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: The Baependi Heart Study — a family-based, highly admixed cohort study in a rural Brazilian town
AUTHORS	Egan, Kieren; von Schantz, Malcolm; Negrão, André; Santos, Hadassa; Horimoto, Andrea; Gonçalves, Guilherme; Soler, Julia; de Andrade, Mariza; Lorenzi-Filho, Geraldo; Vallada, Homero; Taporoski, Tamara; Pedrazzoli, Mario; Azambuja, Ana; de Oliveira, Camila; Alvim, Rafael; Krieger, José; da Costa Pereira, Alexandre

VERSION 1 - REVIEW

REVIEWER	Ailiana Santosa Centre for Demographic and Ageing Research, Umeå University, Umeå, Sweden
REVIEW RETURNED	18-Mar-2016

GENERAL COMMENTS	Reviewer's comments for author: Title: "Cohort Profile- The Baependi Heart Study — a family-based, highly admixed cohort study in a rural Brazilian town." This manuscript is fascinating, concerning study design on a longitudinal study conducted in rural Brazil that tried to cover admixture CVD lifestyle and heritability of Brazilian people aged 18 years and over. This manuscript gives a brief guide to the different study types using The Baependi Heart Study data. I have some major and minor concerns, detailed below. Major concerns are as follows: 1. This cohort profile concerned about CVD as the outcome by examining lifestyle risk factors (physical and psychological factors) and genetic factors, however the authors did not specify CVD as the result (for example CVD deaths or morbidity conditions). Can you elaborate more about this? 2. The authors did not explicitly describe how they recruited the participants in the baseline and the follow-up (they stated 548 additional participants included in the 2010 follow-up). My concern was about the number of recruited participants. The authors reported the inconsistent number of participants either in abstract, under cohort description and in tables 1, 2 & 3). How much response rate for the baseline since the recruitment of study participants started? 3. While the Bapendi Heart study tried to collect many information on socio-demographic, family relationship, health behaviour, physical measurements, and mental health in each interview, I wonder how long the interview took for each participant (the authors did not

	mention about interview time). My concern is associated with the response rate and attrition rate during the follow-up period. It will be good if the authors would describe clearly about the reason for refusals and in which age group they belong to the refusal and the decease case. 4. The general questionnaire was delivered by a trained technician (page 10 line 41), which differs from what Oliveira et al. described (reference 19) in their article. 5. The authors also did not distinctly define overall information collected (step-wise data collection). I suggest having a distinct table of socio-demographic, lifestyle risk factors, mental status, physical and anthropometry measurements with questions used in their study. In addition, they took leukocyte DNA sample for further analysis. But they did not explain more which might confuse the reader with sudden information like that. Medical history that authors stated on page 11, line 15 did not appear in Table 2-4. How they asked the participants about their medical history. What type information they collected for medical history (the type of disease, the period having disease whether the diseases occurred in the last year/month? etc.). Some information about skin and dermatology-related traits have been mentioned in this manuscript. However, the authors did not clearly describe how they collected the information. 6. The authors stated about the aid given by researcher for those participants who were illiterate. How about other groups such disable people, people with some physical problems – vision, hearing, dementia, etc.). 7. I suggest authors can provide more structural description in resource study and variables used, so the reader will be easy to follow the process of data collection in the Baependi Heart study. Minor concern: 1. The aim of this study design paper was, to characterise the contributions to cardiovascular risk in a highly admixed population where the epidemiological transition is more recent than in other parts of the word (stated in abstract), which is quite different from what authors described under introduction section (to develop a longitudinal family-based cohort study that reflects some of the genetic and lifestyle-related peculiarities of the Brazilian populations in order to help evaluate genetic and environmental influences on CVD risk factor traits). I suggest that the authors can clarify the objective or, at least, make it similar what in the abstract and the manuscript.

REVIEWER	Dr Naval K Vikram Department of Medicine, All India Institute of Medical Sciences, New Delhi-110029, India
REVIEW RETURNED	12-Apr-2016

GENERAL COMMENTS	This is an interesting study. The data that the authors wish to report in this manuscript should be defined clearly both in the abstract section and in the 'Findings to date' section.

