ehp ENVIRONMENTAL HEALTH PERSPECTIVES
http://www.ehponline.org

DNA Methylation Score as a Biomarker in Newborns for Sustained Maternal Smoking during Pregnancy

Sarah E. Reese, Shanshan Zhao, Michael C. Wu, Bonnie R. Joubert, Christine L. Parr, Siri E. Håberg, Per Magne Ueland, Roy M. Nilsen, Øivind Midttun, Stein Emil Vollset, Shyamal D. Peddada, Wenche Nystad, and Stephanie J. London

http://dx.doi.org/10.1289/EHP333

Received: 22 October 2015
Revised: 8 April 2016
Accepted: 26 May 2016
Published: 21 June 2016

Environ Health Perspect DOI: 10.1289/EHP333
Advance Publication: Not Copyedited

DNA Methylation Score as a Biomarker in Newborns for Sustained Maternal Smoking during Pregnancy

Sarah E. Reese[1], Shanshan Zhao[2], Michael C. Wu[3], Bonnie R. Joubert[4], Christine L. Parr[1,5], Siri E. Håberg[6], Per Magne Ueland[8,9], Roy M. Nilsen[10,11], Øivind Midttun[12], Stein Emil Vollset[7,10], Shyamal D. Peddada[2], Wenche Nystad[5], and Stephanie J. London[1*]

[1] Epidemiology Branch, National Institute of Environmental Health Sciences, Research Triangle Park, NC, USA

[2] Biostatistics and Computational Biology Branch, National Institute of Environmental Health Sciences, Research Triangle Park, NC, USA

[3] Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA, USA

[4] Population Health Branch, National Institute of Environmental Health Sciences, Research Triangle Park, NC, USA

[5] Department of Chronic Diseases, Norwegian Institute of Public Health, Oslo, Norway

[6] Department of Management and Staff, Norwegian Institute of Public Health, Oslo, Norway

[7] Center for Disease Burden, Norwegian Institute of Public Health, Oslo/Bergen, Norway

[8] Department of Clinical Science, University of Bergen, Bergen, Norway

[9] Laboratory of Clinical Biochemistry, Haukeland University Hospital, Bergen, Norway

[10] Department of Global Public Health and Primary Care, University of Bergen, Bergen, Norway

[11] Center for Clinical Research, Haukeland University Hospital, Bergen, Norway

[12] Bevital A/S, Bergen, Norway

[*]Corresponding Author: Stephanie J. London, NIEHS, PO Box 12233, 111 T.W. Alexander Drive, Building 101, Research Triangle Park, NC 27709. Phone: 919-541-5772; Fax: 919-541-2511; E-mail: london2@niehs.nih.gov

Short running title: Biomarker in Newborns of *in utero* Smoke Exposure

Acknowledgements: We are grateful to all the participating families in Norway who take part in this ongoing cohort study. ØM is employed by Bevital A/S, Bergen, Norway.

Grant information: This work was supported in part by the Intramural Research Program of NIH, NIEHS. The Norwegian Mother and Child Cohort Study is supported by the Norwegian Ministry of Health and the Ministry of Education and Research, NIH/NIEHS (contract no N01-ES-75558), NIH/NINDS (grant no.1 UO1 NS 047537-01), and the Norwegian Research Council/FUGE (grant no. 151918/S10) and the present project by the Norwegian Research Council/BIOBANK (grant no 221097).

Competing financial interests declaration: There are no competing financial interests.

Abstract

Background: Maternal smoking during pregnancy, especially when sustained, leads to numerous adverse health outcomes in offspring. Pregnant women disproportionately underreport smoking and smokers tend to have lower follow-up rates to repeat questionnaires. Missing, incomplete, or inaccurate data on presence and duration of smoking in pregnancy impairs identification of novel health effects and limits adjustment for smoking in studies of other pregnancy exposures. An objective biomarker in newborns of maternal smoking during pregnancy would be valuable.

Objectives: To develop a biomarker of sustained maternal smoking in pregnancy using common DNA methylation platforms.

Methods: Using a dimension reduction method, we developed and tested a numeric score in newborns to reflect sustained maternal smoking in pregnancy from data on cotinine, a short-term smoking biomarker measured mid-pregnancy, and Illumina450K cord blood DNA methylation from newborns in the Norwegian Mother and Child Cohort Study (MoBa).

Results: This score reliably predicted smoking status in the training set (N=1,057; accuracy=96%, sensitivity=80%, specificity=98%). Sensitivity (58%) was predictably lower in the much smaller test set (N=221), but accuracy (91%) and specificity (97%) remained high. Reduced birth weight, a well-known impact of maternal smoking, was as strongly related to the score as to cotinine. A three site score had lower, but acceptable, performance ($accuracy_{train}$=82%, $accuracy_{test}$=83%).

Conclusions: Our smoking methylation score represents a promising novel biomarker of sustained maternal smoking during pregnancy easily calculated with Illumina450K or

IlluminaEPIC data. It may help identify novel health impacts and improve adjustment for smoking when studying other risk factors with more subtle effects.

Introduction

Despite years of health warnings and cessation campaigns, smoking during pregnancy remains an important public health problem (Murin et al. 2011). Women who smoke during pregnancy are more likely to have children with lower birth weight, preterm delivery, reduced lung function, asthma, attention deficit hyperactivity disorder (ADHD), orofacial clefts and other malformations (US Department of Health and Human Services 2014). Emerging evidence links additional health outcomes in children to maternal smoking (Mund et al. 2013). Because of the consistent and important effects of maternal smoking on child health, it is crucial to carefully adjust for smoking when investigating effects of other *in utero* environmental exposures that may have more subtle effects.

