PEER REVIEW HISTORY

BMJ Open publishes all reviews undertaken for accepted manuscripts. Reviewers are asked to complete a checklist review form (**http://bmjopen.bmj.com/site/about/resources/checklist.pdf**) and are provided with free text boxes to elaborate on their assessment. These free text comments are reproduced below.

ARTICLE DETAILS

TITLE (PROVISIONAL)	Cohort Profile: A Prospective Longitudinal Study of the Pregnancy DNA Methylome—United States Pregnancy, Race, Environment, Genes (PREG) Study
AUTHORS	Lapato, Dana; Wagner, Sara; Olivares, Emily; Amstadter, A; Kinser, Patricia; Latendresse, Shawn; Jackson-Cook, Colleen; Roberson-Nay, Roxann; Strauss, Jerome; York, Timothy

VERSION 1 – REVIEW

REVIEWER	Janine LaSalle University of California, Davis USA
REVIEW RETURNED	13-Nov-2017

GENERAL COMMENTS	This manuscript describes the study design and sample collection protocols for a longitudinal study designed to look at socioeconomic and environmental contributors to the higher rates of preterm birth in African American compared to Caucasian births in which a portion of the variability can be explained by a combination of non-genetic risk factors. The blood and cord blood samples collected from this longitudinal study design are currently being analyzed for DNA methylation and gene expression patterns by Illumina (450k and EPIC) arrays, but those molecular data are not included in this manuscript. There are an extensive number of questionnaires collected on the participants designed to test specific hypotheses about the complex multi-factorial components of racial disparities, making this study a potentially very useful one for getting at biological underpinnings of racial disparity and preterm birth. However, there are a number of limitations and oversimplifications that would need to be addressed. Major concerns 1. Table 2 shown that the AA and EA groups within the cohort are clearly significantly different for a number of factors in additional to gestational age (BMI, education, marital status, smoking, education, income, etc). While the authors argue that these differences are what they are examining in the study, the concern is the degree that these break down by race alone. Since the number of subjects in the study that completed and were able to obtain cord blood samples is low, the concern is that the investigators will not be able to tease apart genetic differences from race from those that are defined by these multiple non-genetic covariates with race. Recruitment of a EA group from lower socioeconomic and AA from higher socioeconomic communities would be important as the study

	progresses in order to be more balanced and to increase the numbers of participants. As it stands, it is unclear how any methylation or expression differences between groups will be interpreted. 2. From a large number of genome-wide DNA methylation sequencing studies, it is clear that the relationship between DNA methylation and gene expression is not as simple as portrayed in this manuscript. It is important for the investigators in this study to be aware of the complexities regarding methylation and expression relationships and not expect that their gene expression data set will completely overlap with the methylation data set or try to exclude those genes that do not show up on both platforms. Also, the assumption that methylation changes precede gene expression changes in flawed. Particularly in early life samples such as cord blood, DNA methylation can be a measure of past gene expression patterns as much as it may predict current or future expression patterns. Also the location of the DNA methylation relative to the gene, promoter, or enhancer is an important consideration in the interpretation. Lastly, the presence and percentage of newborn red blood cells (which have nuclei) in the cord blood samples is an important consideration in the DNA methylation patterns since they have a distinct methylation pattern and are influenced by perinatal risk factors. Minor concerns 1. In Table 2, several of the groupings do not add up to 100%, for instance in EA mothers, 94.3 are married but 0% are unmarried. An "other" or otherwise appropriately labelled category should be included in these cases. 2. Also in Table 2, it was unclear when the prenatal vitamin use was assessed. Before conception or first prenatal visit? This should be included as a footnote. 3. A table of the numbers of samples that were collected from each group at each stage should be included, since there were drop-outs and some samples that were unable to be collected.