REVIEWER	Antonio Luiz Pinho Ribeiro UFMG, Brazil
REVIEW RETURNED	09-May-2016

GENERAL COMMENTS	This is the cohort profile of a family-based cohort, established in a small town in Southeast Brazil, ten years ago. It has a difference in relation to other cohort studies in Brazil, since the recruitment included the selection of a proband and the invitation to participate of all his/her first, second and third relatives. This exquisite structure of the cohort turn it specially well-fitted to genetic studies. The study is very interesting, but I have some questions, comments and suggestions. 1. What is the study hypothesis and the main goal? This should be explicitly stated in the methods section. 2. Considering that heritability is a main issue in this study, please define what is exposure, what is outcome in your study. 3. According to the Brazilian Census of 2010, 72% of the population of Baependi lives in urban areas. If your sample was obtained randomly from the resident population of the municipality, we would expect to have most of the participants from urban areas, not rural ones. Thus, the title of the study may be not correct, as well as some of the considerations inside the paper. Please, clarify this issue and/or correct it. 4. In page 11, there is a description of the variables you studied that is almost exactly the same of what is shown in table 3 – as outcomes. Please, choose if they are variables or outcomes. Moreover, it is not necessary to have the same content in text and tables. 5. Page 12: please add the number of the approval of the project by the ethics committee. 6. Table 1: the percentages do not totalize 100% for age groups and education. Please review. 7. Table 2: the units are not needed and this column can be deleted. 8. Table 3: define all dichotomic variables (for example, high fasting glucose, what was the cut-off?)

VERSION 1 – AUTHOR RESPONSE

Reviewer: 1

This manuscript is fascinating, concerning study design on a longitudinal study conducted in rural Brazil that tried to cover admixture CVD lifestyle and heritability of Brazilian people aged 18 years and over. This manuscript gives a brief guide to the different study types using The Baependi Heart Study data. I have some major and minor concerns, detailed below.

Major concerns are as follows:

1. This cohort profile concerned about CVD as the outcome by examining lifestyle risk factors (physical and psychological factors) and genetic factors, however the authors did not specify CVD as the result (for example CVD deaths or morbidity conditions). Can you elaborate more about this?

Response: We have amended the text (pp 11—12) as follows:

"Cardiovascular events (end points such as myocardial infarction, heart failure, coronary insufficiency) and procedures (such as hospitalizations, surgery, and the need for percutaneous coronary

intervention) are recorded regularly; individuals are followed up annually by telephone contactm and every five years during office visits."

2. The authors did not explicitly describe how they recruited the participants in the baseline and the follow-up (they stated 548 additional participants included in the 2010 follow-up). My concern was about the number of recruited participants. The authors reported the inconsistent number of participants either in abstract, under cohort description and in tables 1, 2 & 3). How much response rate for the baseline since the recruitment of study participants started?

Response: Within the abstract and cohort description, we have used the current number of 2,239 participants. The referee is absolutely correct in that we should not have quoted the number of 1,712 individuals included at baseline from the original paper (de Oliveira et al, 2008), where volunteers with incomplete data had not been excluded. The Abstract has been corrected accordingly:

“At baseline, the study evaluated 1,691 individuals across 95 families. Cross-sectional data have been collected for 2,239 participants”

3. While the Bapendi Heart study tried to collect many information on socio-demographic, family relationship, health behaviour, physical measurements, and mental health in each interview, I wonder how long the interview took for each participant (the authors did not mention about interview time). My concern is associated with the response rate and attrition rate during the follow-up period. It will be good if the authors would describe clearly about the reason for refusals and in which age group they belong to the refusal and the decease case.

Response: The length of the interview was dependent on the questionnaires used and the responses of the candidate- for example an interview describing an extensive medical history would permit the subject to explain in full their medical history without time constraints.

With respect to those lost to follow up, we did not record the reasons why individuals refused follow up. Our current procedure in the study is to work around our participants as much as possible to maximize follow up rate: we follow up on missed appointments and if they have any issues with a particular outcome being recorded we will tailor their individual outcomes recorded to meet their request.

The demographics of participants who had died or refused to participate in first follow up is shown in the table below. This information has now been integrated into Figure 1.

