PROTOCOL FOR STATISTICAL ANALYSES OF HEALTH EFFECTS IN THE STUDY ENTITLED "PESTICIDE EXPOSURE, ASTHMA AND DIABETES IN UGANDA (PEXADU)"

Version 1.0

Last edited on Monday, November 25, 2019

Martin Rune Hassan Hansen[1,2*], Erik Jørs[3], Annelli Sandbæk[1,4], Daniel Sekabojja[5], John Ssempebwa[6], Ruth Mubeezi[6], Philipp Staudacher[7,8], Samuel Fuhrimann[9], Torben Sigsgaard[1], Vivi Schlünssen[1,2]

1. Aarhus University, Aarhus C, Denmark
2. National Research Center for the Working Environment, Copenhagen, Denmark
3. Odense University Hospital, Odense, Denmark
4. Steno Diabetes Center Aarhus, Aarhus, Denmark
5. Uganda National Association of Community and Occupational Health, Kampala, Uganda
6. Makerere University, Kampala, Uganda
7. Eawag, Swiss Federal Institute of Aquatic Science and Technology, Dübendorf, Switzerland
8. Institute of Biogeochemistry and Pollutant Dynamics, ETH Zurich, Zürich, Switzerland
9. Utrecht University, Utrecht, Netherlands

* Corresponding author:

Martin Rune Hassan Hansen

Section for Environment, Work and Health, Danish Ramazzini Centre, Department of Public Health, Aarhus University

Bartholins Allé 2, Building 1260

DK-8000 Aarhus C

Denmark

martinrunehassanhansen@ph.au.dk

Contents

1 Introduction

The PEXADU project is a study of the effects of pesticide exposure on glycemic regulation and lung function among 364 small-scale farmers from the Wakiso District in Uganda. The study has been described in detail in a previously published document[1]. We originally planned to use subjective information as our primary exposure metric, but planned to validate the subjective reports before the analyses. In the previous protocol, we presented our analysis plans for validation of subjective information against the acetylcholine esterase (AChE), a biomarker of exposure to organophosphate and carbamate insecticides. The analyses in question have shown poor correlation between subjective pesticide exposure information (results not yet published) and acetylcholine esterase, which may be due to exposure routes not evaluated (diet, re-entry work in sprayed fields, etc.). For the analyses of glycemic regulation and lung function, we have therefore decided that we will use AChE as our exposure metric, as described in detail below.

1.1 Level of significance

p-values ≤ 0.05 will be considered significant.

A relatively high number of statistical tests will be carried out because of the many independent variables we want to examine - e.g. HbA_{1c} (continuous), FPG (continuous), diabetes (yes/no), FEV_1 z-score (continuous), FVC z-score (continuous) and FEV_1/FVC z-score (continuous). The number of tests means that there is a risk of mass significance, i.e. finding statistically significant results where no true differences exist. By definition, this will happen in 5% (= the level of significance) of all tests. While we will not try to account formally for this (e.g. by Bonferroni correction[2]), it will be kept in mind when interpreting results.

1.2 Interdependence of data

Some of the PEXADU participants were relatives. To take this into account, data will be analyzed in random effect models with family as a random effect. A “family” will be defined as a group of people where all members of the group are genetically related to at least one other group member.

2 Analysis plan for lung function tests

2.1 Descriptive statistics

Before more advanced statistical analyses, we will present the following table of demographic information (all data at baseline):

Characteristic	**All participants**	**Conventional farmer's group**	**Semi-organic farmer's group**
Total n			
Sex			
Male, n (%)			
Female, n (%)			
Age in years: Median, IQR			
Educational level (years of full-time schooling): Median, IQR)			
Ethnicity			
Baganda, n (%)			
Other, n (%)			
Body mass index (kg/m2): Median, IQR			
Body height (cm): Median, IQR			
Smoking:			
Ever smoker, n (%)			
Current smoker, n (%)			
Pack-years for ever-smokers: Median, IQR			
Grams of tobacco per day for current smokers: Median, IQR			
Main fuel used for cooking in household: Type 1: n (%) Type 2: n (%) Type 3: n (%)			
Current hours of cooking per week: Median, IQR			
Cumulated life-time hours of cooking: Median, IQR			

Next, we will present the following table of spirometric and biochemical results:

