**Date:** 1/28/2021

**Call for global standardization in micronutrient biomarker statistical analysis, an introduction to the SAMBA R package[[1]](#footnote-1),[[2]](#footnote-2),[[3]](#footnote-3)**

**Authors involved**: Hanqi (BRINDA, UCD), Ty (GAIN), Yaw (CDC, Emory), Charles (UCD statistician), Melissa (Emory), and Parmi (Emory, CDC)

**Hanqi Luo:** Department of Global Health and Global Health Institute, Emory University

, Atlanta, GA, USA; Institute for Global Nutrition and Department of Nutrition, University of California, Davis, CA, USA. ORCID: 0000-0001-6253-5818

**Ty Beal**: GAIN, Washing, DC, USA. ORCID: 0000-0002-0398-9825

**O.Yaw Addo:** Department of Global Health and Global Health Institute, Emory University

, Atlanta, GA, USA; McKing Consulting Corporation Atlanta, GA, USA; Centers for Disease Control and Prevention (CDC), Nutrition Branch, International Micronutrient Malnutrition Prevention and Control Program (IMMPaCt) Unit, Atlanta, GA, USA. ORCID: 0000-0003-1269-759X

**Charles D Arnold**: Institute for Global Nutrition and Department of Nutrition, University of California, Davis, CA, USA. ORCID: 0000-0001-6510-3172

**Melissa F Young**: Department of Global Health, Emory University, Atlanta, GA, USA

**Parminder S Suchdev**: Department of Pediatrics, Emory University, Atlanta, GA, USA; Emory Global Health Institute, Atlanta, GA, USA; Division of Nutrition, Physical Activity and Obesity, US CDC, Atlanta, GA, USA

\* To whom correspondence should be addressed: ???

**Authors’ names for Pubmed indexing:**

**Word count (all text, excluding abstract, figure legends, tables, and reference):** 5376

**Number of Tables:** 2

**Number of Figures:** 2

**Online Supporting Material:** 2

**Running title:** SAMBA package to standardize micronutrient biomarker statistical analysis

**Abstract (structured)**

**Background**:

To justify, plan, and manage successful micronutrient intervention programs, micronutrient biomarkers are the most essential. However, analysis of micronutrient biomarker data typically requires extensive training and skill.

**Objectives:** We introduce the Statistical Apparatus of Micronutrient Biomarker Analysis (SAMBA) package, a new tool to increase accessibility of micronutrient biomarker analysis. We explain the underlying theory behind the tool and provide examples of potential applications.

**Method:** Prior to the SAMBA package, analysts need to understand the biological basis for micronutrient biomarkers, equip with statistical knowledge of analyzing micronutrient biomarkers, and have familiarity with basic computer code. The freely available and open-access SAMBA package written in R software simplifies estimations of micronutrient distribution and prevalence of deficiencies, which can drastically shorten the time from data cleaning to presentation of a final results table.The package and associated documentation are freely available. We used the SAMBA package to analyze data from the US National Health and Examination Survey (NHANES) and a single-site observational study in Kenya to compare with published papers on XX XX XXX.

**Results:**

The SAMBA package generates identical results as the published papers. The examples demonstrate application of the SAMBA package to micronutrient biomarker data with proper statistical and inflammation adjustment implementation.

**Conclusion:** The SAMBA packages may facilitate the analysis of micronutrient biomarker data to inform nutrition research, programs, and policy.

**Word count:**

**Keywords:** Micronutrients, biomarkers, statistical analysis, software, inflammation adjustment

**Introduction**

Hidden hunger, also known as micronutrient deficiencies, has been said to affect more than 2 billion individuals (1). Micronutrient deficiencies can contribute to poor growth, intellectual impairment, and increased risk of morbidity and mortality (2). Micronutrient intervention programs, such as large-scale fortification and supplementation, are available to address this burden. However, to justify, plan, and manage successful micronutrient intervention programs, a broad range of micronutrient information is needed, such as data on food availability, food and nutrient intake, micronutrient status, and intervention program coverage. Of these, analyses of micronutrient biomarkers are the most essential, as these data are only possible method to determine the prevalence of distribution of micronutrient deficiencies (3).