	Died	Refused	Individuals followed up
n	76	316	1,847
Age	71	41	42
Standard deviation	13	16	17

4. The general questionnaire was delivered by a trained technician (page 10 line 41), which differs from what Oliveira et al. described (reference 19) in their article.

Response: In our opinion, our wording "trained technician" overlaps entirely with de Oliveira's terminology “research assistants specifically trained for this task”

5. The authors also did not distinctly define overall information collected (step-wise data collection). I suggest having a distinct table of socio-demographic, lifestyle risk factors, mental status, physical and anthropometry measurements with questions used in their study. In addition, they took leukocyte DNA

sample for further analysis. But they did not explain more which might confuse the reader with sudden information like that. Medical history that authors stated on page 11, line 15 did not appear in Table 2-4. How they asked the participants about their medical history. What type information they collected for medical history (the type of disease, the period having disease whether the diseases occurred in the last year/month? etc.). Some information about skin and dermatology-related traits have been mentioned in this manuscript. However, the authors did not clearly describe how they collected the information.

Response: We agree that some clarification is required. We have altered the text to streamline the depth of information given in the narrative so that the reader can refer to Table 2 as required for further details.

For medical history, we do not unfortunately have the information in codified form, that could be summarised in the way suggested by the reviewer.

With respect to leukocyte DNA, we beg to differ with the referee. In fact, we have explained the purpose of this collection as follows (p 13):

"In addition, leukocyte DNA samples have been collected and used to probe Affymetrix 6.0 arrays, and data imputed using the 1000 Genomes Cosmopolitan Panel. Analysis of genomic ancestry was conducted using the Admixture program. 20"

In response to the comments regarding dermatology related traits, we have attempted to improve the detail regarding such measured, and added the following text to P 14:

"Skin properties of the study participants have also been collected, and a report of significant associations between stratum corneum moisture and sex, age, high sun exposure, and use of sunscreen published. 26"

6. The authors stated about the aid given by researcher for those participants who were illiterate. How about other groups such disable people, people with some physical problems – vision, hearing, dementia, etc.).

Response: We have changed the text as follows (P 12):

"For participants who were illiterate or visually impaired, questionnaires were read in their entirety by the researcher acting as a scribe. Where an individual's capacity was limited to complete a specific question, assistance was given as needed by either the researcher or a caregiver.

Individuals with severe physical or mental handicaps would have been classified as refusals, but as already mentioned, no broken down data were collected categorising refusals."

7. I suggest authors can provide more structural description in resource study and variables used, so the reader will be easy to follow the process of data collection in the Baependi Heart study.

Response: We have amended the "Collaboration" section to read as follows (P 18):

"The database of the project is under continuous development as new data sets are being added. Data will be made available in public repositories as and when required and appropriate, as publications based on data sets are being accepted. The project already involves multiple collaborations, and the investigators welcome new angle of analysis on existing data sets and proposals for new ones. Proposals can be forwarded to Alexandre C. Pereira (e-mail:

alexandre.pereira@incor.usp.br) for discussion at the steering committee meeting."

Minor concern:
1. The aim of this study design paper was, to characterise the contributions to cardiovascular risk in a highly admixed population where the epidemiological transition is more recent than in other parts of the word (stated in abstract), which is quite different from what authors described under introduction section (to develop a longitudinal family-based cohort study that reflects some of the genetic and lifestyle-related peculiarities of the Brazilian populations in order to help evaluate genetic and environmental influences on CVD risk factor traits). I suggest that the authors can clarify the objective or, at least, make it similar what in the abstract and the manuscript.

Response: We agree entirely. The difference was mainly a result of the word count constraints of the Abstract. However, we have now been able to incorporate the same wording by means of minor abbreviations elsewhere. The relevant section of the Abstract now reads:

The Baependi Heart Study was set up in 2005 to develop a longitudinal family-based cohort study that reflects some of the genetic and lifestyle-related peculiarities of the Brazilian populations in order to help evaluate genetic and environmental influences on CVD risk factor traits.