		Phase 1	Phase 2	Phase 3
Raw spirometric indices (not converted to Z-scores): Median, IQR	FEV_1 n			
	FVC n			
	FEV_1/FVC n			
	PEF n			
	FEF_{25} n			
	FEF_{50} n			
	FEF_{75} n			
	FEF_{25-75} n			
Z-scores of spirometric indices: Median, IQR	FEV_1 n			
	FVC n			
	FEV_1/FVC n			
	FEF_{25-75} n			
Acetylcholine esterase levels: Median, IQR	AChE/Hb, U/g n missing			
	Hb, g/dL n missing			
	AChE/Hb, U/g n missing			

2.2 Quality assessment of spirometry

According to American Thoracic Society guidelines on the standardization of spirometry,[3] "acceptable" quality blows may not have any of the following:

1. A cough in the first second of the maneuver
2. Glottis closure or early termination of the blow (i.e., the blow should last ≥ 6 seconds, or the volume-time curve should reach a plateau)
3. Obstruction of mouthpiece
4. Sub-maximal blowing effort
5. Leak around mouthpiece
6. Slow start of exhalation (defined as extrapolated volume ≥5 % of FVC or ≥ 150 ml, whichever is greater).
7. Extra inhalation during the maneuver.

An adequate test (according to the ATS) must have at least three acceptable maneuvers. In addition, the difference between the largest and the second largest values of both FEV_1 and FVC must be ≤ 150 ml (≤100 ml if the largest FVC is ≤ 1.00 liter).[3] To maximize the amount of data available for our analyses, our quality criteria will be less strict. We will call a test adequate if it contains at least two (rather than three) acceptable maneuvers, and reproducibility will be deemed adequate as long as the difference in best and second-best FEV_1 and FVC is ≤250 ml.

When calculating summary measures from spirometry, we will consider a blow “usable”, as long as it does not have problems 1 or 6 mentioned above. Furthermore, blows are not usable for FEV_1 if there is an extra inhalation in the first second, and they are not usable for FVC if there is any extra inhalation during the maneuver. The PEF recorded will be the highest PEF in a blow that is also usable for FEV_1. FEF_{25}, FEF_{50}, FEF_{75} and FEF_{25-75} will be recorded from the blow that has the highest sum of FEV_1 + FVC.

2.3 Outcomes and covariates that will be taken into consideration

The primary outcomes are Z-scores for FEV_1, FVC and FEV_1/FVC, calculated based on the GLI 2012 equations (setting ethnicity as "African-American", as normal values are not available for Ugandans). These three outcomes are considered equally important. The remaining spirometric indices are considered secondary.

Outcome group	**Outcomes in group**
1 (primary outcomes)	FEV_1 Z-score, FVC Z-score, FEV_1/FVC Z-score
2 (secondary outcomes)	FEV_1, FVC, FEV_1/FVC, PEF, FEF_{25}, FEF_{50}, FEF_{75}, FEF_{25-75}, FEF_{25-75} Z-score

Depending on the statistical model (more details below), a priori we have decided that we will consider the following variables possible confounders of pulmonary function: Age, sex, height, socioeconomic status, pack-years of smoking, exposure to biofuel smoke, BMI, and the specific spirometer used for testing. As a large majority (78%) of the study population is of Baganda ethnicity, and the remaining participants are split across 18 different ethnicities, we will not take ethnicity into account in our analyses.

2.4 Description of the statistical models

To take into account the family relations in the study population, and the repeated measurements of both outcome and exposure, our primary statistical model will be a random coefficient model (RCM), i.e. mixed effect model of the following form:

$$y = \beta_{exposure} \times exposure + \beta_c \times c + \theta \times \rho + \alpha + \tau + \varepsilon$$

y is the outcome. $\beta_{exposure}$ is the regression coefficient for the effect of the exposure variable on y; $\beta_{exposure}$ is normally distributed in the study population, and each person has his/her own level of $\beta_{exposure}$. β_c is the regression coefficient for the effect of confounder c on y; all members of the population have the same β_c. θ is a fixed effect for phase of the examination (ρ = 1, 2 or 3), α is a random effect for the family of which the participant is a member, τ is a person-specific random effect, and ε is an error term.

The primary outcomes will also be investigated in a fixed effect (FE) model. This is a mixed effect model of the following form:

$$y_t - y_0 = \beta_x \times (x_t - x_0) + \alpha + \varepsilon$$

, where $y_t - y_0$ is the change in the outcome between two different phases. $x_t - x_0$ is the change in the independent variable x between two phases, and β_x is the regression coefficient for the fixed effect of Δx on Δy (all participants have the same β_x). α is a random effect for the family of which the participant is a member, and ε is an error term. Our primary FE model will compare phases 1+3. Sensitivity analyses of the FE model will compare phases 1+2 and 2+3, respectively.

