection probabilities proportional to . As before, we chose B=10,000. We consider two classifiers to perform significantly different – i.e. we reject the null hypothesis „AUC for classifier 1 is equal to AUC for classifier 2“ – if the confidence interval does not overlap with zero. Analogously, one can test „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 .

###### 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 , it is given by

where denotes the number of trees in the forest, denotes the indicator function, the set of indices for observations not selected for building tree (out-of-the-bag observations), the predictions by the-th tree before, and after permuting the -th variable’s values. Further details can be found in Janitza et al. 2016 [[15](#_ENREF_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](#_ENREF_16)]. First, for the -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 , (ii) fitting a new random forest using the permuted response and (iii) computing again. The latter procedure is repeated times. The p-value is computed as the fraction of null importance values that exceed the originally computed importance over the 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](#_ENREF_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 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.

# Tables

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% (Basel) 27.8% (Munich) 37.7% (Ulm) | 15.6% (Basel) 32.6% (Munich) 31.9% (Ulm) | 0.123 |
| female (female gender)\*\* | 39.70% | 49.40% | 0.002 |
| age (age in years at 2007-01-01)\*\* | 8.32 (se=0.06) | 8.19 (se=0.06) | 0.15 |
| BMI (body mass index)\*\* | 17.11 (se=0.11) | 16.99 (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 (se=0.04) | 2.37 (se=0.04) | 0.921 |
| adults-household (people 18 years and older live currently in your household) | 2.13 (se=0.06) | 2.18 (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% (Spring) 27.2% (Autumn) 25.4% (Winter) | 22.6% (Spring) 23.5% (Autumn) 25.6% (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  (Figure 2, random forest) | Best synergy model (Figure 3, random forest) | lower bound of  95%-confidence interval  for AUC difference |
| all children | family history | family history   + demographics   + environment | 0.0051 |
| non-farm  children | family history | family history   + demographics   + environment | 0.0078 |
| farm  children | family history | family history   + demographics   + 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.

**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 (se=0.03) | 6.12 (se=0.01) | 0.823 |
| BMI (body mass index) | 15.73 (se=0.17) | 15.78 (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% (Spring) 23.1% (Autumn) 32.1% (Winter) | 27% (Spring) 22.8% (Autumn) 26.2% (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

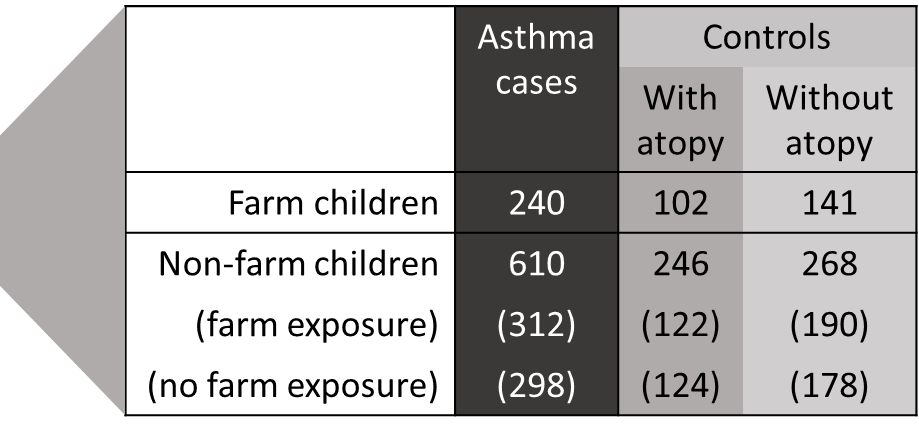


Figure E2: Variable importance by IPF-LASSO for the best model in farm children



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 (, left panel) with respective effect sizes (, right panel). Positive βi values represent risk factors; negative values represent protective factors; λ and βivalues are averaged over the 5 imputation datasets.

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



Mean weighted ROC-curves of the 5 imputations (black) with 1000 weighted ROC curves obtained via bootstrap (gray) for predictions on the Austrian GABRIELA arm. A. Best prediction model for non-farm children (random forest) validated on the Austrian GABRIELA arm. B. Best prediction model, resulting from model averaging over prediction scores of random forest and IPF-LASSO, for farm children. Here, for guaranteeing robustness of AUC confidence intervals, high class imbalance (occurring for farm children on the Austrian GABRIELA arm) was avoided by forcing at 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.

15. Janitza S, Celik E, Boulesteix A-L. A computationally fast variable importance test for random forests for high-dimensional data. Springer Berlin Heidelberg; 2016.

16. Altmann A, Toloşi L, Sander O, Lengauer T. Permutation importance: a corrected feature importance measure. *Bioinformatics* 2010; 26: 1340-1347.