Various newborn adverse health outcomes related to maternal smoking, including reduced birth weight, have been shown to be mitigated by cessation (US Department of Health and Human Services 2014), suggesting that sustained smoking during pregnancy rather than simply any smoking during pregnancy is the important parameter to assess in epidemiologic studies. Using a genome-wide approach (Illumina HumanMethylation450 BeadChip, henceforth Illumina450K), Joubert et al. (2012) reported that maternal smoking during pregnancy is associated with differential DNA methylation in newborns at specific loci that replicated in a second population. Subsequent reports have consistently confirmed and extended these findings (Joubert et al. 2016). Joubert et al. (2014) subsequently reported that the DNA methylation signals observed in newborns reflect sustained smoking, defined by cotinine measured at about 18 weeks of gestation, rather than transient smoking; they were not seen when women quit earlier in pregnancy.

Smoking during pregnancy is generally assessed in epidemiologic studies by questionnaires. Studies vary in the number of time points at which smoking information is collected and, even when complete histories across pregnancy are sought, missing questionnaire data at one or more time points decreases sample size for assessment of sustained smoking. Smokers tend to have lower response rates to follow-up questionnaires (Jacobsen and Thelle 1988). While a positive self-report of smoking is reliable, pregnant women are more likely to underreport smoking compared to women of the same age who are not pregnant (Dietz et al. 2011; Kvalvik et al. 2012), likely due to the well-known negative impacts of this exposure on the child. Cotinine is the best biomarker of smoking status available (Benowitz 1996; Kvalvik et al. 2012), but has a half-life of only 17 hours in non-pregnant women (Benowitz 1996) and 9 hours in pregnant women (Dempsey et al. 2002). There have been recent attempts to develop biomarkers of long-term smoking exposure in adults using the Illumina450K platform (Shenker et al. 2013; Zhang et al. 2015). However, this has not been done in newborns to reflect exposure to maternal smoking during pregnancy.

The goal of this paper is to develop, using the Illumina450K methylation platform, a biomarker in newborns of sustained smoking by the mother during pregnancy that can be easily applied to other newborn studies with either Illumina450K or Illumina Infinium MethylationEPIC BeadChip (henceforth IlluminaEPIC) methylation data. A biomarker of this nature will be useful in studies of childhood health outcomes to fill in the inevitable missing data on whether or for how long a mother smoked, when limited data were collected on timing of smoking, and to validate self-reports of non-smoking. While statistical methods exist to fill in missing data, such as multiple imputation, these are inferior to a direct and objective biomarker. Further these

methods involve assumptions about the random nature of the missing data (Sterne et al. 2009) that are unlikely to hold for smoking, especially during pregnancy. We used two existing datasets with Illumina450K methylation measured in newborns and cotinine measured in maternal plasma during pregnancy to develop and test a methylation score to predict smoking. We also examined the association between the resulting methylation score and reduced birth weight, a well established consequence of maternal smoking during pregnancy (US Department of Health and Human Services 2014).

Methods

Study Population

The Norwegian Mother and Child Cohort Study (MoBa) is a large population-based pregnancy study conducted by the Norwegian Institute of Public Health targeting all women who gave birth in Norway from 1999 to 2008 (Magnus et al. 2006; Ronningen et al. 2006). Blood samples were obtained from the mother during pregnancy and from newborns (cord blood). Here we analyzed a subcohort of MoBa participants (born 2002-2004) with Illumina450K methylation data measured from newborn DNA and cotinine measured from maternal plasma at about gestational week 18 of pregnancy (Joubert et al. 2012). The Illumina450K methylation data were generated in two different analytic batches: MoBa1 (N=1,068, generated in 2011) and MoBa2 (N=222, generated in 2013). We used the first dataset (MoBa1) analyzed by Joubert et al. (2012), as our training set. The second dataset (MoBa2) served as our test dataset.

Exposure to nicotine from sources other than cigarette smoking could be reflected in cotinine levels, but are not expected to generate the same methylation signals (Besingi and Johansson

2014); therefore, we excluded the 10 subjects from the training set and one subject from the test set who reported use during pregnancy of snuff/chewing tobacco, nicotine gum, nicotine patch, or nicotine inhaler. One additional subject was excluded from the training set due to missing cotinine data. This left us with 1,057 subjects in the training set and 221 in the test set for analyses.

The MoBa study has been approved by the Regional Committee for Ethics in Medical Research, the Norwegian Data Inspectorate, and the Institutional Review Board of the National Institute of Environmental Health Sciences. Written informed consent was obtained from all participants.

Laboratory Measurements

Cotinine

Cotinine concentrations were measured in maternal plasma collected at approximately 18 weeks gestation (Kvalvik et al. 2012) using liquid chromatography-tandem mass spectrometry at BEVITAL AS (www.bevital.no) (Midttun et al. 2009).

Methylation Data

We measured DNA methylation in cord blood samples at 485,577 CpG sites using the Infinium HumanMethylation450 BeadChip (Bibikova et al. 2011; Sandoval et al. 2011). Bisulfite conversion was performed using the EZ-96 DNA methylation kit. All quality control and data processing was done as described previously (Joubert et al. 2012). Briefly, samples were omitted if the average detection p-value across all probes was less than 0.05 and/or they were labeled as failed by Illumina, they were identified as a gender outlier, or they were a blind duplicate of another sample included in the dataset. CpGs that were missing chromosome data, were missing

more than 10% of data across samples, or were on chromosome X or Y were omitted. Joubert et al. (2012) found no evidence of batch effects in these data, which were generated over less than four weeks. Beta values, β, were calculated in Illumina's GenomeStudio methylation software as the ratio of the intensity of the methylated allele to the sum of the intensities of the methylated and unmethylated alleles plus a constant. The beta values were additionally logit transformed to obtain the log ratio, $\ln\left(\frac{\beta}{1-\beta}\right)$.