The point of the FE model is to remove effects of unknown confounders on the relationship between x and y. If the effect of confounder c on y is additive (i.e., it does not interact with the effect of x on y), and if c is constant between phases, then c should have no effect on Δy in the FE model. Furthermore, if the effect of Δx is modelled non-linearly, the FE model allows us to evaluate if an effect of *x* on *y* is reversible or irreversible (within the timespan between examinations).

Because we have incomplete information on family relationships in the study population, including a random effect for family ID may be insufficient to completely account for the interdependence of observations, leading to a risk of falsely low standard errors in our estimates of effects. In some sensitivity analyses of the RCM, we will therefore account for the interdependence of observations (family relations) not only by including a random effect for family, but also by a bootstrapping procedure as implemented in the Stata command **`bootstrap`**, with 200 repetitions. To keep the time required to run each iteration of the regression model manageable, the model used in the bootstrap procedure will assume that all continuous variables influence the outcome in a linear fashion.

Regarding missing data, all analyses will be performed as “full information”, i.e. we will only include individuals with non-missing values for all the variables in the model.

2.5 Overview of statistical analyses

	Outcome metrics	**Exposure metric**	**Statistical model**	**Assumptions regarding linearity**	**Handling of interdependent data (family)**	**Handling of multiple AChE measurements in same phase***	**Classification**	**Covariate set**	**Reported where?**
Primary model for primary outcomes	Group 1	AChE/Hb	RCM	Cubic splines, 4 knots	RE	First measurement used	Primary analysis	Basic set	Main article
							Primary analysis	Unadjusted	
							Sensitivity analysis	Extended set	
Secondary model for primary outcomes	Group 1 (phase 1+3)	AChE/Hb	FE	Cubic splines, 4 knots	RE	First measurement used	Secondary analysis	Basic set	Main article
								Unadjusted	
								Extended set	
Primary and only model for secondary outcomes	Group 2	AChE/Hb	RCM	Cubic splines, 4 knots	RE	First measurement used	Secondary analysis	Basic set	Main article
							Sensitivity analysis	Extended set	Online appendix
							Sensitivity analysis	Unadjusted	
Additional sensitivity analyses for primary outcomes	Group 1	AChE/Hb	RCM	Cubic splines, 4 knots	RE	If two measurements were made, the second one is used	Sensitivity analysis	Basic set	Online appendix
	Group 1	AChE/Hb	RCM	Assuming linearity	RE + bootstrap	First measurement used			

	Group 1	AChE/Hb	RCM	Cubic splines, 3 knots	RE	First measurement used			
	Group 1	AChE/Hb	RCM	Cubic splines, 5 knots	RE	First measurement used			
	Group 1 (calculated using stricter quality criteria: difference in best and second-best FEV1 and FVC must be ≤ 0.15l, or ≤ 0.1l if best FVC ≤ 1l)	AChE/Hb	RCM	Cubic splines, 4 knots	RE	First measurement used			
	Group 1	AChE/Hb (calculated using adjusted Hb to account for measurement errors)	RCM	Cubic splines, 4 knots	RE	First measurement used			
	Group 1 (excluding anyone with a self-reported prior diagnosis of tuberculosis, and a participant with goiter)	AChE/Hb	RCM	Cubic splines, 4 knots	RE	First measurement used			
	Group 1 (excluding all observations where the AChE gave a warning that delays had happened during analysis)	AChE/Hb	RCM	Cubic splines, 4 knots	RE	First measurement used			

	Group 1 (phase 1+2)	AChE/Hb	FE	Cubic splines, 4 knots	RE	First measurement used			
	Group 1 (phase 2+3)	AChE/Hb	FE	Cubic splines, 4 knots	RE	First measurement used			

RCM = random coefficient model. FE = fixed effect model. RE = random effect term. AChE = acetylcholine esterase. Hb = hemoglobin.

* Handling of multiple AChE measurements in same phase: Each participant had his/her AChE measured in each phase. In some cases, the primary investigator suspected that an error had occurred during analysis (e.g., due to very low or very high measured hemoglobin values), and a second measurement was therefore made. Both results were saved. The decision to re-do the AChE analysis or not may have been biased unintentionally. Therefore, in the primary analyses we will always use the first measurement, as measurement errors are assumed to happen at random. In some sensitivity analyses, we will instead use the result from the second analysis.

