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)	Gut microbiome, enteric infections, and child growth across a rural-urban gradient: A prospective cohort study protocol
AUTHORS	Lee, Gwenyth; Eisenberg, Joseph; Uruchima, Jessica; Vasco, Gabriela; Smith, Shanon; Van Engen, Amanda; Victor, Courtney; Reynolds, Elise; MacKay, Rebecca; Jesser, Kelsey; Castro, Nancy; Calvopina, Manuel; Konstantinidis, Konstantinos; Cevallos, William; Trueba, Gabriel; Levy, Karen

VERSION 1 – REVIEW

REVIEWER	Thahir, Andi The University of Sydney, Central Clinical School
REVIEW RETURNED	12-Jan-2021

GENERAL COMMENTS	The reviewer provided a marked copy with additional comments. Please contact the publisher for full details.

REVIEWER	Robertson, Ruairi Blizard Institute, Queen Mary University of London, UK
REVIEW RETURNED	19-Jan-2021

GENERAL COMMENTS	Here, the authors describe a protocol for a study that aims to identify the associations between enteric pathogen carriage, the gut microbiome and child growth in a low-income setting. Overall, this manuscript is well-written and includes the details required in a study protocol manuscript. This study will provide some important insights into the interaction between enteric pathogens, microbiome and growth in children at risk of undernutrition. There is some clarification required regarding the criteria for the case-control sub-study and further detail on specific pathogens that will be assessed. Other comments are largely minor. Major comments 1. The criteria and details of the diarrhoea case-control study require much further clarification. The manuscript states that 200 cases will be selected. Will these be 200 unique participants? It is likely that a large number of participants will have more than one episode of diarrhea. Therefore, which diarrheal episode would be chosen in these circumstances? It would be more suitable to select 200 unique participants rather than including more than one episode from a single participant, to ensure a true n=200. Furthermore, the controls in this sub-study need to be clarified. As there are 480 total children in the larger study, this 'sub-study' will include nearly every child (n=200 cases and n=200 controls). The

	criteria for selecting controls are those carrying the 'same pathogen infection but without diarrhea'. This assumes that, once diarrheal cases are selected, almost every other child will carry the 'same pathogen'. Which pathogens will be considered? Children will likely carry multiple pathogens at different time points. These could contribute to diarrhea at a later time point, so would their presence in an asymptomatic child at an earlier time point still be considered as a 'control' child? There is no detail on whether the pathogen attributed to the diarrheal epsidoe will be identified (although this is recommended). Finally, please elaborate on whether/how these samples will be age-matched. The gut microbiome at 6 months will be quite different to that at 12 months. Minor comments - It would be helpful to provide data/references to stunting/growth/diarrhea metrics in this region in the introduction - Page 7 (of 28), line 15 – this initial sentence is a bit simplified. Infection/pathogen invasion can still occur in 'healthy' microbiome, therefore best to modify this sentence slightly. - Page 7, line 46 – what does 'differential responses to infection' mean? This is important to clarify if it is the primary hypothesis. May be better to state "… the gut microbome mediates the effect of enteric infections on diarrhea, EED and growth in the first 2 years of life". - Page 7, line 47 – Similarly, when stating "…contribute to acute and chronic health outcomes?", probably best to state 'associated with' health outcomes. - Page 7, line 49 – the gut microbiome is "associated with" rather than "modifies" the short and long term health outcomes - Page 8, line 41 – "We will perform subsequent sensitivity analyses to determine how data from these participants should be used." - sensitivity analyses should be clarified here rather than planned post-hoc - Please provide details on how gestational age will be determined (LMP?) as this will be important for eligibility criteria of 37 weeks. - Please provide details of which pathogens will be screened for on Taq cards. This is an important aspect of the study and one that has received little detail in this manuscript. There will likely be carriage of multiple pathogens at a time. It would also be beneficial to attempt to identify the pathogen attributed to the diarrheal episode (as has been reported in MAL-ED studies using the same technology). See major comment regarding this point. - Presumably models will be adjusted for site (urban and rural sites), however this has not been mentioned - Page 12, line 34 – What are meant by "founder community characteristics" and "maturity indices"? No detail is provided

	previously about these statements. Maturity indices presumably refers to MAZ scores. If so, please provide details. - Page 13 line 3 – Similar to previous comments, how will enteric infections be defined if more than one pathogen is present (which is likely)? Simple presence/absence or confirmed pathogen-attributable diarrhea? - Please provide some details on shotgun sequencing methods – sequencing and bioinformatic pipelines etc - Table 3 - The available sample sizes column is unclear? The study has n=480. How are there available samples sizes of >1000 for some models e.g 1a and 1b examining the association between baseline water and microbiome/infection? - Figure 3 - EED fecal biomarkers are only being assessed at 6 months. Therefore how will this be able to meet objectives of aim 2? Is this a mistake in the gantt chart? - There is much emphasis on the importance of the rural-urban gradient in this cohort but very little detail on how this will be considered statistically or interpreted. Will analyses simply be split by the 3 different sites and compared or will there be more complex considerations as to how/why this gradient is influential