Definition of Sustained Smoking in this Analysis

We used the term "sustained smoking" as in our previous report (Joubert et al. 2014) where we found that the methylation signals were observed in newborns with mothers in this group but not in mothers who quit early in pregnancy. Among the 1,278 pregnancies across the training and test sets, we examined the timing of quitting smoking during pregnancy using questionnaire data collected at two time points in pregnancy (approximately weeks 17 and 30 of gestation). Among these women, 127 reported smoking at the beginning of pregnancy and did not report quitting. Among the 253 who reported quitting during pregnancy, there were 54 who did not report in which week of pregnancy they quit, 184 who reported quitting by 18 weeks, and 15 who reported quitting after 18 weeks. Thus the vast majority of women who reported that they stopped smoking during pregnancy did so by 18 weeks. Our cotinine value measured at about 18 weeks identifies women who are still smokers at this time point. When considered in combination with our questionnaire data, a cotinine value in the active smoking range both reflects smoking into the second trimester, as opposed to smoking that stopped early in pregnancy, and, for the vast majority of women who smoked at the onset of pregnancy, correlates with smoking through most

of pregnancy. We therefore refer to smoking detected by cotinine >56.8 nmol/L (Shaw et al. 2009) at about 18 weeks or self-reported later in pregnancy (17 or 30 weeks) as "sustained" smoking during pregnancy in this analysis.

Cotinine-based Classification of Sustained Smoking During Pregnancy

We refer to the smoking variable based solely on cotinine dichotomized at 56.8 nmol/L as "cotinine-based sustained smoking."

Self-report based Classification of Sustained Smoking During Pregnancy

The "self-reported sustained smoking" variable was created from data from two questionnaires, one administered at about 17 weeks of pregnancy and one administered at about 30 weeks, supplemented with information collected from mothers at birth from the Medical Birth Registry of Norway (MBRN). This variable classifies mothers who reported that they were sometimes or daily smokers as smokers, and mothers who reported that they never smoked, quit before pregnancy, or stopped smoking early in pregnancy as non-smokers.

Combined Classification of Sustained Smoking During Pregnancy

We also created a "combined sustained smoking" variable that classifies mothers based on cotinine levels above 56.8 nmol/L as smokers combined with mothers who self-reported as daily smokers whether or not their cotinine value exceeded this threshold. This "combined sustained smoking" variable reclassified as smokers ten individuals in the training set and one in the test set who were nonsmokers according to the "cotinine-based sustained smoking" variable.

Statistical Methods

Development of Smoking Biomarker on Training Data

We performed a genome-wide robust linear regression (Fox and Weisberg 2011) on the training set (MoBa1) using the "combined sustained smoking" variable as the dichotomous predictor and the log ratios of the DNA methylation data as the response variable. These were non-normalized as in Joubert et al. (2012) so as to closely replicate these results. The top 200 most significant CpGs were selected, consistent with the sure independent screening approach suggested by Fan and Lv (2008). We then cross-referenced the 200 CpGs with lists of potentially problematic probes (Chen et al. 2013), including those that have single nucleotide polymorphisms nearby. We visually inspected the distributions of all CpGs that overlapped with these lists and removed 5 CpGs with non-unimodal distributions from further analysis. The remaining 195 CpGs were used in the logistic least absolute shrinkage and selection operator (LASSO) model to choose a set of CpGs for use in the calculation of the smoking score (Hastie et al. 2009; Tibshirani 1996).

We used the untransformed methylation beta values as the predictors of maternal smoking because it has become more common to analyze Illumina450K data on the natural scale. In previous studies, results of classification methods were not significantly different when using beta values versus log ratios for large sample sizes (Zhuang et al. 2012). To account for the randomness of the LASSO procedure (Hastie et al. 2009; Tibshirani 1996), we performed it 100 times. After running the 100 iterations, we selected the subset of CpGs that appeared in all 100 to choose a robust subset of CpGs that might be more applicable to other studies. A smoking score was then calculated as the linear combination of the subset of CpGs and the logistic LASSO regression coefficients.

Receiver operating characteristic (ROC) analysis (Metz 1978) was used to establish a threshold, based on the logistic LASSO regression coefficients, for the smoking methylation score to classify newborns according to exposure to a mother with sustained smoking during pregnancy using the combined variable described above. We set the threshold to minimize the sum of false positives and false negatives with the restriction that the sensitivity had to be at least 80%. False positives are subjects misclassified as offspring of smoking mothers, whose mothers did not smoke according to their combined self-report and cotinine measurements. False negatives are subjects misclassified as offspring of mothers who did not smoke, who appear to smoke based on their combined self-report and cotinine values. We calculated the area under the curve (AUC) and used the threshold to classify samples and to calculate the accuracy, sensitivity, and specificity.

Validation of Smoking Biomarker on Test Data

Using the same logistic LASSO regression coefficients and threshold value, we calculated the smoking methylation score for the test set (MoBa2) and performed ROC analysis at the threshold established above to calculate the accuracy, sensitivity, and specificity.

Comparing Different Smoking Variables to Train the Score

Several additional analyses were performed to assess how the LASSO regression results changed when using other smoking variables to train the model rather than combined sustained smoking. We focused on combined sustained smoking because, although cotinine is an objective measure and the best available biomarker of smoking, it is relatively short term. Most pregnant women in our study do not smoke heavily and many do not smoke daily. Thus, if a woman refrained from smoking on the day of her clinic visit when blood for cotinine was drawn, the value might be in

the non-smoking range. Because pregnant women are exceedingly unlikely to claim to be smokers when they are not, it seems imprudent to overwrite a positive self-report of smoking because of a cotinine value below our cutoff. In addition to this primary smoking variable ("combined sustained smoking"), we trained our model using two additional smoking variables: "cotinine-based sustained smoking," which is based only on the cotinine measurement and "self-reported sustained smoking," based only on questionnaire data.