2.6 Definitions of covariate sets

Outcome group	**Random coefficient model**		**Fixed effect model**		**Comment**
	Basic covariate set	**Extended covariate set**	**Basic covariate set**	**Extended covariate set**	
1 (primary outcome)	Age (continuous) Sex (dichotomous) Spirometer used (categorical) Pack-years of smoking (continuous) Cumulated lifetime hours of cooking (proxy for exposure to biofuel smoke, continuous)	Basic set + BMI (continuous) Years of full-time education (proxy for socioeconomic status, continuous)	Δage (continuous) Spirometers used (categorical) Δpack-years (continuous) Hours of cooking in the last week (proxy for exposure to biofuel smoke, continuous)	Minimal set + ΔBMI (continuous)	Sex and education level (as proxy for socio-economic status) is not included in the fixed effect model, as they are assumed constant. Height is implicitly included in the models, as height is used for Z-score calculation.
2 (secondary outcomes)	Basic set for group 1 + height (continuous)	Basic set + BMI Years of full-time education	N/A	N/A	For FEF_{25-75} Z-score, height is not included in either of the covariate sets.

Our previously published analysis protocol[1] for validation of questionnaire data on pesticide exposure included considerations on confounders, based on Directed Acyclic Graphs (DAGs). One DAG suggested that one could not obtain an unbiased estimate of the effect of AChE on pulmonary function by covariate adjustment (as closing one backdoor path by adjustment would unblock other backdoor paths). The DAG has now been simplified to show which the most important confounders to adjust for are. We have made the following simplifications during the revision:

1) The only parental factors causally influencing the offspring are genetics and socioeconomic status.
2) Organophosphate insecticide exposure only influences pulmonary function mediated by AChE.

While assumption 2 might be an oversimplification, it is deemed acceptable. In the PEXADU project, we are not interested in the effect of AChE on health *per se*, but rather in the effects of organophosphate insecticide exposure. If exposure to organophosphate insecticides influences pulmonary health through other mechanisms than AChE, the association between AChE and pulmonary health might be biased away from the null if these other mechanisms are ignored. However, in this case AChE can be considered a proxy variable for all biochemical effects of organophosphate exposure.

The simplified DAG can be found in "Appendix A: Directed Acyclic Graph for AChE vs. pulmonary function". To decide which potential confounders to adjust for, we first specified that we wanted to adjust for sex, age and which specific spirometer was used during examination. We then let the DAGitty software suggest a complete set of covariates, the inclusion of which allowed estimation of the causal effect of AChE on lung function. This is the "basic set" of confounders for RCM described above. The "extended set" of confounders for RCM was created by adding BMI and socioeconomic changes to the basic set. While our DAG did not suggest that it was necessary to adjust for these two factors, adjusting for them did not open any backdoor paths. The DAG showed that adjusting for parents' history of lung disease (asthma or COPD) would open backdoor paths and lead to bias, which is why it is not included in either set of covariates.

3 Analysis plan for glycemic regulation

3.1 Descriptive statistics

Before more advanced statistical analyses, we will present the following table of demographic information:

Characteristic	**All participants**	**Conventional farmer's group**	**Semi-organic farmer's group**
Total n			
Sex			
Male, n (%)			
Female, n (%)			
Age in years: Median, IQR			
Educational level (years of full-time schooling): Median, IQR)			
Alcohol consumption in the last week (grams): Median, IQR			
Ethnicity			
Baganda, n (%)			
Other, n (%)			
Body mass index (kg/m2): Median, IQR			
Fruit consumption in last week (number of servings)			
Vegetable consumption in last week (number of servings)			

Next, we will present the following table of biochemical results:

	Phase 1	Phase 2	Phase 3
HbA1c, mmol/mol: Median, IQR HbA1c, %NGSP: Median, IQR n			
FPG: Median, IQR n			
AChE, U/ml: Median, IQR n missing			
Hb, g/dL: Median, IQR n missing			
AChE/Hb, U/g: Median, IQR n missing			

3.2 Outcomes and covariates that will be taken into consideration

The primary outcome is glycosylated hemoglobin A (HbA1c), measured in mmol/mol. Fasting plasma glucose (FPG) is considered secondary.