The 2017 Global Nutrition Report (GNR) proposed a “nutrition data revolution”, calling for “better use of data that is collected to create a more responsive information system” for policy decision-making (3). This revolution does not only apply to low- and middle- income countries (LMICs), but is also pertinent to high-income countries, where there is likewise a need to address micronutrient deficiencies in certain groups such as iron deficiencies in pregnant women. During the nutrition data revolution, a “nutrition data value chain” was introduced, which has five critical processes – data prioritization, creation and collection, curation, analysis, interpretation/recommendation, and decision-making (3,33). To date, enormous effort has been spent overcoming the early challenges in this chain, data prioritization, collection, curation, and lab analysis (4–6). The recent published micronutrient survey manual & toolkit provides comprehensive information on planning a national micronutrient survey (4); the Hemoglobin Measurement Laboratory Protocol outlined the standard laboratory method for obtaining accurate results of hemoglobin in population surveys (5). Even though these guidelines provide recommendation on data analysis and presentation, little has been achieved in *streamlining the analysis* of micronutrient biomarker data. This analysis still requires intermediate knowledge of statistics (e.g., analysis of surveys with complex design), high-level of proficiency in computer coding, and a deep understanding the biological bases of micronutrient biomarkers (e.g., the mechanism of how inflammation influences micronutrient biomarkers) (7). Due to the complex nature of micronutrient data, organizations need to hire senior statisticians with a specialization in micronutrient data, who might not always be available. Alternatively, organizations need to invest in building this capacity in their analysts; however, the time lag may result in lost opportunities to make micronutrient biomarker data available during critical decision-making periods.

To improve the utilization of micronutrient data, we provide an easy-to-use and “all-in-one” tool, Statistical Apparatus for Micronutrient Biomarker Analysis (SAMBA). Three primary advantages of this tool are that 1) it provides a single tool that streamlines biomarker analysis for MULTIPLE datasets, which reduces the time required for analysts to clean and analyze biomarker datasets; 2) without any modification of the R package, this tool allows for analyzing additional biomarkers beyond the six example micronutrient biomarkers built in this package; 3) the package is built in R, a free and open-access software, which means no financial investment in software needed. The SAMBA package and example codes and example data are available online (Open Science Framework link: XXXX); the corresponding SAMBA package user manual is provided with this article (**supplemental user manual 1**). We hope the detailed user manual and examples will allow analysts to modify the example codes for their own micronutrient biomarker data.

The objectives of this paper are to describe the current challenges in micronutrient biomarker analysis, display the general structure and features of the SAMBA package by explaining the underlying analysis algorithm, present example applications of the SAMBA package for micronutrient biomarker analysis using a single-site study in Kenya [REF] and a nationally representative survey, the United States’ National Health And Nutrition Examination Survey (NHANES)(8), and at the end, describe the strengths and limitations of the SAMBA package in the discussion.

**Challenges in micronutrient biomarker analysis**

Numerous challenges exist to analyze global micronutrient data. We have listed some challenges in the following section, also illustrated in **Fig. 1** .