VERSION 1 – AUTHOR RESPONSE

	Comment	**Response**
R1.1	First, I would like to thank you for the opportunity to review this manuscript which describes the rationale and methods of the gut microbiome study concerning child growth across Ecuador. The manuscript deals with a very relevant topic for developing countries (stunting), which is appropriate for the Journal scope and interest area. Although the study has clearer originality, the authors need to make more clarity in some areas. My suggestion is to improve the manuscript's clarity, including the time points of data collection and dietary intake questionnaires, before further considering it for publication.	We appreciate the reviewer's useful comments. We assume that the reviewer's comments are those embedded in the PDF, but please let us know if that was incorrect or if any comments were missed. We respond to the reviewer's comments on clarity in our responses to specific comments below.
R1.2	The abstract is well-written summarizing the problem and methods. However, it is not clear what statistical analysis will be used either. Please include a short sentence for this.	We have added further detail to the methods: "**Methods and analysis:** We are conducting a community-based birth cohort study to

		examine interactions between gut microbiome conditions and enteric infection, and how environmental conditions affect the development of the gut microbiome. We will follow 360 newborns from three sites along a rural-urban gradient in northern coastal Ecuador, **characterizing** enteric infections and gut microbial communities in the children every three to six months over their first two years of life. **We will use longitudinal regression models to assess the correlation between environmental conditions and gut microbiome diversity and presence of specific taxa,** controlling for factors that are known to be associated with the gut microbiome**, such as diet**. From six to 12 months of age **we will collect weekly stool samples to compare** microbiome conditions **in diarrhea cases versus controls** prior to, during, and after acute enteric infections**, using principal-coordinate analysis and other multivariate statistical methods**." (LINES 40-53).
R1.3	(6 to twelve months of age in the abstract). Meaning that there will be 24 fecal samples collected for each individual? If so, are these samples for determination infant gut microbiome? While in the figure 3, gut microbiome from six to 12 months were measured only twice. Please be clear.	The reviewer is correct, 24 weekly samples will be collected from 6 to 12 months of age. We have clarified this in the legend of Figure 3: **"Figure 3 legend:** Primary study activities are programmed according to the age of the child. **Squares indicate a single sampling point, except for intensive samples, which are**

		collected weekly from 6-12 months of age (a total of 24 samples), and diarrheal symptoms, which are collected weekly for the entire period of the study." (LINES 754-757).
R1.4	The introduction has been well-written providing compact review literatures that explain background and the gaps. However, maternal factors have been insufficiently clarified. As far as I understood the authors have also observed maternal aspects (and probably will be going to do some analysis). I therefore thought that explaining maternal factors that may be relevant to the study would be better linking the whole study, pregnancy (the starting point for data collection), EED, and child nutritional status.	Although the relationship between maternal factors and the infant microbiome is not the focus of our study, we acknowledge the importance of these relationships and the need to adjust for maternal factors in analysis. We have modified the introduction and methods of the manuscript in several locations to clarify this. Introduction: "These patterns in bacterial community diversity may be driven by **the maternal diet and microbiome[23], infant dietary** habits and nutrition[24,25], exposure to animals[26] and chemicals[27], or other lifestyle factors[28–30]. **Prenatal maternal weight, diet, antibiotic usage, have all been shown to impact the infant gut microbiota**.[31]" (LINES 99-102). Methods: "We also comprehensively evaluate environmental conditions that are associated with enteric pathogen exposure and assess other factors that are known to be