Depending on the statistical model (more details below), a priori we have decided that we will consider the following variables possible confounders for glycemic regulation: Age, sex, alcohol consumption, physical activity level, consumption of fruit and vegetables, tobacco smoking, socioeconomic status, and BMI. As mentioned in section 2.2, we will not take ethnicity into account.

3.3 Description of the statistical models

As for pulmonary function, results will be analyzed in RCM and FE models, as described in section 2.4 above.

3.4 Handling of HbA1c <4% NGSP

The device used to measure HBA_{1c} in the PEXADU project (HemoCue HbA1c 501) has a lower limit of quantitation (LLOQ) of 4% NGSP, and an upper limit of quantitation of 14% NGSP. Outside of the quantifiable range, the device on only reports that measurements are “<4%” or “>14%”. In the PEXADU population, no values of “>14%” were seen, but 3% of observations were “<4%”. As we primarily want to analyze HbA1c as a

continuous metric, we will have to impute values for the observations listed as "<4%". We will assign the same value to all observations "<4%".

According to the National Glycohemoglobin Standardization Program (NGSP, http://www.ngsp.org/convert1.asp), an HbA1c = 3% NGSP is equivalent to an estimated average glucose of 2.2 mmol/l. We will therefore assume that the lower limit of physiological HbA1c values is 3% NGSP, and we will further assume that values <4% NGSP follow a triangular distribution between 3% NGSP and 4% NGSP. Hence, the imputed value (in % NGSP) will be

$$z = 3 + \frac{LLOQ - 3}{\sqrt{2}} = 3 + \frac{4 - 3}{\sqrt{2}} = 3.71$$

or, equivalently in mmol/mol

$$z = 9 + \frac{LLOQ - 9}{\sqrt{2}} = 9 + \frac{20 - 9}{\sqrt{2}} = 16.8$$

3.5 Overview of statistical analyses

Outcome metrics	Exposure metric	Statistical model	Assumptions regarding linearity	Handling of interdependent data (family)	Handling of multiple AChE and HbA1c measurement in same phase*	Handling of HbA1c < 4% NGSP	Classification	Covariate set	Reported where?
HbA1c (continuous)	AChE/Hb	RCM	Cubic splines, 4 knots	RE	First measurement used	Imputed	Primary analysis	Basic set	Main article
							Primary analysis	Unadjusted	
							Sensitivity analysis	Extended set	
HbA1c (continuous), phase 1+3	AChE/Hb	FE	Cubic splines, 4 knots	RE	First measurement used	Imputed	Secondary analysis	Basic set	Main article
							Secondary analysis	Unadjusted	
							Sensitivity analysis	Extended set	
FPG (continuous)	AChE/Hb	RCM	Cubic splines, 4 knots	RE	First measurement used	N/A	Secondary analysis	Basic set	Main article
							Sensitivity analysis	Extended set	Online appendix
							Sensitivity analysis	Unadjusted	
HbA1c (continuous)	AChE/Hb	RCM	Cubic splines, 4 knots	RE	First measurement used	Excluded	Sensitivity analysis	Basic set	Online appendix
HbA1c (continuous)	AChE/Hb	RCM	Cubic splines, 4 knots	RE	If two measurements were made,	Imputed			Online appendix
									Online appendix

					the second one is used				Online appendix
HbA1c (continuous)	AChE/Hb	RCM	Assuming linearity	RE + bootstrap	First measurement used	Imputed			Online appendix
HbA1c dichotomized into normal (≤38 mmol/mol) vs. raised (≥ 39 mmol/l)	AChE/Hb	RCM (logistic)	Cubic splines, 4 knots	RE	First measurement used	Imputed			Online appendix
HbA1c (continuous)	AChE/Hb	RCM	Cubic splines, 3 knots	RE	First measurement used	Imputed			
HbA1c (continuous)	AChE/Hb	RCM	Cubic splines, 5 knots	RE	First measurement used	Imputed			
HbA1c (continuous)	AChE/Hb (calculated using adjusted Hb to account for measurement errors)	RCM	Cubic splines, 4 knots	RE	First measurement used	Imputed			
HbA1c (continuous)	AChE/Hb	RCM	Cubic splines, 4 knots	RE	First measurement used	Imputed	Sensitivity analysis	Extended set + Hb	
FPG (continuous) (excluding all observations where the AChE gave a warning that delays had happened during	AChE/Hb	RCM	Cubic splines, 4 knots	RE	First measurement used	N/A	Sensitivity analysis	Basic set	