* **Datasets stored in different file formats:** It iscommon that survey coordinators who implement surveys are not also data analysts; therefore, analysts might receive data files in a format that is not compatible to the analysis software that they used. Although certain packages were developed to convert different types of data files (e.g., STATA, SAS, and excel files), analysts still need to learn the related packages to carry out this conversion.
* **Different variable names:** One biomarker can be stored in different variables names. For example, biomarker serum ferritin can be stored under the variable name, serum\_ferritin, serum.ferritin, sf, ferritin, or FER. This can cause confusion for analysts, especially with absence of a comprehensive codebook.
* **Different units:** Biomarkers can be stored in either conventional units or the International System of Units (SI) (9). Most often, analysts will need use additional information such as molar mass of the biomarker to convert between conventional units and SI, so that they can compare the biomarker with a cutoff established with consistent units.
* **Different methods to adjust for inflammation:** Inflammation is known to affect many micronutrient biomarkers, which can thus lead to overestimate or underestimate the prevalence of deficiencies in a population (7). The inflammation adjustment methods, such as Thurnham (10) or Biomarkers Reflecting Inflammation and Nutrition Determinants of Anemia (BRINDA) method (7), have been well-documented; however, analysts still need to learn and apply those methods, which requires additional learning time.
* **Different cutoffs for micronutrient deficiencies:** Micronutrient deficiencies are calculated based on varying cutoffs. Some cutoffs are based on manufactures’ recommendation, and some are based on international consensus. In addition, physiological or environmental factors can also influence the cutoffs. For example, anemia cutoff using hemoglobin values can be affected by smoking status and altitude (11,12); zinc deficiency cutoff using serum zinc can be influenced by fasting status (13).
* **Different data analysis approaches based on the study design:** Analysts should have the knowledge required for analyzing studies with different study design. For a multi-stage complex survey design, analysts should understand how strata, clusters, and survey weights play a role in estimate parameters of interest (e.g., mean, geometric mean, prevalence of deficiencies).
* **Time and financial commitment required for capacity building of in-country analysis:** In-country analysis of micronutrient data is essential for capacity building in public health nutrition (14), which can facilitate timely reporting and thus lead to effective policy decision. However, such capacity building requires time and financial commitment as well as multi-sectional collaboration.
* **Reproducibility of scientific results cannot be always ensured:** Inconsistent and contradictory results from analyzing the same study by different analysts continues to emerge (15) and, therefore, compromise the validity of the study. Inaccessible analytical method and codes make evaluation and diagnosis of the flawed analyses difficult.

**Introduction to the bio\_analysis R package**

The SAMBA R package is an easy-to-use and “all-in-one” tool that facilitates estimation of micronutrient biomarker distributions and prevalence of micronutrient deficiencies for multiple datasets simultaneously. The SAMBA package also includes additional features, such as carrying out checks on the input datasets and implementing a variety of specific analyses, including adjusting for inflammation for certain biomarkers (7) and taking into consideration of multi-stage complex survey design if relevant. Prior to this paper, the associated codes and analytical details of those analyses have been published but in different papers and reports, which requires analysts to learn them separately, and may thus limit the extent to which analysts can then apply these methods to their own data.

An overview of the SAMBA package inputs, core contents, and outputs is illustrated in **Figure 2.** Two input .csv files are required: 1) biomarker dataset template and 2) biomarker cutoff template. For the biomarker dataset template, analysts need to fill in information on the location of the dataset in the computer, survey-related information (if relevant, e.g., strata, cluster, and survey weight variable names), demographics information (e.g., age and sex variable names), and micronutrient biomarker variable names and units. In this template, analysts can include information for multiple datasets (one column per dataset). For the biomarker cutoff template, analysts need to review the existing values of the biomarker cutoffs that vary by age and population groups (men, women, and pregnant women), and then they can make modifications on the cutoff template in case that the cutoffs in the template differ from their desired values.

The SAMBA package has multiple steps. SAMBA will first load the biomarker dataset template and cutoff template and preload the biomarker dataset, and then use the information to check for dataset specification errors before starting analysis; if any detectable errors are found, SAMBA provides guidance for correcting them. If no errors exist, the biomarker datasets will be officially loaded. Currently, SAMBA supports SAS, SPSS, STATA, CSV, and EXCEL files. SAMBA will change the variable names listed in biomarker dataset template into standardized variable names (predefined in the biomarker dataset template), as well as change the units of biomarkers into SI. Then, SAMBA will adjust for inflammation using the BRINDA method for selected nutrients (7). After adjustment, a cleaned dataset in csv format per dataset analyzed will be exported, which contains standardized variable names and units and adjusted biomarker values for inflammation for all study participants (output 1). SAMBA will also estimate parameters based on study design to generate a formatted csv file suitable for use in reports or manuscripts containing estimated results of the micronutrient biomarker distribution, such as the mean, geometric, median, IQR, and prevalence of deficiencies with associated standard errors and confidence intervals (output 2).