REVIEWER	M. Plusquin Hasselt University Agoralaan, building D, 3590 Diepenbeek Belgium
REVIEW RETURNED	19-Nov-2017

GENERAL COMMENTS	The paper describes the PREG cohort that aims to study racial health disparities in perinatal outcomes. The cohort entails a combination of the collection of social determinants and biological measures. The paper is well written and my comments are mainly to provide a better understanding of the paper. The title contains only the abbreviation of the cohort but not the full name, the paper would be clearer if the authors also include "Pregnancy, Race, Environment, Genes study" in the title. Why did the authors exclude mothers older than 40 years of age? The cohort focusses on environmental factors during the pregnancy and although the NLfETy provides information about the neighbourhood including traffic there is no information about

	exposure levels of for example air pollution. Will the authors also consider environmental exposure to contaminants? That exposures are not included in an environmental study may be included as one of the limitations of the study. The part of the study design lacks details on data analytic approaches. And do the authors have a plan for data sharing? The authors should clarify in the title which population table 2 represents (all recruited, prenatal, postpartum). Table 1 could be clarified by adding the number of participants that have these measurements. Does birth refer to cord blood, this should be clarified under the table. Why is telomere length twice in the table? This could perhaps be combined on 1 line if they indicated which samples are maternal of the new-born? The postpartum population includes a selection of the mothers, the authors should provide a comparison between the PREG and MDP populations. The authors should add on table 2 that gestational age is expressed in days. Did the authors actually measured global DNA methylation as reported on page 14? Or do they want to refer to epigenome-wide methylation here? If they measured global methylation they should include the technique used (for example LINE 1, LUMA). The authors should provide a list of abbreviations. Which technique was used for the telomere length measurements? This should be clarified in the text.

REVIEWER	Panagiotis Georgiadis National Hellenic Research Foundation Athens, GREECE
REVIEW RETURNED	22-Nov-2017

GENERAL COMMENTS	The present report is a description of the PREG cohort which focuses on how early life nutritional, environmental and social stressors as well as genetic determinants may affect child health and lead to preterm birth. The study was very well designed, well executed and the current manuscript would be a good reference for future experimental PREG cohort reports. My only concern is that no direct measurement of the 'environmental stressors' is described as part of the study design. The authors should not rely only in questionnaire-derived measurements and some short of actual quantitation of environmental or dietary constituents should be included in the future plans. The cohort is a medium-sized one in terms of the number of recruited participants, however, it is fairly large if we consider all the samples collected during the preterm and post-term gestational periods. The authors decided in favour of an in depth preterm epigenetic and possibly other omics analyses (4 maternal blood samples!) instead of an increase of the number of mothers taking part in the study (page 3 bullet 4). Although I am quite sceptical about this part of the study design, their explanation (page 5 second

	paragraph) is acceptable. The manuscript is quite long and difficult to read. A study flow diagram is necessary and can substitute some of the text

VERSION 1 – AUTHOR RESPONSE

Dear Dr. Sucksmith,

We are grateful to the reviewers for their thoughtful comments and suggestions and have made several changes to the original manuscript as outlined below:

Editorial Requests:

"Please provide a full justification for the sample size of your cohort, including a power calculation."

RESPONSE: Power for the PREG study relies on the theory that small correlations between environmental risk and clinical outcome can be resolved into chains of much larger individual correlations between intervening epigenetic pathways, which can then be estimated with greater reliability using smaller samples than are typical in epidemiological investigations. Using calculations based on samples of 2,000,000 random multivariate normal response vectors, power ranged from 70-90% for testing partial regressions >0.2, even when the direct correlation between a distal covariate and clinical outcome might be too small to be detected reliably (e.g., <0.1). In this way, the repeated measures allow us to characterize the pathways that mediate effects of the environment on PTB that might be too small to be detected individually. A brief description of the rationale for PREG and MDP sample sizes has been added to pages 9 & 10, respectively.

"[I]nclude a copy of the STREGA checklist indicating the page/line numbers of your manuscript where the relevant information relating to the GWAS aspect of your study."

RESPONSE: We have also included a STROBE checklist for cohort profiles instead of a STREGA checklist because the genome-wide association study to identify methylation quantitative trait loci has not been conducted. Therefore, most of the STREGA extension criteria cannot be addressed (e.g., allele calling algorithm used, error rates for allelic variant calls, tests of Hardy-Weinberg equilibrium, number of individuals attempted to be genotyped compared to the total number passing quality control, etc.). Additionally, a footnote has been added to Table 1 clarifying that the GWAS is intended to identified methylation quantitative trait loci.