analysis, or temperature at FPG analysis > 27 °C)									
HbA1c (continuous) (excluding all observations where the AChE gave a warning that delays had happened)	AChE/Hb	RCM	Cubic splines, 4 knots	RE	First measurement used	Imputed			
HbA1c (continuous), phase 1+2	AChE/Hb	FE	Cubic splines, 4 knots	RE	First measurement used	Imputed			
HbA1c (continuous) , phase 2+3	AChE/Hb	FE	Cubic splines, 4 knots	RE	First measurement used	Imputed			

RCM = random coefficient model. FE = fixed effect model. RE = random effect term. AChE = acetylcholine esterase. Hb = hemoglobin.

* Handling of multiple AChE and HbA_{1c} measurements in same phase: Please see explanation in section 0 above.

3.6 Definitions of covariate sets

Random coefficient model		**Fixed effect model**		
Basic covariate set	**Extended covariate set**	**Basic covariate set**	**Full covariate set**	**Comment**
Age (continuous) Sex (dichotomous) Alcohol consumption (grams of alcohol in the last week, continuous) MET-minutes per week of physical activity (continuous) Servings of fruit and vegetables consumed per week (continuous) Tobacco smoking (grams of tobacco per day in the last week, continuous)	Basic set + BMI (continuous) Years of full-time education (proxy for socioeconomic status, continuous)	Δage Δ(alcohol consumption) Δ(MET-minutes) Δ(consumption of fruit and vegetables) Δ(tobacco smoking)	Minimal set + ΔBMI	Sex and education level (as proxy for socioeconomic status) is not included in the fixed effect model, as they are assumed constant.

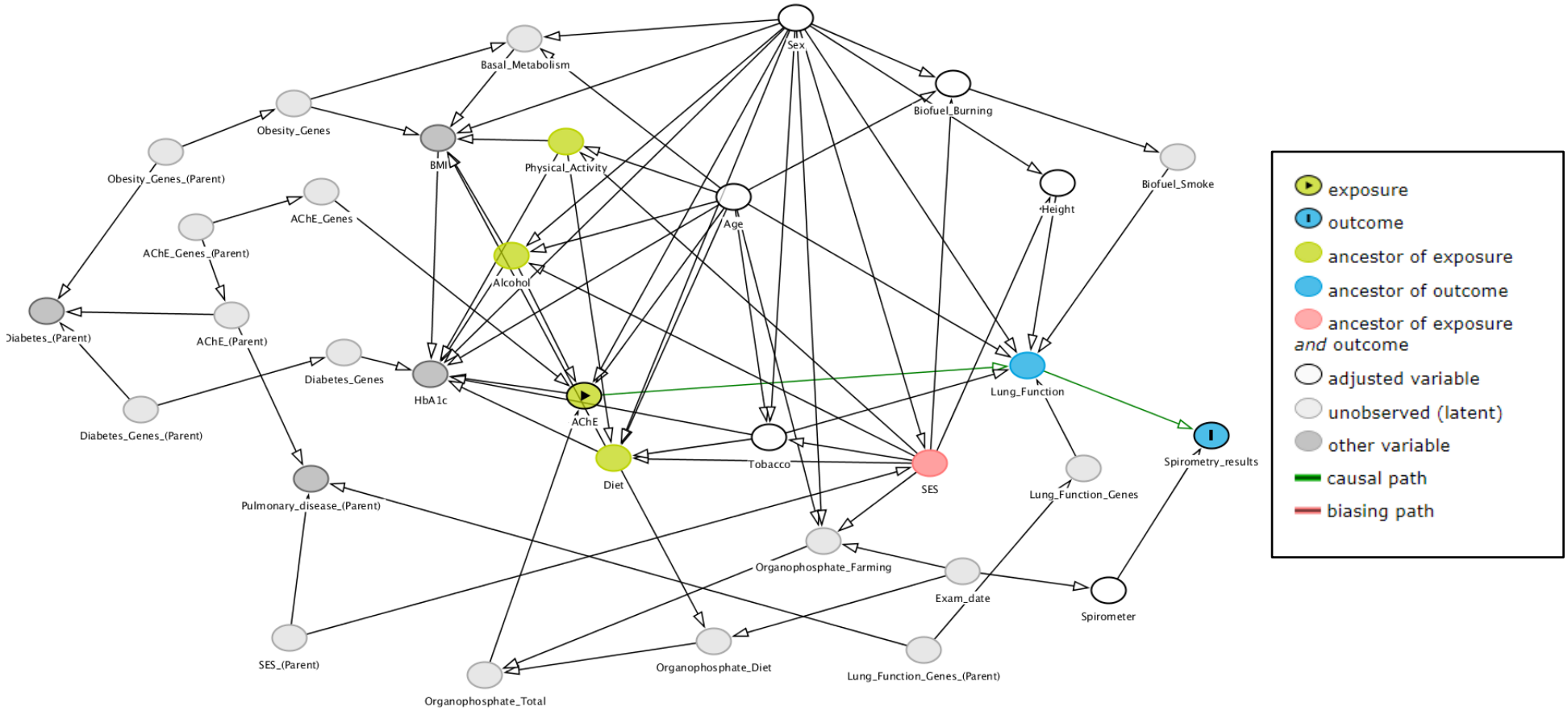
The covariate sets were determined as described for pulmonary function in section 2.6. The DAG used for the analysis is shown in "Appendix B: Directed Acyclic Graph for AChE vs. HbA_{1c}". We first specified that we wanted to adjust our analyses for age and sex. The DAGitty software suggested additional covariates to adjust for to obtain an unbiased estimate of the effect of AChE on HbA_{1c}, and this set is the "basic set" for the RCM shown above. The "extended set" of covariates were created by adding BMI and socioeconomic status as confounders, and confirming in DAGitty that adjusting for these variables would not open backdoor paths and introduce bias. Our DAG showed that we should not adjust for parents' history of diabetes mellitus, as this would open a backdoor path and introduce bias in our results.