**Extension of the bio\_analysis R package**

The SAMBA package was built to be generic to analyze any additional biomarkers that can be dissected with a binary outcome. Without any modification of the R package, this tool permits analyzing additional biomarkers, such as blood pressures and cholesterols, beyond the six default micronutrient biomarkers built in this package. Analysts just need to need to add the variable name, units, and survey weight variable (if relevant) of additional biomarkers in the biomarker dataset template, and add the cutoffs of the biomarkers (e.g., the cutoff of high blood pressure using diastolic blood pressure). After a successful run of the SAMBA, a cleaned dataset for all study participants per dataset analyzed (output 1) will include information of the additional biomarker, and the result file (output 2) will contain the distribution and prevalence of deficiency of this biomarker.

In addition, researchers can merge the outputs of the SAMBA package with other variables, such as underweight and Body Mass Index (BMI) to explore a wide range of research questions. For example, the cleaned dataset (output 1) can be merged with individual’s weight, height, and other health comes to explore the association between micronutrient deficiencies and Non-Communicable Diseases (NCD). As micronutrient deficiencies is an important component of double-burden malnutrition, the result file (output 2) of several nationally representative surveys can be merged with aggregated data on over- and under-nutrition to explore malnutrition in all forms globally (16–18).

**Validation of the SAMBA package**

We applied the SAMBA package to a single-site study in Kenya [REF] and NHANES (8). The results generated by the SAMBA package were identical to the results published before [Ref].

* Description of the results (TBD)

**Discussion**

The SAMBA package provides a streamlined and efficient structure for micronutrient biomarker analysis for multiple datasets. Prior to the SAMBA package, analysts need to understand the biological basis for micronutrient biomarkers, obtain statistical knowledge of analyzing micronutrient biomarkers, and have familiarity with basic computer code. With this background, the new, freely available, and open-access tool simplifies estimations of micronutrient distribution and prevalence of deficiencies. In all, the efficient programming capacity of this tool can drastically shorten the time from data cleaning to presentation of a final results table.

Micronutrient deficiencies is an important component in malnutrition for both LMICs and high-income countries (16), especially when micronutrient deficiencies are also prevalent in normal-weight and overweight population (19). However, micronutrient deficiencies are always overlooked when assessing all forms of malnutrition, due to both the scarcity and limited utilization of data (16). The GNR report and other initiatives have emphasized to fill in the data gap in a holistic approach by improving all components in the data values chains, from data prioritization, creation and collection, curation, analysis, to interpretation/recommendation (20,21). The SAMBA package is a software tool that can be applied to diverse global contexts to fill the gap in the nutrition data value chain between data collection and interpretation for decision-making. The SAMBA package reduces the time spent on 1) learning basic biological knowledge of micronutrient biomarkers (e.g., how inflammation influences biomarker values), 2) intermediate statistical skills (e.g., analysis of complex survey design), and 3) time to write codes to analyze each biomarker. Researchers and analysts can focus on the actual research questions instead of learning different pieces of micronutrient biomarker analysis from a variety of published papers and reports. The feature of analyzing multiple micronutrient datasets is especially useful for countries that continuously collect micronutrient data in their nutrition surveillance system and organizations that monitor the global micronutrient status. Furthermore, the SAMBA package is built in R, a free and open-access software, so that Institutions and analysts do not need to make additional investment in the software purchase, which can improve the accessibility of the SAMBA package. In addition, there is an increasing push to improve research capacity at institutions in LMICs. The SAMBA package can facilitate these efforts by providing a user-friendly platform for analyst to conduct micronutrient biomarker data analysis, as well as as the ability to automatically output a spreadsheet containing the information typically requested for national and regional micronutrient biomarker surveys (e.g., mean, geometric mean, median, and prevalence of micronutrient deficiencies) to local governments and the international community, with substantial reductions in the time and cost required to do so. Moreover, extensions of the tool permit analyses of additional biomarkers beyond the default six biomarkers and connecting the outputs of the SAMBA package (either the aggregated result file or clean datasets) with other nutrition information (over- and under-nutrition) and health outcomes to explore malnutrition in all forms and association between micronutrient deficiencies and NCDs.