		associated with the **infant** gut microbiome, **including the maternal microbiome in late pregnancy**, cesarean versus vaginal delivery[42], **infant** diet[24,25], and **infant** nutritional status[43]." (LINES 145-149)
R.1.5	(comment on text: "whether the gut microbiome modifies the short- and long-term health outcomes associated with enteric pathogen infections (i.e., diarrhea, EED, and child growth)") In some literature, malnutrition and gut microbiome have bi-directional relationship. For instance, not only does gut microbiome associate with malnutrition, but also malnutrition may affect gut microbiome. Do you agree? How do you address this in the point 2?	We agree with the reviewer about this point. Further in the text, we make the point that longitudinal data are particularly critical to disentangling these bidirectional relationships. We have modified the statement to include malnutrition: "Longitudinal data are needed to characterize conditions prior to infection, and to compare those data to the commensal gut microbiome response following infection, enabling us to better disaggregate cause and effect for interactions between infection and the microbiome. **Similarly, longitudinal data are useful in identifying whether alterations to the microbiome precede, accompany, or follow from, EED and growth faltering."** (LINES 121-125).
R1.6	(comment on text: "Women are recruited and enrolled in late pregnancy, from 37-weeks onward.") The number of studies correlate mother and her offspring gut microbiome. Did the author measure the maternal gut microbiome in this study? If not, why?	Yes, we do measure the maternal microbiome. We have clarified this in the methods section. "**Maternal stool is collected once, at 37 weeks pregnancy**." (LINE 221)

R1.7	Microbiome	We have corrected the misspelling (“microbioem")
R1.8	(comment on text: “…high-risk pregnancy according to Ministry of Health guidelines” I agree that high-risk pregnancy is such a criteria that should be avoided and therefore they should be excluded. However, I see the justification may be more due to it may affect mother's health rather than directly affect gut microbiome of the infant. I meant, maternal factors are important to include in order to justify aim and objectives.	Our primary rationale for excluding high-risk pregnancies was that high-risk pregnancies are more likely to result in unplanned cesarean sections, we have clarified this in the text: “study exclusion criteria (**Table 1**) include planned cesarean section, or high-risk pregnancy according to Ministry of Health guidelines**, as this is a risk factor for unplanned cesarean section**.” (LINES 154-165).
R1.9	(comment on section “Withdrawal”) It is very good point here where the authors clearly explain on how the participants can withdraw themselves and in what criteria they then be categorizing dropped out from the study. Just a thought, there is missing information about what benefit the participants may get from participating this study. Can it be possible to include here?	We have added the following just prior to the “Withdrawal” section: “**Benefits of study participation include child growth monitoring, the timely identification of parasitic infections and anemia, for which appropriate treatment is coordinated through the Ministry of Health, and information about household water quality from environmental testing. We also provide mothers with general health advice. Study incentives include objects such as soap, baby oil, and small toys, of approximately USD 3-5 dollars, distributed every 3 to 6 months**.” (LINES 180-184).

R1.10	(Comment on "Within seven days of birth, information about delivery mode (vaginal versus cesarean), delivery location, and breastfeeding initiation is captured, a child stool sample is collected, and child anthropometry is measured. Further stool and serum samples, nutrition data, and environmental assessments are then collected") Information such as pregnancy outcome of birth weight and gestational age, in some literature, is associated with the study outcomes (growth and microbiome), why they are not collected in this study?	We also collect these variables and have modified the section accordingly: "Within seven days of birth, information about delivery mode (vaginal versus cesarean), **gestational age, birthweight, and** delivery location **is recorded based on information recorded on the child's vaccine card**, breastfeeding initiation is captured **based on maternal report**, a child stool sample is collected, and child anthropometry is measured." (LINES 195-197).
R1.12	(Comment on "stool collection") is better for the authors to mention the advantages of using the method of storing at -196C degrees in liquid nitrogen tanks rather than using DNA/RNA shield reagent that can still store the samples in the long-term. So, all participant are given two different tubes for collecting their fecal samples? If the frozen samples are processed, what happens to those stored in Zymo?	The mother is given a container to collect the fecal sample, and it is only subsequently aliquoted into several tubes by the study team. We have clarified this in the text. "**Mothers are given a small container to store stool and asked to store this in a cooler provided by the study until the study team member arrives to collect the sample."** (LINES 222-224). Any frozen samples and Zymo-stored samples that remain after the initial processing is complete are retained for the sample biorepository, which is described further in the text.
R1.13	(Comment on "stool collection")	Please see response to comment R1.6