4 Definitions of independent variables

The independent variables mentioned in the text above will be defined and modelled in the following manner:

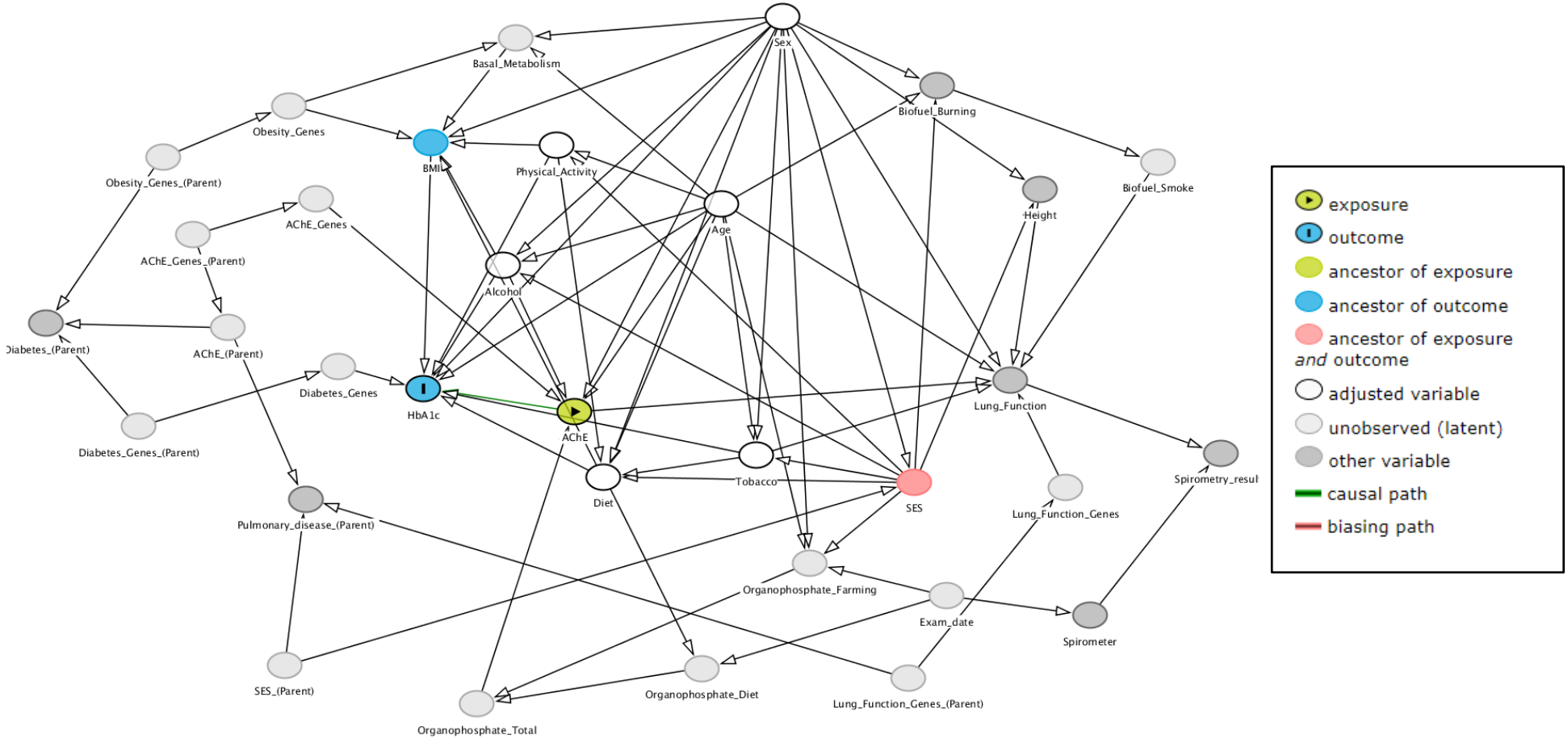
- AChE/Hb: Hemoglobin-adjusted acetylcholine esterase activity. Will be included as continuous variable, modelled using restricted cubic splines with four knots, the location of which will be decided by the distribution of values.
- Body height / stature: Defined as the average body height at all examinations for a specific participant. We use this approach instead of using the measurements in individual phases directly, in attempt to account for measuring errors. The majority of the PEXADU population are women, and many Ugandan women have hairstyles (braiding, hair extensions) that may make it difficult to measure height accurately. Modelled as continuous variables, modelled using restricted cubic splines with four knots.
- Body mass index (BMI): Defined as $BMI = \frac{w}{h^2}$, where w is body weight in kilograms and *h* is body height in meters. Will be included as continuous variable, modelled using restricted cubic splines with four knots.
- Sex: Dichotomous variable.
- Age: Included as continuous variable. Modelled using restricted cubic splines with four knots.
- Educational level (years of full-time education completed): Included as continuous variable. Modelled using restricted cubic splines with four knots.
- Biofuel smoke: The vast majority of participants reported that their households use wood or charcoal for cooking. Exposure to biofuel smoke will be expressed as hours of cooking carried out by the participant (continuous), modelled using restricted cubic splines with four knots.
    - Cumulated exposure (used in RCM): Lifetime hours of cooking.