We performed an additional sensitivity analysis using a naïve CpG selection approach including only the three loci replicated at strict Bonferroni significance in Joubert et al. (2012) to form the smoking methylation score. This approach used the most significant CpG from each of these three loci (*AHRR*, *GFI1*, and *CYP1A1*) from our genome-wide analysis and the corresponding robust linear regression coefficients to compute the smoking methylation score.

Illumina recently released the EPIC BeadChip which covers over 850K CpG sites (Moran et al. 2016). Approximately 42,000 of the Illumina450K CpGs are not included on IlluminaEPIC. Because we do not have IlluminaEPIC data, we assessed the performance of the score trained on the Ilumina450K data after deleting CpGs that do not overlap between the two platforms.

The AUC, accuracy, sensitivity, and specificity were used to evaluate the performance of the methylation score created in these different additional analyses.

Birth Weight in Relation to the Different Smoking Variables

We examined how our methylation score relates to a known newborn health outcome of having a mother who smoked during pregnancy. We chose birth weight because of the well-established inverse association with maternal smoking during pregnancy (US Department of Health and

Human Services 2014). We performed a linear regression analysis to compare the association between birth weight and various smoking variables: sustained smoking based on our newly created smoking methylation score, cotinine-based sustained smoking, self-reported sustained smoking, combined sustained smoking (using both self-report and cotinine), and a self-report variable for *any* (yes or no) smoking during the pregnancy whether sustained or not. We appreciate that there is some circularity because we developed the score in the training portion of the data using the combined sustained smoking variable as the gold standard.

The birth weight variable came from the Medical Birth Registry of Norway (MBRN) (Irgens 2000). Covariates included in all birth weight models were gender, gestational age, maternal education, maternal age, parity, and the selection variable for the data set. We also created a crude model without the smoking variable for comparison.

We assessed fit of the birth weight models using likelihood ratio tests (LRT) comparing models including a smoking variable to the crude model. We used root mean square error (RMSE) to assess how well each model estimated birth weight. The smaller the RMSE the better the model estimated birth weight.

All analyses were performed in R version 3.0.2 (R Core Team 2013) (Packages used: glmnet, pROC, MASS, sandwich, and lmtest).

Results

The percentage of mothers positive for combined sustained smoking during pregnancy was similar in the training and test sets (Training: 13.0%; Test: 14.0%; p-value=0.34; Table 1).

Among these smokers, the amount smoked was low (median=5 cigarettes per day) in both the training and test sets (Table 1).

The iterative logistic LASSO AUC cross-validation procedure, a procedure to choose the CpGs most predictive of combined sustained smoking, identified 28 CpGs retained in all 100 runs in the training set (Supplementary Table S1). As expected, there was substantial overlap of the CpGs on this list and those reported by Joubert et al. (2012) – 5 of the original 10 loci were identified. The distributions of the calculated smoking methylation score for the training set by levels of our combined sustained smoking variable are displayed in Supplementary Figure S1A. In the ROC analysis for the training set (N=1,057), the smoking methylation score compared well to the combined sustained smoking variable (AUC=0.96; 95% Confidence Interval (CI)=[0.95, 0.98]; Supplementary Figure S2). The resulting threshold value for the smoking methylation score was -0.37 with an accuracy of 96%, sensitivity of 80% and specificity of 98% (Table 2 model c). At this threshold, there were 19 (1.8%) false positives (non-smokers that were classified as smokers) and 27 (2.6%) false negatives in the training set.

For the test set (N=221) the AUC was 0.90 (CI=[0.83, 0.97]; Supplementary Figure S2), using the same regression coefficients from the LASSO to calculate the smoking methylation score (Supplementary Figure S1B) and the same threshold value for the ROC analysis. As expected, the performance of the smoking methylation score was not as high in this much smaller test set (Table 2 model c): sensitivity was reduced to 58%, although accuracy (91%) and specificity (97%) were only slightly lower than in the training set. In the test set there were 6 (2.7%) false positives and 13 (5.9%) false negatives.

Additional Analyses

As expected, cotinine-based sustained smoking and self-reported sustained smoking differed slightly (Supplementary Table S2; phi coefficient=0.79 (Training set) and 0.81 (Test set)). Therefore, we compared our main analysis ("combined sustained smoking," Table 2 model c and Supplementary Table S1) to models where we trained the smoking methylation score using the "cotinine-based sustained smoking" variable (Table 2 model a and Supplementary Table S3) or separately, the "self-reported sustained smoking" variable (Table 2 model b and Supplementary Table S4). Table 2 shows the number of CpGs (q) used to calculate the smoking methylation score and the results of the ROC analysis for the smoking methylation scores calculated using the three different smoking variables. The predictive ability of the smoking methylation score was best when trained on the combined sustained smoking. As expected, in all models the sensitivity in the smaller test set was substantially reduced compared to the larger training set. The specificity remained high, only slightly reduced, for the test set compared to the training set.

The naïve approach using only the three replicated CpGs does not predict smoking status as reliably as the LASSO model trained on combined sustained smoking and resulted in lower sensitivity and considerably lower specificity in both the training and test sets (Table 2 model d) although it had acceptable performance (training set AUC=0.89, test set AUC=0.82).

Only two of the 28 CpGs identified in the combined sustained smoking score are not included in the IlluminaEPIC array (cg00709966 and cg11864574). Leaving these two CpGs out made very little difference in the performance of the score (Supplementary Tables S5 and S6).