	For blood, maternal information are sufficiently provided. Why is the information missing for the maternal gut microbiome? If the samples is not going to be collected or analysis, why?	
R1.14	(Comment on "bacterial microbiome assessment" Very clear! What about the quality control of the extracted DNA samples? How the authors ensure that the DNA is amplifiable?	We have added the following further details. Full results of our stool validation study will be published in a forthcoming manuscript. **"The quality of DNA extracted from both frozen samples and samples preserved in Zymo DNA/RNA shield was previously confirmed during piloting of the laboratory methods (data not shown)."** (LINES 229-231) **"TAC assays will be tested for linearity and matrix inhibition using positive control plasmids designed[48] as well as limits of detection, repeatability, reproducibility, and analytical accuracy using reference strains or genomic DNA/RNA[46]. Extrinsic controls phocine herpesvirus and MS2 bacteriophage will be used to measure extraction efficiency and matrix inhibition[46]."** (LINES 258-262).
R1.15	(comment on "myeloperoxidase, alpha-1-antitrypsin, neopterin, and calprotectin)" Are any of these going to be using? or all of them?	We will assess all of these biomarkers. We have clarified the text:

		"Environmental enteric dysfunction will be assessed by analyzing the stool for **four** fecal biomarkers of intestinal inflammation and permeability". (LINES 262-265).
R1.16	(Comment on "three questionnaires") I believe that the authors should elaborate more details about the questionnaires. The first two questionnaires are developed? Are these modified FFQ, 24-h recalls or food weigh? Please be clear.	We have elaborated as follows: Dietary data is collected using three questionnaires. First, from zero to 24 months of age, weekly surveillance visits include **three** questions to characterize **the presence or absence of** breast-feeding, the intake of non–breast milk liquids, and intake of semi-solid and solid foods in the past week. Second, on a quarterly basis, **a modified dietary diversity** questionnaire is administered to characterize 1) dietary diversity, 2) indicators of complementary feeding such as feeding frequency and minimal acceptable diet, and 3) the child's usage of common micronutrient supplements (vitamin A, iron, or zinc)[55]. **This questionnaire queries about the presence or absence of food groups consumed in the past 24 hours, with some additional questions that capture common regional complementary foods, such as baby porridges frequently made with and without milk.** Third, 24-hour dietary recalls are conducted with the child's primary caregiver at six, 12, and 18 months of age to assess the

		child's intake of non-breastmilk macro- and micro-nutrients and to allow for the calculation of nutrient adequacy ratios and summary measures of overall dietary quality[56,57]. At each time point, this consists of three recalls conducted on non-consecutive days over a one-week period. **The** macro- and micro- nutrient **density of complementary foods** is calculated using a standard reference food composition database for Ecuador[58]. This database has been expanded to include local dishes and recipes specific to the study population. (LINES 286-302).
R1.17	(Comment on "macro and micro-nutrients") Any specific name for the tools or instruments? Does the authors calculate intakes from breastfeeding and adding to the total child's intake? I believe the contribution of breastmilk for children age 6-12 months is still significant. It is possible that the children around this month obtain energy little from complementary feeding but sufficiently from breastmilk. How the authors analyse the breastmilk intake?	The reviewer is correct that the contribution of breastmilk for children 6-12 in this population is significant. Around 25% of infants are fully weaned by 12 months, but most remain partially breastfed by this age. We chose not to measure breastmilk intake. Instead, as a primary analysis, we plan to focus on the energy density of non-breastmilk complementary foods from 12 months onwards. As a secondary analysis, we may estimate intake from breastfeeding following the method recently reported by Morseth et al, among others. This method assumes a total energy intake based on the weight of the child. Dietary recalls collected at 6 months will not be part of the

		primary analysis, as they are instead collected to validate the modified dietary diversity questionnaire, which is implemented more frequently (quarterly) to understand the intake of non-breastmilk foods at complementary ages. We have elaborated on this paragraph (quoted in response to reviewer comment 1.16) to make it clearer that dietary recalls will be used to calculate the nutrient density of complementary foods, rather than absolute nutrient intakes. Morseth MS, Torheim LE, Chandyo RK, Ulak M, Shrestha SK, Shrestha B, Pripp AH, Henjum S. Severely inadequate micronutrient intake among children 9–24 months in Nepal—The MAL-ED birth cohort study. Maternal & child nutrition. 2018 Apr;14(2):e12552.
R1.18	(Comment on "enrolled subjects were interrupted" What protocol has been applied for the enrolled participants? Are they withdrawn or being eliminated from the study? Or still continue with, of course, missing some data.	We have clarified this by changing the word "interrupted" to "modified" and added the following: "**As a result, some data continued to be available during this period, but other data are missing**." (LINES 325-326).
R1.19	(Comment on: Outcome measures and Statistical Analysis)	We have added the following further detail related to