The SAMBA package also has several limitations. First, the SAMBA package can only generate unbiased results when the micronutrient biomarker data are of high quality and represent the population’s micronutrient biomarker distribution. It can neither correct for errors in data collection nor removing extreme or unreasonable values of micronutrient biomarkers. By default, the SAMBA package will recode zero values to 0.0001 so that log transformation of the biomarker values can be made possible for inflammation adjustment and calculation of the confidence interval of geometric means; SAMBA will not remove any outliners, because for lots of surveys, such as NHANES, biomarker values are already cleaned and within a reasonable range. To ensure the validity of micronutrient data analysis, besides rigorous, standardized training of laboratory technicians, we recommend analysts carefully clean data, such as changing the zero values of micronutrient biomarker to the minimal detectable values based on manufacturers’ laboratory protocols and remove any outliners before applying the SAMBA package to their own data. Second, the current version of the SAMBA package cannot calculate percent of people with more than nutrient deficiencies, such as the prevalence of both vitamin A and iron deficiencies. To address this limitation, we recommend people use the cleaned dataset (output 1) from the SAMBA package to calculate the overlapping micronutrient deficiencies. Third, the SAMBA package has specific technical requirements. Analysts need to have R software installed and equip with basic R programming skills, which requires time and access to appropriate training materials. These user requirements may block the usage of the SAMBA package from a wider audience.

To facilitate use of the current SAMBA package version, we have designed training materials and a thorough user manual to enhance the usage of the SAMBA package and build the capacity of researchers and policy analysts (training materials are available upon request). Plans to expand the current SAMBA tool include the addition of new functionalities, such as analysis of overlapping of micronutrient deficiencies and generating a codebook of datasets based on the information filled in biomarker dataset template as well as the results of analysis. In the longer term, we envision development of a web-based tool based on the SAMBA package to further increase the accessibility of the method and decrease the time and resources required to make micronutrient biomarker data results available.

**Acknowledgement**

The authors appreciated comments from Micronutrient Forum Advisory Panel led by Kenneth Brown, Lynnette Neufeld and Mduduzi Mbuya from GAIN, Reina Engle-Stone and Stephen Vosti from University of California Davis, and Juna Carlos Sequel. The authors gratefully acknowledge BRINDA team members for their contribution to collection and interpretation of the datasets used in this paper. All authors contributed to the model design and interpretation of the data; **H.L.** developed the method and wrote the paper; **T.B.** proposed the idea and assisted in method development and editing; **C.A.** and **OY.D.** provided statistical support. **P.S.** and **M.Y.** provided overall guidance on the project and participated in the paper editing. All authors critically revised the manuscript and approved the final version.

Creation of the statistical code presented in this paper was made possible through support from the U.S. Agency for International Development (USAID) under the terms of contract 7200AA18C00070 awarded to JSI Research & Training Institute, Inc. (JSI). The contents are the responsibility of JSI and do not necessarily reflect the views of USAID or the U.S. Government. This research was also supported by the Bill & Melinda Gates Foundation (OPPXXXX).

**Reference**

1. FAO, WFP, IFAD. The State of Food Insecurity in the World 2012. Economic growth is necessary but not sufficient to accelerate reduction of hunger and malnutrition. [Internet]. Rome: FAO; Available from: http://www.fao.org/3/i3027e/i3027e00.htm

2. Bailey RL, West Jr. KP, Black RE. The Epidemiology of Global Micronutrient Deficiencies. Ann Nutr Metab. 2015;66:22–33.

3. Brown KH. Future needs and challenges for generating better data on micronutrient status [Internet]. Online; 2020 [cited 2021 Jan 21]. Available from: https://www.eventscribe.com/2020/MNF-CONNECTED/fsPopup.asp?efp=QU9WU1BJREMxMzcwMQ&PresentationID=780929&rnd=5.262113E-02&mode=presinfo

4. CDC, Nutrition International, UNICEF. Micronutrient Survey Manual & Toolkit [Internet]. 2020 [cited 2021 Jan 22]. Available from: https://mnsurvey.nutritionintl.org/

5. USAID Advancing Nutrition. Protocol for Comparative Evaluation of Blood Sampling Methods and Analytical Devices in the Measurement of Hemoglobin in Population Surveys—A Laboratory Study [Internet]. Arlington, VA: USAID Advancing Nutrition; Available from: https://www.advancingnutrition.org/sites/default/files/2020-06/py2\_deliverable\_24d\_hemeprotocol\_laboratory\_20200430\_submitted.pdf

6. O’Callaghan KM, Roth DE. Standardization of laboratory practices and reporting of biomarker data in clinical nutrition research. The American Journal of Clinical Nutrition. 2020;112:453S-457S.