Birth Weight Analysis

Using linear regression models, we compared the association between birth weight and smoking, classified variously as exposed based on the smoking methylation score (12.7% prevalence), cotinine-based sustained smoking (12.2%), self-reported sustained smoking (11.6%), combined sustained smoking (13.2%), and an additional variable for self-report of any smoking during pregnancy whether sustained or not (yes versus no; yes = 28.3%; Supplementary Table S7). Supplementary Tables S7 and S8 give descriptive statistics in the training data for the smoking variables and covariates included in the models. Table 3 shows the resulting coefficients and standard errors from the linear regression models, the Akaike information criterion (AIC), log-likelihood, and p-value resulting from the likelihood ratio test to the crude model. The RMSE did not distinguish much between models (range 444.06-445.62). This is not surprising given that maternal smoking leads only to a modest decrement in birth weight, and thus, is not its major determinant; in these data the maximum percent of variation explained was 33.2% (range 32.6%-33.2%). Although the differences were miniscule, the sustained smoking models all performed significantly better than the crude model (Table 3) whereas the any smoking variable did not perform better than the crude model.

Discussion

We developed a novel biomarker in newborns of sustained maternal smoking in pregnancy using methylation values in newborns from the Illumina450K platform. This biomarker is a smoking methylation score that incorporates the subset of 28 CpGs we found to be most predictive of maternal smoking status from a logistic LASSO model. The sensitivity was high in the training

set but lower, as expected, in the much smaller separate test set; however, the specificity remained high in both. When we evaluated the relationship with reduced birth weight, a well-established health effect of maternal smoking, we found that our smoking methylation biomarker performs about the same as the cotinine-based sustained smoking, self-reported sustained smoking, combined sustained smoking incorporating self-report and cotinine, and substantially better than self-report of any smoking in pregnancy.

The score that we developed is intended for studies with Illumina450K methylation data. For studies with the new IlluminaEPIC array, the score can be directly applied using the CpGs from our score that overlap with those on the IlluminaEPIC array with little loss of performance. Our work also allows comparison with a naïve method based not on any dimension reduction method but simply on three replicated top loci from Joubert et al. (2012). Interestingly, this naïve three CpG score performed relatively well given how little epigenetic information was included (training accuracy 82% versus 96% from the LASSO). For studies without Illumina450K or IlluminaEPIC data, this score could be implemented by assessing methylation at these three loci using pyrosequencing or other methods (Roessler and Lehmann 2015; Wani and Aldape 2016; Wiencke et al. 2014).

Previous studies have developed biomarkers of smoking in adults from methylation data. Shenker et al. (2013) developed a methylation index based on a linear combination of methylation values of four CpGs and the coefficients from their genome-wide analysis. Zhang et al. (2015) developed a biomarker based on two CpGs that were strongly associated with all-cause, cardiovascular, and cancer mortality. Philibert et al. (2015) investigated the use of five

CpGs as potential indicators of smoking for use in clinical settings. We developed a biomarker of sustained smoking in pregnancy using genome-wide data, which retained a larger number of CpGs (q=28). While there are several dimension reduction methods to choose from, we chose LASSO because it generally selects a more parsimonious set of features and it is difficult to show a significant difference in performance between the methods (Hastie et al. 2009). This smaller set of CpGs expected to be selected by the LASSO allows the smoking methylation score to be more easily implemented in other studies.

Recent studies have shown that many of the smoking methylation signals seen in newborns persist into childhood. For example, the three CpGs in our naïve score are also related to sustained maternal smoking during pregnancy in several studies of older children, but the effects are attenuated with the passage of time (Küpers et al. 2015; Ladd-Acosta et al. 2016; Lee et al. 2015; Richmond et al. 2015).

The smoking methylation score provides studies that lack cotinine values or have incomplete self-reported smoking histories with an easy to calculate, objective biomarker in newborns of having a mother who smoked during most of the pregnancy as well as a validation of self-reported non-smoking. It can be used to fill in missing data on smoking or its timing throughout pregnancy. A biomarker is superior to statistical methods to fill in missing data, such as multiple imputation. Our score is simple to compute in other newborn datasets with Illumina450K or IlluminaEPIC methylation data to generate a biomarker in newborns of sustained smoking in pregnancy. The score is a simple linear combination of the methylation values of 28 CpGs and a vector of logistic LASSO regression coefficients, which we have provided in Supplementary

Table S1. It is known that positive self-reports of smoking are reliable but that some smokers may falsely deny smoking. Because of the well-publicized adverse effects of smoking during pregnancy on offspring, pregnant smokers are more likely to deny smoking than other smokers of reproductive age who are not pregnant (Dietz et al. 2011; Kvalvik et al. 2012). Thus in studying effects of maternal smoking in pregnancy on health outcomes in children or adjusting for smoking effects in studies of other risk factors that often have more subtle effects, having an objective biomarker to aid in classification of smoking status is useful.

A biomarker of sustained smoking during pregnancy will also be useful in studies of childhood health outcomes where DNA can be obtained from routinely collected neonatal blood spots. Concomitant information on smoking in birth certificates or medical charts is often limited to yes or no during pregnancy and may have large numbers of missing values. Smoking during pregnancy queried several years later when children have had time to develop conditions that are known to be related to parental smoking is subject to biased reporting.

We previously reported that sustained maternal smoking during pregnancy has a much greater effect on newborn methylation than smoking that ceased early in pregnancy (Joubert et al. 2014). Here we show that "sustained smoking during pregnancy" had a greater effect on birth weight than "any smoking during pregnancy" which was not significantly related to birth weight. The smoking methylation score we developed, which reflects sustained rather than any smoking, may better capture health effects of maternal smoking on the newborn as our birth weight analysis suggests.