	The way of the authors dividing the study aims was very clear to explain what the study is about. I thought, it still requires the authors to more explain about the statistical approach used or going to be used for SA2 and SA3.	proposed analysis in SA2 and SA3. "**We will develop regression models with random effects for site and individual. e will adjust for** other covariates related to each outcome, such as breastfeeding, reported antibiotic usage across 24 months, and socio-economic status." (LINES 369-372) "Within a given child, we will compare 16S bacterial microbiome communities the week prior to infection with the week during infection. We additionally will compare 16S microbiome communities the week during infection with two weeks **or more** post infection to characterize recovery. **We will examine the impact of community, controlling for other covariates, on microbiome community similarity using ADONIS permutation models, based on Unifrac distances, and visualize differences using NMDS plots."** (LINES 383-388).
R1.20	(Comment on: Relative to 16S sequencing, shotgun metagenomics will provide improved resolution to study pathogens and pathogen genotypes, measure relative in-situ abundance of pathogen populations62, and investigate changes in functional and virulence gene abundances in the gut microbiome in response to pathogen presence.)	We have added a citation.

	Excellent justifications!! Can add a citation(s) for the point "investigate changes in functional and virulence gene abundances"	
R1.21	(Comment on: We plan to enroll up to 480 pregnant women (160 per site)). Some times it is difficult to reach targeted number in enrolment. Do the authors consider; what anticipation or step is made if the target is not reached? what is the pregnancy rate in the study sites? how long the target for recruitment or enrollment process takes?	We had added the following further detail: "**Our current enrollment rate is ~18 dyads per month."** (LINES 421-422).
R1.23	(Comment on: In the region) Insert a citation(s)	We have added this.
R1.24	(Comment on Discussion) The discussion section is written with appropriate portion and academic language that clarifies what are the potential significances and implications for the study.	We appreciate the reviewer's comment.
R1.25	(Comment on Figure 3) Does the author measure anthropometry exact in the specific months? Or using a range for each time point? Because somehow, in the urban area, it is challenging to monitor the infants' growth and development due to high mobility.	We have added further detail, "**Measurements are made within two weeks (14 days) of the targeted day, or up to 6 weeks after the target day if the family is traveling during the desired window**." (LINES 279-280).

		We do experience many challenges given that the population is quite mobile. One help for us has been that families living in more rural areas often relocate to the urban center for short periods of time. Since our study also has a team operating in that urban center, we have been able to maintain contact with some families even as they move between more rural and urban communities.
R2.0	Comments to the Author: Here, the authors describe a protocol for a study that aims to identify the associations between enteric pathogen carriage, the gut microbiome and child growth in a low-income setting. Overall, this manuscript is well-written and includes the details required in a study protocol manuscript. This study will provide some important insights into the interaction between enteric pathogens, microbiome and growth in children at risk of undernutrition. There is some clarification required regarding the criteria for the case-control sub-study and further detail on specific pathogens that will be assessed. Other comments are largely minor.	We appreciate the reviewer's thoughtful reading of the manuscript and address specific comments below.
R2.1	1. The criteria and details of the diarrhoea case-control study require much further clarification. The manuscript states that 200 cases will be selected. Will these be 200 unique participants? It is likely that a large number of participants will have more than one episode of diarrhea. Therefore, which diarrheal episode would be chosen in these circumstances? It would be more suitable to select 200 unique participants rather than including more than one episode from a single participant, to ensure a true n=200.	**We had added additional detail:** "After all intensive stool samples have been collected, we will conduct a case-control study**, using the banked samples. We will select stool samples from 200 diarrhea cases and test for enteric infections using a TaqMan array card. Cases will be selected based on the onset of reported symptoms preceded by seven days of no reported diarrhea. Pathogens that are potentially linked to the diarrheal disease will be based on relative cycle**