    - Current exposure (used in FE model): Hours of cooking in the week before examination.
- Diet: Modelled as the sum of servings of fruit and vegetables consumed in the week before examination. Continuous variable, modelled using restricted cubic splines with four knots.
- Physical activity level: MET-minutes in the last week. Continuous variable, modelled using restricted cubic splines with four knots.
- Tobacco smoking: Continuous metrics that will be modelled in a linear fashion, as the number of ever-smokers is too low to allow the use of splines.
    - Cumulated exposure (used in RCM): Pack-years of smoking.
    - Current exposure (used in FE model): Grams of tobacco smoked per day in the week before examination.
- Alcohol consumption: Grams of alcohol consumed in the last week before examination. Continuous variable, modelled using restricted cubic splines with four knots.

5 Appendix A: Directed Acyclic Graph for AChE vs. pulmonary function



The DAG was drawn and analyzed using the DAGitty[4] software, freely available from dagitty.net

6 Appendix B: Directed Acyclic Graph for AChE vs. HbA_{1c}



The DAG was drawn and analyzed using the DAGitty[4] software, freely available from dagitty.net

References

1 Hansen MRH, Jørs E, Sandbæk A et al. Protocol for statistical analyses in the study entitled "Pesticide exposure, asthma and diabetes in Uganda (PEXADU)". *Zenodo* 2019.

2 Armstrong RA. When to use the Bonferroni correction. *Ophthalmic Physiol Opt* 2014;34:502-8.

3 Miller MR, Hankinson J, Brusasco V et al. Standardisation of spirometry. *European Respiratory Journal* 2005;26:319-338.

4 Textor J, van der Zander B, Gilthorpe MS, Liskiewicz M, Ellison GT. Robust causal inference using directed acyclic graphs: the R package 'dagitty'. *Int J Epidemiol* 2016;45:1887-1894.