Given the large and reproducible impact of maternal smoking on the newborn methylome, there is great interest in whether these signals mediate health outcomes causally linked to this exposure, such as reduced birth weight (Küpers et al. 2015). However, regardless of whether they are mediators, these methylation signals are useful biomarkers of *in utero* exposure. The success of this approach for smoking, where methylation signals are abundant, augurs well for the use of the methylation data to develop objective biomarkers of *in utero* exposures that are harder to measure and may have subtler effects on the epigenome and child health outcomes.

We note that the smoking methylation score was developed using data from a homogenous population from Norway. Therefore we do not know how generalizable it would be to other ethnic groups. However, the training and test methylation datasets were generated at different time points in different analytic batches spaced about two years apart. Thus our finding of good performance of the score in the test set incorporates the effects of laboratory variability increasing the applicability to other studies.

To develop the score, we used data that were not normalized (not corrected for the fact that the Illumina450K includes two probe types). We did this both for comparability with our previous publication (Joubert et al. 2012) and to increase generalizability to studies that may not have normalized or used varying normalization procedures. We found that normalizing using the popular BMIQ method (Teschendorff et al. 2013) does not influence the smoking results in our data (Joubert et al. 2014). In addition, Wu et al. (2014), using our data, found that when examining an association with a high level of statistical significance, such as maternal smoking in pregnancy, results using raw versus normalized data are very similar. In addition, we did not

batch correct the test and training sets which were analyzed at different points in time. We did this to better approximate how the score will behave in other studies to increase generalizability of our results. For investigators who might want to normalize to our data, we provide the mean methylation values for the set of CpGs used in the score in our Supplementary Materials (Supplementary Table S9).

As a supplemental analysis, we performed the LASSO method using the log ratios, rather than the untransformed methylation beta values, and the model performance was virtually identical (Training: untransformed accuracy=0.96 vs. log ratio accuracy=0.95; Test: untransformed accuracy=0.91 vs. log ratio accuracy=0.91; see Supplementary Table S6), however it retained more CpGs (37 vs. 28). A score with fewer elements is easier to use, but for users who prefer to analyze their data on the log ratio scale we provide a supplementary table with the 37 CpGs and their coefficients (Supplementary Table S10).

We refer to our primary exposure metric, based on the combination of a positive self report and cotinine measured in samples taken at approximately 18 weeks, as "sustained smoking" because most women who reported that they had smoked in early pregnancy but quit later, had done so by 18 weeks. However, to determine sustained smoking, it would have been better to have measured cotinine again near the end of pregnancy.

A limitation in developing a methylation score biomarker of sustained smoking during pregnancy is that there is no clear gold standard. Cotinine is only a reliable biomarker of recent smoking. We primarily used cotinine to train the model (since only a few cotinine-based non-smokers were switched to smokers based on self-report) and thus our score cannot perform better

than cotinine. This removes our ability to discern whether the methylation score is truly superior to cotinine, a short-term biomarker, in predicting health effects of sustained maternal smoking on birth weight or other outcomes.

Conclusions

We have developed a novel biomarker in the newborn of exposure to sustained maternal smoking during pregnancy using Illumina450K DNA methylation data. This methylation score is an objective biomarker that reflects much longer-term exposure than cotinine, the best available smoking biomarker. The score can be easily implemented in other studies with similar methylation data. It provides a means to validate self-reported non-smoking status during pregnancy and enables the ascertainment of sustained smoking when limited time course information was collected. This biomarker of sustained smoking during pregnancy should facilitate better adjustment for maternal smoking in studies of other *in utero* exposures with more subtle effects and may improve the ability to capture novel health effects caused by this important prenatal exposure.

References

Benowitz NL. 1996. Cotinine as a biomarker of environmental tobacco smoke exposure. Epidemiol Rev 18:188-204.

Besingi W, Johansson A. 2014. Smoke-related DNA methylation changes in the etiology of human disease. Hum Mol Genet 23:2290-2297.

Bibikova M, Barnes B, Tsan C, Ho V, Klotzle B, Le JM, et al. 2011. High density DNA methylation array with single CpG site resolution. Genomics 98:288-295.

Chen YA, Lemire M, Choufani S, Butcher DT, Grafodatskaya D, Zanke BW, et al. 2013. Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. Epigenetics 8:203-209.

Dempsey D, Jacob P, 3rd, Benowitz NL. 2002. Accelerated metabolism of nicotine and cotinine in pregnant smokers. J Pharmacol Exp Ther 301:594-598.

Dietz PM, Homa D, England LJ, Burley K, Tong VT, Dube SR, et al. 2011. Estimates of nondisclosure of cigarette smoking among pregnant and nonpregnant women of reproductive age in the United States. Am J Epidemiol 173:355-359.

Fan J, Lv J. 2008. Sure independence screening for ultrahigh dimensional feature space. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 70:849-911.

Fox J, Weisberg S. 2011. Robust regression in R. In: An R companion to applied regression, Part 2nd Ed. Thousand Oaks, CA:Sage.

Hastie T, Tibshirani R, Friedman J. 2009. The elements of statistical learning: Data mining, inference, and prediction:Springer.

Irgens LM. 2000. The medical birth registry of Norway. Epidemiological research and surveillance throughout 30 years. Acta Obstet Gynecol Scand 79:435-439.

Jacobsen BK, Thelle DS. 1988. The tromso heart study: Responders and non-responders to a health questionnaire, do they differ? Scand J Soc Med 16:101-104.

Joubert BR, Haberg SE, Nilsen RM, Wang X, Vollset SE, Murphy SK, et al. 2012. 450K epigenome-wide scan identifies differential DNA methylation in newborns related to maternal smoking during pregnancy. Environ Health Perspect 120:1425-1431.