		threshold values from the TaqMan results[45]. A control will be selected for each case based on the criterion that the individual has no reported symptoms at the time of stool collection for at least seven days prior and is matched by age and by infection with the same pathogen. For each selected case or control, we will **also** assay three stool samples for 16S **rRNA gene amplicon sequencing** and enteric infections, **using a TaqMan array card (TAC; ThermoFisher) further described below**: one collected the week before the diarrheal episode, one during the week of the episode, and one two **or more** weeks following the episode." (LINES 232-243).
R2.2	Furthermore, the controls in this sub-study need to be clarified. As there are 480 total children in the larger study, this 'sub-study' will include nearly every child (n=200 cases and n=200 controls). The criteria for selecting controls are those carrying the 'same pathogen infection but without diarrhea'. This assumes that, once diarrheal cases are selected, almost every other child will carry the 'same pathogen'. Which pathogens will be considered? Children will likely carry multiple pathogens at different time points. These could contribute to diarrhea at a later time point, so would their presence in an asymptomatic child at an earlier time point still be considered as a 'control' child? There is no detail on whether the pathogen attributed to the diarrheal episode will be identified (although this is recommended). Finally, please elaborate on whether/how these samples will be age-matched. The gut microbiome at 6 months will be quite different to that at 12 months.	Please see our response to comment R2.1
R2.3	- It would be helpful to provide data/references to stunting/growth/diarrhea metrics in this region in the introduction	We have added the following,

		"**Previous studies have placed the prevalence of stunting in our study region at around 15%[38–40] and the two-week prevalence of diarrhea in children under five at around 10%[41]**" (LINE 139-140).
R2.4	- Page 7 (of 28), line 15 – this initial sentence is a bit simplified. Infection/pathogen invasion can still occur in 'healthy' microbiome, therefore best to modify this sentence slightly.	We have modified the sentence, it now reads, "In a healthy microbiome, resident microorganisms **may reduce the risk of** pathogen invasion" (LINES 85-86)
R2.5	- Page 7, line 46 – what does 'differential responses to infection' mean? This is important to clarify if it is the primary hypothesis. May be better to state "… the gut microbome mediates the effect of enteric infections on diarrhea, EED and growth in the first 2 years of life".	We have made this change.
R2.6	- Page 7, line 47 – Similarly, when stating "…contribute to acute and chronic health outcomes?", probably best to state 'associated with' health outcomes.	We have made this change.
R2.7	- Page 7, line 49 – the gut microbiome is "associated with" rather than "modifies" the short and long term health outcomes	We have made this change.
R2.8	- Page 8, line 41 – "We will perform subsequent sensitivity analyses to determine how data from these participants should be used." - sensitivity analyses should be clarified here rather than planned post-hoc	We have removed the word 'sensitivity' so that the sentence now reads, "We will perform subsequent analyses to determine how data from these participants should be used." (LINES 159-160)

R2.9	- Please provide details on how gestational age will be determined (LMP?) as this will be important for eligibility criteria of 37 weeks.	We have modified this to reflect the fact that gestational age is based on maternal report, we also wish to note that this information is subsequently confirmed through information about the gestational age at delivery, which is reported by the Ministry of Health in a vaccine card given to the mother. "Women are recruited and enrolled in late pregnancy, from 37-weeks onward, **where gestational age is based on the mother's report.**" (LINES 142-143). "Within seven days of birth, information about delivery mode (vaginal versus cesarean), **gestational age, birthweight, and** delivery location **is recorded based on information recorded on the child's vaccine card**, breastfeeding initiation is captured **based on maternal report**, a child stool sample is collected, and child anthropometry is measured." (LINES 194-197)
R2.10	- Please provide details of which pathogens will be screened for on Taq cards. This is an important aspect of the study and one that has received little detail in this manuscript. There will likely be carriage of multiple pathogens at a time. It would also be beneficial to attempt to identify the pathogen attributed to the diarrheal episode (as has been reported in MAL-ED studies using the same technology). See major comment regarding this point.	We have added the following, "**The specific pathogens we will test for are: Enteroaggregative Escherichia coli (EAEC)AEC, Diarrheagenic Escherichia coli (DAEC), Shiga toxin-producing**

		Escherichia coli (STEC), Enteropathogenic Escherichia coli (EPEC), Enterotoxigenic Escherichia coli (ETEC), Shigella, Enteroinvasive Escherichia coli (EIEC), Campylobacter jejuni, Campylobacter coli, Salmonella Typhi, Adenovirus, Astrovirus, Enterovirus, Norovirus GI, Norovirus GII, Rotavirus, Sapovirus, SARS-CoV-2, Cryptosporidium hominus, Cryptosporidium parvum, Cyclospora cayetanensis, Giardia lamblia, Entamoeba histolytica, Ascaris lumbricoides, and Trichirus trichiura." (LINES 151-158) We will identify pathogens attributed to the episode using Ct values, please see our full response to comment R2.1.
R2.11	- Presumably models will be adjusted for site (urban and rural sites), however this has not been mentioned	Please see response to comment R1.19.
R2.12	- Page 12, line 34 – What are meant by "founder community characteristics" and "maturity indices"? No detail is provided previously about these statements. Maturity indices presumably refers to MAZ scores. If so, please provide details.	We have added further detail: **"We will also use stool samples collected at one week of age to assess whether the initial colonization of the microbiome influences assembly and subsequent composition at later ages (founder effects)."** (LINES 167-169).