Reviewer 1

"Since the number of subjects in the study that completed and were able to obtain cord blood samples is low, the concern is that the investigators will not be able to tease apart genetic differences from race from those that are defined by these multiple non-genetic covariates with race…. As it stands, it is unclear how any methylation or expression differences between groups will be interpreted."

RESPONSE: The concern that the degree of demographic differences by race may make some analyses intractable is valid, and any investigations using this data must be mindful of its limitations. PREG is an epidemiological study. Enrollment was based on a few health criteria and self-identified race. No actions were taken to ensure that the African-American and European-American women were matched for any demographic information. This strategy resulted in a representative sample of the city of Richmond, including noticeable demographic differences by race. That said, having multiple time points of data does provide leverage to disentangle some biological and environmental factors that may be influencing gestational age at birth or maternal mood. Moreover, ancestry principal

components can be derived from methylation microarray data and can be leveraged to control for genetic differences between super populations (see Accounting for Population Stratification in DNA Methylation Studies. Barfield et al. (2014)), and preliminary principal component analysis with PREG data shows two distinct groups clustering by self-reported race. These ancestry-relevant principal components will be included as covariates in regression models to control for allelic group differences. We have added this information to the manuscript on page 15.

"[T]he relationship between DNA methylation and gene expression is not as simple as portrayed in this manuscript."

RESPONSE: We amended the text in the conceptual model section (p.6) to emphasize possible feedback mechanisms between GE and DNAm, which is illustrated in the conceptual model figure but was not emphasized in the text.

"It is important for the investigators in this study to be aware of the complexities regarding methylation and expression relationships and not expect that their gene expression data set will completely overlap with the methylation data set or try to exclude those genes that do not show up on both platforms."

RESPONSE: We understand that combining DNAm and gene expression data presents many challenges; indeed, relatively few papers have attempted to harmonize multi-omic data and have instead relied on correlations or looked only at genes with significant results from both platforms. We believe that utilizing data from these two platforms will provide more robust results and that some directionality can be inferred (e.g., DNAm marks present at visits 1 and 2 were not caused by gene expression profiles in visit 3). Additional information regarding the U133 chip has also been added to the manuscript to highlight that both the 450k and U133 probe sets cover a reasonable proportion of the genome and transcriptome and should have considerable overlap in genes assayed (p.15).

"The assumption that methylation changes precede gene expression changes in flawed. Particularly in early life samples such as cord blood, DNA methylation can be a measure of past gene expression patterns as much as it may predict current or future expression patterns."

RESPONSE: As stated above, the conceptual model does not presuppose that all gene expression changes follow methylation changes. In the case of maternal DNAm measurements, the repeated measures will allow some inference of causality, given that gene expression measured at visit 4 in late pregnancy cannot have caused the DNA methylation patterns measured in visit 1. The cord blood analysis will not have the advantage of repeated measures; however, the in utero environment phenotyping for the PREG cord blood samples includes not only a wide range of exposures but also some leverage to estimate the duration of exposure. Timing, type, and duration of exposure have each been speculated to influence health outcomes, and the PREG study includes a dataset capable of assessing the impact of all three.

"The presence and percentage of newborn red blood cells (which have nuclei) in the cord blood samples is an important consideration in the DNA methylation patterns since they have a distinct methylation pattern and are influenced by perinatal risk factors."

RESPONSE: We are aware of algorithms by Gervin et al. (2016) and Bakulski et al. (2016) that estimate cord blood cell proportions and intend to use those to account for cell type heterogeneity in cord blood samples.

"In Table 2, several of the groupings do not add up to 100% and it was unclear when the prenatal vitamin use was assessed."

RESPONSE: Table 2 now includes an "other" category for relationship status, and footnotes have been added to clarify that all demographic information, including prenatal vitamin usage, was assessed at visit 1.

"A table of the numbers of samples that were collected from each group at each stage should be included."