Joubert BR, Haberg SE, Bell DA, Nilsen RM, Vollset SE, Midttun Ø, et al. 2014. Maternal smoking and DNA methylation in newborns: In utero effect or epigenetic inheritance? Cancer Epidemiol Biomarkers Prev 23:1007-1017.

Joubert Bonnie R, Felix Janine F, Yousefi P, Bakulski Kelly M, Just Allan C, Breton C, et al. 2016. DNA methylation in newborns and maternal smoking in pregnancy: Genome-wide consortium meta-analysis. The American Journal of Human Genetics.

Küpers LK, Xu X, Jankipersadsing SA, Vaez A, la Bastide-van Gemert S, Scholtens S, et al. 2015. DNA methylation mediates the effect of maternal smoking during pregnancy on birthweight of the offspring. Int J Epidemiol.

Kvalvik LG, Nilsen RM, Skjaerven R, Vollset SE, Midttun Ø, Ueland PM, et al. 2012. Self-reported smoking status and plasma cotinine concentrations among pregnant women in the Norwegian Mother and Child Cohort Study. Pediatr Res 72:101-107.

Ladd-Acosta C, Shu C, Lee BK, Gidaya N, Singer A, Schieve LA, et al. 2016. Presence of an epigenetic signature of prenatal cigarette smoke exposure in childhood. Environ Res 144:139-148.

Lee KW, Richmond R, Hu P, French L, Shin J, Bourdon C, et al. 2015. Prenatal exposure to maternal cigarette smoking and DNA methylation: Epigenome-wide association in a

discovery sample of adolescents and replication in an independent cohort at birth through 17 years of age. Environ Health Perspect 123:193-199.

Magnus P, Irgens LM, Haug K, Nystad W, Skjaerven R, Stoltenberg C, et al. 2006. Cohort profile: The Norwegian Mother and Child Cohort Study (MoBa). Int J Epidemiol 35:1146-1150.

Metz CE. 1978. Basic principles of ROC analysis. Semin Nucl Med 8:283-298.

Midttun Ø, Hustad S, Ueland PM. 2009. Quantitative profiling of biomarkers related to B-vitamin status, tryptophan metabolism and inflammation in human plasma by liquid chromatography/tandem mass spectrometry. Rapid Commun Mass Spectrom 23:1371-1379.

Moran S, Arribas C, Esteller M. 2016. Validation of a DNA methylation microarray for 850,000 CpG sites of the human genome enriched in enhancer sequences. Epigenomics 8:389-399.

Mund M, Louwen F, Klingelhoefer D, Gerber A. 2013. Smoking and pregnancy--a review on the first major environmental risk factor of the unborn. Int J Environ Res Public Health 10:6485-6499.

Murin S, Rafii R, Bilello K. 2011. Smoking and smoking cessation in pregnancy. Clin Chest Med 32:75-91, viii.

Philibert R, Hollenbeck N, Andersen E, Osborn T, Gerrard M, Gibbons FX, et al. 2015. A quantitative epigenetic approach for the assessment of cigarette consumption. Front Psychol 6:656.

R Core Team. 2013. R: A language and environment for statistical computing. Part 3.0.2. Vienna, Austria:R Foundation for Statistical Computing.

Richmond RC, Simpkin AJ, Woodward G, Gaunt TR, Lyttleton O, McArdle WL, et al. 2015. Prenatal exposure to maternal smoking and offspring DNA methylation across the lifecourse: Findings from the Avon longitudinal study of parents and children (ALSPAC). Hum Mol Genet 24:2201-2217.

Roessler J, Lehmann U. 2015. Quantitative DNA methylation analysis by pyrosequencing(R). Methods Mol Biol 1315:175-188.

Ronningen KS, Paltiel L, Meltzer HM, Nordhagen R, Lie KK, Hovengen R, et al. 2006. The biobank of the Norwegian Mother and Child Cohort Study: A resource for the next 100 years. Eur J Epidemiol 21:619-625.

Sandoval J, Heyn H, Moran S, Serra-Musach J, Pujana MA, Bibikova M, et al. 2011. Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome. Epigenetics 6:692-702.

Shenker NS, Ueland PM, Polidoro S, van Veldhoven K, Ricceri F, Brown R, et al. 2013. DNA methylation as a long-term biomarker of exposure to tobacco smoke. Epidemiology 24:712-716.

Sterne JA, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, et al. 2009. Multiple imputation for missing data in epidemiological and clinical research: Potential and pitfalls. BMJ 338:b2393.

Teschendorff AE, Marabita F, Lechner M, Bartlett T, Tegner J, Gomez-Cabrero D, et al. 2013. A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450K DNA methylation data. Bioinformatics 29:189-196.

Tibshirani R. 1996. Regression shrinkage and selection via the LASSO. J Roy Stat Soc B Met 58:267-288.

US Department of Health and Human Services. 2014. The health consequences of smoking: 50 years of progress. A report of the surgeon general. Atlanta, GA: US Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health 17.

Wani K, Aldape KD. 2016. PCR techniques in characterizing DNA methylation. Methods Mol Biol 1392:177-186.

Wiencke JK, Bracci PM, Hsuang G, Zheng S, Hansen H, Wrensch MR, et al. 2014. A comparison of DNA methylation specific droplet digital PCR (ddPCR) and real time qPCR with flow cytometry in characterizing human T cells in peripheral blood. Epigenetics 9:1360-1365.

Wu MC, Joubert BR, Kuan PF, Haberg SE, Nystad W, Peddada SD, et al. 2014. A systematic assessment of normalization approaches for the Infinium 450K methylation platform. Epigenetics 9:318-329.

Zhang Y, Schottker B, Florath I, Stock C, Butterbach K, Holleczek B, et al. 2015. Smoking-associated DNA methylation biomarkers and their predictive value for all-cause and cardiovascular mortality. Environ Health Perspect.