		We have removed the reference to maturity indices.
R2.13	- Page 13 line 3 – Similar to previous comments, how will enteric infections be defined if more than one pathogen is present (which is likely)? Simple presence/absence or confirmed pathogen-attributable diarrhea?	Please see our full response to comment R2.1.
R2.14	- Please provide some details on shotgun sequencing methods – sequencing and bioinformatic pipelines etc	We have added the following detail, "**Shotgun metagenomes will be sequenced using Illumina sequencing chemistry as previously described[69]. Raw FASTQ reads from the sequencing runs will be quality checked, trimmed, assembled, and annotated using tools implemented in the MiGA (Microbial Genomes Atlas) pipeline[70], which was developed for efficient processing and management of microbial metagenomes. We will use read-based mapping to quantify annotated metagenome features, and will normalize abundances based on the sequencing depth of an external spike-in control[71]. Additional open-source or internal software tools will be used to remove human read contamination, annotate gene functions and taxonomy, and bin and quality check metagenome-assembled genomes, as we have done with other analyses[69]."** (LINES 441-419).

R2.15	- Table 3 - The available sample sizes column is unclear? The study has n=480. How are there available samples sizes of >1000 for some models e.g 1a and 1b examining the association between baseline water and microbiome/infection?	We had added the following example: "**For example, household water access and the microbiome will each be measured at least 5 times per child, resulting in an estimated 1800 paired measurements. We similarly estimate that, assuming an intraclass correlation coefficient of 1.1 to account for within-child correlation between measurements, 1268-1800 measurements would be needed to detect differences in the Chao1 richness related to water access, if the overall prevalence of access to improved water in the study communities is between 25 and 50%."** (LINES 432-437).
R2.16	- Figure 3 - EED fecal biomarkers are only being assessed at 6 months. Therefore how will this be able to meet objectives of aim 2? Is this a mistake in the gantt chart?	Currently, we only plan to measure EED fecal biomarkers at a single time point (6 month) and to assess associations between the microbiome and EED biomarkers based on that single sample. We feel that a single time point, around the age when growth faltering is likely to begin, is appropriate for this relatively exploratory aim 2b.
R2.17	- There is much emphasis on the importance of the rural-urban gradient in this cohort but very little detail on how this will be considered statistically or interpreted. Will analyses simply be split by the 3 different sites and compared or will there be more complex considerations as to how/why this gradient is influential.	Please see response to comment R1.19. In our primary analysis, we plan to include a random effect for site. However, combined analyses will be preceded by site-specific stratified analysis to assess the extent to which

		observed associations are consistent or inconsistent between sites.

VERSION 2 – REVIEW

REVIEWER	Thahir, Andi The University of Sydney, Central Clinical School
REVIEW RETURNED	10-Jul-2021

GENERAL COMMENTS	Dear authors I'd like to say thank you for your efforts to revise the manuscript. I understand that my concerns, including lack of information about maternal factors, statistical analysis, gut microbiome methods, and questionnaires as well as the abstract part have been addressed in this latest version. Therefore, I recommend this paper to accept for publication. One last concern about dietary factors – lines 299 – 300. What does "the macro and micronutrient density of complementary foods" mean? Is that dietary intake will be analyzed as nutrients or foods (food groups)? It is probably important to decide how the dietary data will be analysed, as the association between these two and the gut microbiome might be different (food-microbiome vs nutrient-microbiome).