Reviewer: 2

Reviewer Name

Dr Naval K Vikram

Institution and Country

Department of Medicine, All India Institute of Medical Sciences, New Delhi-110029, India

Please state any competing interests or state ‘None declared’:
None declared

Please leave your comments for the authors below
This is an interesting study.
The data that the authors wish to report in this manuscript should be defined clearly both in the abstract section and in the 'Findings to date' section.

Response: We thank the reviewer for his kind assessment. As per our previous comment, the inclusion of data in the Abstract is inevitably constrained by the total word count. The "Findings to date" section summarises the published datasets, with reference to the original publication in each case.

Reviewer: 3

Reviewer Name

Antonio Luiz Pinho Ribeiro

Institution and Country

UFMG, Brazil

Please state any competing interests or state ‘None declared’:
None declared

Please leave your comments for the authors below
This is the cohort profile of a family-based cohort, established in a small town in Southeast Brazil, ten years ago. It has a difference in relation to other cohort studies in Brazil, since the recruitment included the selection of a proband and the invitation to participate of all his/her first, second and third relatives. This exquisite structure of the cohort turn it specially well-fitted to genetic studies.
The study is very interesting, but I have some questions, comments and suggestions.
1. What is the study hypothesis and the main goal? This should be explicitly stated in the methods section.

Response: We have added the following text (p 9):

"The overall goal was to quantify and characterise the inter-individual variation in common cardiovascular risk factors, and disentangle its genetic and environmental components. By collecting and analysing a uniquely wide range of phenotypic and phenomic information, the study aims to serve as a generator of hypotheses for future investigation."

2. Considering that heritability is a main issue in this study, please define what is exposure, what is outcome in your study.

Response: We have added the following to clarify this issue (P 11):

"Described in the following are outcomes (incidence information derived from longitudinal data collection) and exposures (cross-sectional data collected at baseline or at subsequent study visits)."

3. According to the Brazilian Census of 2010, 72% of the population of Baepandi lives in urban areas. If your sample was obtained randomly from the resident population of the municipality, we would expect to have most of the participants from urban areas, not rural ones. Thus, the title of the study may be not correct, as well as some of the considerations inside the paper. Please, clarify this issue and/or correct it.

Response: We fully agree that this data point should have been mentioned, in order to explain the distribution of the population between the urban and the rural zone. We have clarified this as follows (pp 10—11):

"After the proband's first contact, first-degree relatives were invited to participate by phone; these included all living relatives in the municipality of Baependi (the urban zone, where 72% of the population lives4, and the geographically larger rural zone) and beyond."

4. In page 11, there is a description of the variables you studied that is almost exactly the same of what is shown in table 3 – as outcomes. Please, choose if they are variables or outcomes. Moreover, it is not necessary to have the same content in text and tables.

Response: We take the reviewer's point, and the text (now on p 10) has been abbreviated as follows:

"The following types of information was collected (for details of variables and distribution estimates, see Table 3): General lifestyle and health, Cardiovascular health, Biochemistry, Anthropometry, Mental health, and Sleep and circadian rhythms."

5. Page 12: please add the number of the approval of the project by the ethics committee

This has now been added to the manuscript under the “ethics approval section” (now on p 19):

"and was approved by the Ethics committee of the Hospital das Clínicas, University of São Paulo, Brazil (approval number 0494/10)."

6. Table 1: the percentages do not totalize 100% for age groups and education. Please review.

Response: We apologise for this mistake, which has been rectified. We have also updated these information using a higher-quality dataset than was available previously.

7. Table 2: the units are not needed and this column can be deleted.

Response: This has now been amended.

8. Table 3: define all dichotomic variables (for example, high fasting glucose, what was the cut-off?)

Response: This information has now been added.

Again, we are grateful to you and reviewers for the swift and fair review process. We hope the revised manuscript is to your satisfaction, and look forward to hearing from you in due course.

VERSION 2 – REVIEW

REVIEWER	Ailiana Santosa Centrr for Demographic and Ageing Research, Umeå University, Sweden
REVIEW RETURNED	02-Jul-2016

GENERAL COMMENTS	Overall, the authors have revised and improved this manuscript sufficiently. I do not have any further comments. Therefore I think this manuscript is suitable for publication.