Zhuang J, Widschwendter M, Teschendorff AE. 2012. A comparison of feature selection and classification methods in DNA methylation studies using the Illumina Infinium platform. BMC Bioinformatics 13:59.

Table 1: Descriptive statistics of sustained smoking variables, cotinine, and quantity smoked

Variable	Category	Training (MoBa1; N=1,057)	Test (MoBa2; N=221)
Cotinine-based sustained smoking; n (%)[a]	No	930 (88.0)	191 (86.4)
	Yes	127 (12.0)	30 (13.6)
Self-reported sustained smoking; n (%)	No	936 (88.6)	191 (86.4)
	Yes	121 (11.4)	30 (13.6)
Combined sustained smoking; n (%)	No	920 (87.0)	190 (86.0)
	Yes	137 (13.0)	31 (14.0)
Cotinine values by sustained smoking category[b] (mean ± SD)	No	0.7 ± 3.4	0.7± 2.2
	Yes	424± 337	497 ± 301
Number of cigarettes per day (among smokers[b]; median (IQR))		5 (2-10)	5 (3-7)

[a] Based on cotinine measured in maternal plasma collected at about 18 weeks of pregnancy, values >56.8 nmol/L classified as Yes.

[b] Based on the combined sustained smoking variable.

Table 2: Logistic LASSO results for main and additional analyses – The number of CpGs (*q*) used to calculate the smoking methylation score, area under the curve (AUC) and 95% confidence interval (CI), smoking methylation score threshold, accuracy and CI, sensitivity and CI, specificity and CI, and number and percentage of false negatives (FN) and false positives (FP).

Model			q	AUC (CI)	Threshold	Accuracy (CI)	Sensitivity (CI)	Specificity (CI)	FN (%)	FP (%)
a	Cotinine-based Sustained Smoking	Training[c]	24	0.97 (0.95,0.99)	-9.09	0.95 (0.94,0.97)	0.83 (0.76,0.89)	0.97 (0.96,0.98)	22 (2.1)	27 (2.6)
		Test[c]		0.88 (0.80,0.96)		0.90 (0.86,0.93)	0.63 (0.47,0.80)	0.94 (0.90,0.97)	11 (5.0)	12 (5.4)
b	Self-reported Sustained Smoking	Training[c]	12	0.93 (0.90,0.96)	-11.71	0.92 (0.90,0.94)	0.81 (0.74,0.88)	0.93 (0.92,0.95)	23 (2.2)	64 (6.0)
		Test[c]		0.82 (0.74,0.91)		0.90 (0.86,0.93)	0.47 (0.30,0.63)	0.96 (0.94,0.99)	16 (7.2)	7 (3.2)
c	Combined Sustained Smoking[a]	Training[c]	28	0.96 (0.95,0.98)	-0.37	0.96 (0.94,0.97)	0.80 (0.74,0.87)	0.98 (0.97,0.99)	27 (2.6)	19 (1.8)
		Test[c]		0.90 (0.83,0.97)		0.91 (0.88,0.95)	0.58 (0.39,0.74)	0.97 (0.94,0.99)	13 (5.9)	6 (2.7)
d	Naïve CpG selection[b]	Training[c]	3	0.89 (0.86,0.92)	-0.47	0.82 (0.80,0.84)	0.81 (0.74,0.87)	0.82 (0.80,0.85)	24 (2.3)	166 (15.7)
		Test[c]		0.82 (0.73,0.91)		0.83 (0.78,0.88)	0.60 (0.43,0.77)	0.87 (0.82,0.92)	12 (5.4)	25 (11.3)

[a] In the combined sustained smoking variable, a woman's positive report of daily smoking during pregnancy overrides a cotinine value of ≤56.8 nmol/L used to classify a woman as a non-smoker in the cotinine-based sustained smoking variable.

[b] Naïve CpG selection refers to the smoking score calculated using the three CpGs from the loci replicated at strict Bonferroni significance in Joubert et al. (2012). These CpGs and corresponding coefficients are cg05575921 (-0.558; *AHRR*), cg14179389 (-0.555; *GFI1*), cg18092474 (0.205; *CYP1A1*).

[c] Training N=1,057; Test N=221.

Table 3: Birth weight regression analysis results on the training data (N=1,039) – Coefficient, standard error (SE), and linear model p-value (P_{LM}) for each model with a smoking variable, and Akaike information criterion (AIC), log likelihood (logL), likelihood-ratio test to the crude model, p-value (P_{LRT}), and root mean squared error (RMSE) for each model[a].

Model	Coefficient	SE	P_{LM}[b]	AIC	logL	P_{LRT}[c]	RMSE
Crude	NA	NA	NA	15645.5	-7812.8	NA	446.11
Methylation Score Class	-130.9	42.81	0.0023	15638.1	-7808.1	0.0022	444.10
Cotinine-based Sustained Smoking	-133.3	44.08	0.0026	15638.3	-7808.2	0.0024	444.14
Self-reported Sustained Smoking	-120.4	44.95	0.0075	15640.3	-7809.2	0.0072	444.56
Combined Sustained Smoking[b]	-131.4	42.62	0.0021	15638.0	-7808.0	0.0020	444.06
Self-reported Any Smoking	-48.2	32.18	0.1348	15645.3	-7811.6	0.1329	445.62

[a] All models adjusted for gender, gestational age, maternal education, maternal age, parity, and selection.

[b] In the combined sustained smoking variable, a woman's self-report of daily smoking during pregnancy overrides a cotinine value of ≤56.8 nmol/L used to classify a woman as a non-smoker in the cotinine-based sustained smoking variable.

[c] p-values<0.05 were considered to represent statistical significance.