RESPONSE: Table 1 now includes counts of each measure at each visit, and the labels have been reorganized so that they are easier to read.

Reviewer 2

"The paper would be clearer if the authors also include "Pregnancy, Race, Environment, Genes study" in the title."

RESPONSE: We agree and have changed the manuscript title accordingly.

"Why did the authors exclude mothers older than 40 years of age?"

RESPONSE: An explanation for excluding mothers over the age of 40 has been added to the Participant eligibility and recruitment section (p.8).

"[T]here is no information about exposure levels of for example air pollution. Will the authors also consider environmental exposure to contaminants?"

RESPONSE: Regional daily and weekly data on environmental contaminants, such as lead, sulfur dioxide and ozone, have been obtained from the US Environmental Protection Agency & Virginia Department of Environmental Quality archive for 2013-2016. A description of this data has been added to p.18.

"The part of the study design lacks details on data analytic approaches. [D]o the authors have a plan for data sharing?"

RESPONSE: Additional details were not added to the data sharing plans because we felt like the current data sharing statement included sufficient information for potential collaborators interested in using the data. Regarding analysis plans, we elected not to focus on future analysis plans but to describe the rationale for the PREG study and the breadth of data available. Descriptions of the electronic data capture and processing and how each are relevant to data sharing and reproducibility are available in the Transparent Data Processing section (p.19-20).

"The authors should clarify in the title which population table 2 represents (all recruited, prenatal, postpartum)."

RESPONSE: We have clarified in the text (p.13) that Table 2 refers to all participants who met no pregnancy or birth exclusion criteria.

"Table 1 could be clarified by adding the number of participants that have these measurements."

RESPONSE: Table 1 now includes the number of samples and questionnaires collected at each time point.

"The postpartum population includes a selection of the mothers, the authors should provide a comparison between the PREG and MDP populations."

RESPONSE: The reported between group comparisons has been expanded to include prenatal vitamin use, education, household income, unemployment status, relationship status, and student status in addition to gestational age at birth, race, and maternal age which were presented in the manuscript (p.19)

"The authors should add on table 2 that gestational age is expressed in days."

RESPONSE: A footnote has been added to Table 2 clarifying that gestational age at birth is measured in days.

"Did the authors actually measured global DNA methylation as reported on page 14?"

RESPONSE: Global DNAm measurements were not collected, and the word "global" has been changed to "genome-wide" on p.14

"The authors should provide a list of abbreviations."

RESPONSE: A complete list of abbreviations has been added to the end of the manuscript.

"Which technique was used for the telomere length measurements? This should be clarified in the text."

RESPONSE: quantitative PCR was used to measure global telomere lengths. This detail has been added to p.15 in the text.

Reviewer 3

"My only concern is that no direct measurement of the 'environmental stressors' is described as part of the study design. The authors should not rely only in questionnaire-derived measurements and some short of actual quantitation of environmental or dietary constituents should be included in the future plans."

RESPONSE: We acknowledge that additional variables such as neighborhood levels of pollution and laboratory tests for hormone levels and nutritional intake would have been useful; however, resources did not permit collecting those data. That said, a fair amount of non-self-report data is available, including regional levels of certain environmental contaminants and neighborhood environmental ratings in addition to numerous biological indicators. Also, nurse module notes abstracted from clinical visits during gestation and through delivery are available for a majority of the participants and include non-self-report information about substance use, maternal weight gain, and medical tests ordered.

"The manuscript is quite long and difficult to read. A study flow diagram is necessary and can substitute some of the text."

RESPONSE: A study flow diagram has been added as a Supplementary figure per Reviewer 3's suggestion. Additionally, the text in the Introduction and in the Results sections has been reduced (see pages 3, 6, 17).

Again, we appreciate the time and effort each reviewer took to provide helpful feedback.

Best,
Timothy P. York & Dana Lapato

VERSION 2 – REVIEW

REVIEWER	Michelle Plusquin Hasselt University Agoralaan, building D BE-3590 Diepenbeek
REVIEW RETURNED	25-Jan-2018

GENERAL COMMENTS	The authors adequately responded to the remarks. It is not clear whether the study has been registered.