**Peer Review File**

Article Information: http://dx.doi.org/10.21037/pm-20-100

**Reviewer 1**

Comment 1: Symptoms of GER were measured using the Infant Gastroesophageal Reflux Questionnaire – Revised (I-GERQ-R). The I-GERQ-R is a 12 item parent-reported questionnaire about symptoms of GER in the seven days prior to the questionnaire being completed. Is the questionnaire objective and accurate? Does the education level of parents affect the completion of the questionnaire?

Reply 1: Yes, the I-GERQ-R is objective and accurate. It has been well-studied in a variety of populations with varying educational levels. The psychometric properties of the tool are described under the Measurement section.

Comment 2: Infants born full-term showed improvements in GER symptoms with increasing corrected age, but none of the preterm infant groups showed these same improvements over the first 6 months of life. Is there an effective treatment for gastroesophageal reflux in preterm infants?

Reply 2: Treatment and management of gastroesophageal reflux in preterm infants is a controversial topic. A full description of this is beyond the scope of this article, but there are many other publications available on this topic. One of the references (Pados & Davitt 2020) provides readers with non-pharmacologic strategies.

Comment 3: The findings indicate that, in addition to preterm birth and younger corrected gestational age, a family history of allergy is a factor that may be helpful in identifying infants at particularly high risk for developing significant gastroesophageal reflux symptoms. What are the underlying causes of more gastroesophageal reflux symptoms in infants with a family history of allergy?

Reply 3: We have added several sentences to the discussion section about underlying mechanisms for the relationship between family history of allergy and gastroesophageal reflux symptoms.

Comment 4: The relationship between GER and microbiota has been mentioned many times in this manuscript. Can probiotics improve the symptoms of gastroesophageal reflux? Although infants born 32–37 weeks did not exhibit more symptoms than infants born full-term, they failed to have the same improvement in symptoms with increasing age. What is the biggest harm of not improving GER symptoms?

Reply 4: There have been limited studies of the use of probiotics for the management of GER symptoms. This is another area in need of more study, particularly for preterm infants who may have disruptions in the intestinal barrier. As indicated in the introduction section, GER symptoms are associated with feeding difficulties, which may

impact growth and development.

Comment 5: Infants born prior to 32 weeks also frequently experience respiratory disease and growth faltering, which may require increased feeding volumes and nutritional changes, that may contribute to GER symptoms. Can high nutritional density food or semi liquid food reduce GER symptoms?

Reply 5: Several sentences and references were added to the discussion about nutrition, nutritional density, and GER symptoms.

**Reviewer 2**

Comment 1: One of the difficulties with the study is the fact that while almost 600 children were recruited the large majority were full term infants, only about 100 were <37 weeks gestation. Furthermore, by dividing the preterm group into 4 cohorts of different gestational ages the number of infants in each group was rrelatively small, especially in the 4-6-month group (which included only 2-3 infants for each preterm cohort), which perhaps is the most important group in their final conclusions.

Reply 1: We acknowledge in the limitations that the sample sizes within the preterm infant cohorts were small. We have combined gestational age groups to create three categories by gestational age (i.e., < 32 0/7 weeks, 32 0/7 – 36 6/7 weeks, and ≥ 37 weeks). This has increased our sample sizes within the cohorts for analysis. We have also added a statement in the conclusion indicating that additional research with larger samples is needed.

Comment 2: Another concern is the statement that a difference of 1 in the I-GERQ-R is “clinically significant.” The authors give no reference for this statement. A recent “quantitative synthesis” of multiple studies (Smith et al. Patient Related Outcome Measures 11:87-93, 2019), while finding that the I-GERQ-R is a useful tool nevertheless make the statement that “The instrument developers have determined a change score of 5 to 6 to be of clinical importance7,8 based on parental/caregiver, and clinician rating of symptom change. A change of 3 on the I-GERQ-R has been deemed by the developers to represent a minimally important difference (MID), although the basis for this is not known. Therefore, in this study, an MID was defined as the smallest perceptible change perceived as beneficial (or deleterious) to the patient.”12 As a clinician, I find a difference of 1 to be highly unlikely to be of clinical significance. Because the authors chose 1 to be their benchmark, it is thus likely that many of their “significant differences” were NOT clinically significant. In fact, it is only in the 4-6-month group that mean differences of >3 (perhaps a “minimally important difference” (see above)) were found (Figure 2) … and this is the preterm group with the smallest “n’s” infants. The authors should clarify the basis for their decisions and perhaps rewrite the results, discussion, and conclusion sections accordingly. The authors do acknowledge some of these limitations in the appropriate section.

Reply 2: We have used the reference by Smith et al. (2020) to calculate an average of

the mean standardized differences across the seven studies reported in that meta-analysis. We used this average (1.3) as the effect size in our power analysis. Because the effect size is calculated as the mean difference between groups divided by the standard deviation, we used the reported mean standardized differences instead of the change score. As reported in Smith et al. (2020), the mean standardized difference was calculated as the change score divided by the standard deviation at baseline, so this was appropriate. Of note, an effect size of 1.3 with a standard deviation of 4, which was common across the seven studies reported, would equate to a difference in I-GERQ-R scores of 5.2, which is within the range of reported clinical importance. We have removed the statement about clinical significance.

Comment 3: Looking at Figure 1, it appears that the "very preterm" and "extremely preterm infants" are essentially similar, and the older preterm & full-term cohorts are also similar. Perhaps comparing those 2 combined groups will yield clues as to why the two groups differ.

Reply 3: We have combined the extremely and very preterm groups and have combined the moderate and late preterm groups.

Comment 4: In the discussion, the authors should compare and contrast previous studies that have addressed the same problems (for example, Martin et al. Pediatr. 109:1061, 2002, Curien-Chotard et al., BMC Pediatr. 20:152, 2020; Campanozzi et al., Pediatr. 123:779, 2009, etc.).

Reply 4: We have add references to Curien-Chotard and Campanozzi's studies, which both found improvement in GER symptoms over the first 6 months of life in otherwise healthy children.

Comment 5: After addressing the questions above, the authors should indicate in Figures 1 and 2 (with *) which groups are significant.

Reply 5: This change has been made to Figure 1. Figure 2 became very busy and confusing when the comparisons were overlayed on the figure, so we chose to not make this change to figure 2. Readers can refer to the text for which comparisons are statistically significant.

Comment 6: Re-doing Table 2 by gestational age may also reveal differences that could explain GER-score differences.

Reply 6: Table 2 is intended to provide general characteristics of the sample. Since these factors were not included in the analysis as potential explanatory factors for I-GERQ-R we have chosen not to report table 2 by gestational age group.