REVIEWER	Robertson, Ruairi Blizard Institute, Queen Mary University of London, UK
REVIEW RETURNED	01-Jun-2021

GENERAL COMMENTS	The manuscript has been much improved in its revised from. There are just two remaining comments: 1. It is still necessary to clarify the case-control sub-study with more detail. It appears that the authors will be selecting 200 diarrheal stool samples rather than 200 unique infants with diarrhea. Therefore, it is likely that the 200 'cases' selected will contain multiple stool samples from the same infants, rather than 200 unique infants. In this sense, these are not true cases. Therefore, the authors should refer to these as 200 diarrheal 'episodes' or diarrheal 'stools'. It is also still not clear how these will be selected. Will these 200 be selected at random? Or will age be considered when selected these 200 samples in order to obtain an even distribution across 6-12 months of age? This approach also suggests that a child could act as both a case and a control if it has a diarrheal stool/episode chosen as a case from an earlier time point and a non-diarrheal stool at a later time point or vice versa. Therefore, it is necessary to state if this will be the case. How will age-matching be conducted between cases and controls (same month of age? Same week of age?)? These criteria need further consideration and clarification. 2. There is a typo on line 362 of the clean manuscript version: "… random effects for site and 362 individual. e will adjust…"

VERSION 2 – AUTHOR RESPONSE

	Comment	Response
R2.1	It is still necessary to clarify the case-control sub-study with more detail. It appears that the authors will be selecting 200 diarrheal stool samples rather than 200 unique infants with diarrhea. Therefore, it is likely that the 200 'cases' selected will contain multiple stool samples from the same infants, rather than 200 unique infants. In this sense, these are not true cases. Therefore, the authors should refer to these as 200 diarrheal 'episodes' or diarrheal 'stools'. It is also still not clear how these will be selected. Will these 200 be selected at random? Or will age be considered when selected these 200 samples in order to obtain an even distribution across 6-12 months of age? This approach also suggests that a child could act as both a case and a control if it has a diarrheal stool/episode chosen as a case from an earlier time point and a non-diarrheal stool at a later time point or vice versa. Therefore, it is necessary to state if this will be the case. How will age-matching be conducted between cases and controls (same month of age? Same week of age?)? These criteria need further consideration and clarification.	The reviewer is correct that some children may contribute more than one sample to the sub-study, so we have adjusted our language so that we no longer use the terms 'case' or 'control' to describe this. We have also clarified in our text that diarrhea samples will be selected to evenly represent a rural-urban gradient, but not an equal distribution by age. This is because one of our prespecified aims is to assess characteristics of the microbiome from children living across a range of urbanicity characteristics. The revised text now reads as follows: After all intensive stool samples have been collected, we will conduct a **sub analysis** using the banked samples. We will first randomly select 200 stool samples from **children who experienced an episode of** diarrhea **at the time of the sample collection, where an episode is defined as an onset of diarrhea** preceded by seven days of no reported diarrhea**. These samples will be randomly selected among geographic strata to ensure equal representation of children across the rural-urban gradient of the study. A matched stool sample from a child without diarrhea** will be selected based on the criterion that the individual has no reported symptoms at the time of stool collection for at least seven days prior and is matched by age **(+/- 1 month of the symptomatic child**) and by infection with the same pathogen. **We will test these samples for enteric infections using a TaqMan array card. Pathogens will be linked to the diarrheal disease episode based on relative cycle threshold values from the TaqMan results (45). Children may contribute both diarrheal and asymptomatic stool samples to the study at different ages. This repeated sample design will be accounted for in the analysis.**

R2.2	2. There is a typo on line 362 of the clean manuscript version: "… random effects for site and 362 individual. e will adjust…"	We have corrected this.
R1.1	I'd like to say thank you for your efforts to revise the manuscript. I understand that my concerns, including lack of information about maternal factors, statistical analysis, gut microbiome methods, and questionnaires as well as the abstract part have been addressed in this latest version. Therefore, I recommend this paper to accept for publication. One last concern about dietary factors – lines 299 – 300. What does "the macro and micronutrient density of complementary foods" mean? Is that dietary intake will be analyzed as nutrients or foods (food groups)? It is probably important to decide how the dietary data will be analysed, as the association between these two and the gut microbiome might be different (food-microbiome vs nutrient-microbiome).	We decided to delete the reference to micronutrient density as those details are not necessary in this protocol paper. We did add clarification to our dietary assessment plan to emphasize our focus on nutrients as follows: **We will calculate** macro- and micro- nutrient **intakes from** complementary foods using a standard reference food composition database for Ecuador[62]. This database has been expanded to include local dishes and recipes specific to the study population. **Macro and micronutrient intake data will be used to create covariates for the analyses described below. We will use data reduction techniques to summarize intake data. These techniques may include *a priori* methods to assess specific characteristics of the diet, such as its inflammatory potential[63], as well as *a posteriori* data-driven approaches such as principal components analysis and reduced rank regression[64].** (LINES 294-300).