7. Suchdev PS, Namaste SM, Aaron GJ, Raiten DJ, Brown KH, Flores-Ayala R, on behalf of the BRINDA Working Group. Overview of the Biomarkers Reflecting Inflammation and Nutritional Determinants of Anemia (BRINDA) Project. Advances in Nutrition. 2016;7:349–56.

8. National Center for Health Statistics, CDC. NHANES - National Health and Nutrition Examination Survey Homepage [Internet]. 2020 [cited 2020 Sep 24]. Available from: https://www.cdc.gov/nchs/nhanes/index.htm

9. Newell DB, Tiesinga E. The international system of units (SI):: 2019 edition [Internet]. Gaithersburg, MD: National Institute of Standards and Technology; 2019 Aug p. NIST SP 330-2019. Report No.: NIST SP 330-2019. Available from: https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.330-2019.pdf

10. Thurnham DI, Northrop-Clewes CA, Knowles J. The Use of Adjustment Factors to Address the Impact of Inflammation on Vitamin A and Iron Status in Humans. The Journal of Nutrition. 2015;145:1137S-1143S.

11. Nordenberg D. The Effect of Cigarette Smoking on Hemoglobin Levels and Anemia Screening. JAMA. 1990;264:1556.

12. Windsor JS, Rodway GW. Heights and haematology: the story of haemoglobin at altitude. Postgraduate Medical Journal. 2007;83:148–51.

13. Hess SY, Peerson JM, King JC, Brown KH. Use of Serum Zinc Concentration as an Indicator of Population Zinc Status. Food Nutr Bull. 2007;28:S403–29.

14. Geissler C. Capacity building in public health nutrition. Proc Nutr Soc. 2015;74:430–6.

15. Peng R. The reproducibility crisis in science: A statistical counterattack. Significance. 2015;12:30–2.

16. Osendarp SJM, Brown KH, Neufeld LM, Udomkesmalee E, Moore SE. The double burden of malnutrition—further perspective. The Lancet. 2020;396:813.

17. Davis JN, Oaks BM, Engle-Stone R. The Double Burden of Malnutrition: A Systematic Review of Operational Definitions. Current Developments in Nutrition. 2020;4:nzaa127.

18. Luo H, Zyba SJ, Webb P. Measuring malnutrition in all its forms: An update of the net state of nutrition index to track the global burden of malnutrition at country level. Global Food Security. 2020;26:100453.

19. Laillou A, Yakes E, Le TH, Wieringa FT, Le BM, Moench-Pfanner R, Berger J. Intra-Individual Double Burden of Overweight and Micronutrient Deficiencies among Vietnamese Women. Alemany M, editor. PLoS ONE. 2014;9:e110499.

20. Development Initiatives. 2018 Global Nutrition Report: Shining a light to spur action on nutrition. Bristol, UK; 2018.

21. Piwoz E, Rawat R, Fracassi P, Kim D. Strengthening the Nutrition Data Value Chain for Accountability and Action. Sight and Life. 2019;33:6.

1. GNR, Global Nutrition Report; IS, International System of Units; NCD, Non-Communicable Diseases; NHANES, National Health and Nutrition Examination Survey; SAMBA, Statistical Apparatus of Micronutrient Biomarker Analysis; [↑](#footnote-ref-1)
2. This analysis was supported by award [↑](#footnote-ref-2)
3. Conflict of Interest and Funding Disclosure:XXX, XXX, XXX: No conflicts of interest. [↑](#footnote-ref-3)