Gender data gaps: structural equation modeling offers an alternative to collecting more data

Sarah E. Oerthera,*

aSchool of Nursing, St. Louis University, USA

Abstract

An enhanced commitment to achieving gender equality, women’s rights and women’s empowerment in the post-2015 development framework is important. Gender data gaps exist because gender biases have inhibited and altered data collection resulting in policy formation that has insufficiently addressed gender issues. Structural Equation Modelling (SEM) is an effective statistical tool used to estimate correlational relations using limited input data confined to a hypothetical, researcher-specified framework of relationship. This paper highlights the use of SEM to evaluate progress on gender equality and women’s empowerment during the period of the first 1,000 days of life (from pregnancy through 2 years of age for infants).

Keywords: data gender gaps; structural equation modelling; post-2015 development

1. Introduction

The United Nations Millennium Development Goals (MDGs) are eight global targets, aimed to be achieved by 2015, with the ultimate purpose of reducing global poverty, chronic disease, and stalled development. Identified by the United Nations General Assembly in 2000, and promoted by the UN for 15 years, the MDGs focus on poverty and hunger, education, gender equality, child and maternal health, combating HIV/AIDS, protecting the environment and promoting global partnerships. The target date for achieving the MDGs is rapidly approaching, and the United Nations Entity for Gender Equality and the Empowerment of Women has called for a reaffirmation of the
commitment to achieving gender equality, women’s rights and women’s empowerment in the post-2015 development framework and Sustainable Development Goals (SDGs) [1].

Since the establishment of the MDGs in 2000, progress has been made on improving the abundance of, quality of and approaches to collecting data to support policy creation as well as to track progress towards achieving the MDGs. Moving forward in the post 2015 agenda, this improved data stream needs to be evaluated and used to help establish priorities for the advancement of statistical infrastructure for gender equality [9]. This paper will identify existing data gaps on gender equality, discuss statistical models that can generate new insight, and share knowledge on new methodological approaches to address complex areas.

Gender data gaps exist because gender biases have inhibited and altered data collection (i.e., because males are predominant in many work places, surveys that collect socio-economic information targeting work places are disproportionately focused on male responses) [1]. To understand impediments to achieving gender equality, data gaps that obfuscate important trends relevant to women and young girls needs to be filled. Currently very little is understood about the lives of females and the challenges they encounter because data disaggregated by gender is limited, especially in developing countries [1]. Different methods of collecting and analyzing “big data” could potentially assist in closing existing global gender gaps. This data would make it possible to define and measure priorities to help establish policies promoting gender equality as well as to monitor progress towards achieving goals. Also, improved data collection can inform improved models (i.e., decision support tools) to help to determine the size and nature of social and economic problems and opportunities as well as the efficacy and cost-effectiveness of alternative policies [9].

There are a wide variety of questions that need to be explored. For example, how can extension offices in developing countries work with female farmers because a disproportionate fraction of males are migrating away from the smallholder farms and moving to the urban setting in search of jobs? Another example of a gender specific question could involve evaluating the most cost effective means of delivering social safety nets, including money, to women by evaluating mobile payments or traditional cash transfers. A third question is to identify how many girls are forced into child marriage when adolescent development is underway and delayed onset of initial pregnancy could be a helpful policy [1]? Good data can provide important insight into these questions and contribute to closing the gender gap in development interventions [9].

2. Methods

While improvements in data collection and an increase in the total amounts of data collected are important aspects of this problem, data without analysis is of limited utility. Therefore, an iterative process of evaluating existing data using improved statistical tools and collecting additional data would help to ensure that data collection targets information that serves as input to analyses that inform effective policy decisions. Structural Equation Modelling (SEM) has been shown to be a useful statistical tool to evaluate social, economic, and health data. SEM is a statistical technique used to estimate correlational relations using data confined to a hypothetical, researcher-specified framework of relationship.

For example, in the health field an example of a project includes the use of SEM to examine the relationship between increased sugar consumption in developing countries and the prevalence of health care providers/infrastructure that is available to provide caesareans and related care for pregnant women. In developing countries issues that are not observed in developed countries complicate the performance of caesarean section. This is because the caesarean operation is often technically more difficult to perform due to lack of resources in developing countries. Developing countries have a shortage of resources, such as sufficient, affordable, appropriate health care infrastructure and trained health care providers [3].

Research strongly suggests that the “Western diet” (i.e., high sugar and high fat) is causing people all around the globe to gain weight and to suffer from chronic, non-communicable disease (NCDs) including overweight, obese, and metabolic syndrome [4]. As developing countries replace traditional foods with modern, processed foods high in sugar, refined flour and vegetable oils, the populations of developing countries increasingly suffer from NCDs [4]. The central hypothesis to be explored through SEM is, “There is a looming public health disaster that will occur when sugar consumption increases in developing countries, and health care providers/infrastructure is unavailable to provided caesareans and related care.”
To avoid an epidemic wherein “big babies” increase the risk of morbidity and mortality associated with labour and delivery problems, effective country-specific policies need to be developed and implemented through efforts such as the Scaling Up Nutrition Movement (SUN). Evaluating potential solutions to this looming epidemic is exacerbated due to a paucity of data on the nutritional status of women, young girls, and infants in many developing countries during the critical 1,000 days of life (from pregnancy through 2 years of age) as well as an overall lack of political focus on the issue of nutrition and obesity in the populations of many developing countries (i.e., see the Rome Declaration on Nutrition and the accompanying Framework for Action adopted by global consensus at the Second International Conference on Nutrition in Rome in November, 2014).

The goal of this ongoing study is to measure the correlation of sugar consumption and the size of babies and the prevalence of births by cesareans using women in Mumbai, India as an initial target population to represent the global community of women and young girls. The results of this study will illustrate if sugar consumption during pregnancy has a significant, positive effect on the birth weight of babies and can increase the prevalence of cesareans. Additionally these effects of sugar on the size of babies and the prevalence of cesareans are being correlated to the increase risk for maternal mortality and morbidity due to lack of health care providers/infrastructure.

Due to the complexity of the system under study and to be able to test the hypothesis, “there is a looming public health disaster that will occur when sugar consumption increases in developing countries, and health care providers/infrastructure is unavailable to provided cesareans and related care”, the following is being evaluated and the results are being analyzed collectively, namely: (1) cesarean sections are unavailable now and will remain unavailable for the foreseeable future, (2) there is low demand for cesareans sections currently, (3) birth weight of babies is the primary cause for cesarean sections, (4) the current distribution of baby birth weights, (5) sugar intake is positively correlated to babies’ birth weight, and (6) sugar intake is increasing.

SEM is being used to analyze all of the data collected in the study. Specifically SEM is being used to explore the relationship between sugar intake and increased risk for cesarean. Cesarean are assumed to correlate with the occurrence of larger babies; hence birth weight of babies is used as a directly observed measure while sugar intake is being assessed indirectly as a latent variable.

In general, an SEM approach follows two-steps, namely the creation of a hypothetical model, followed by data and testing of a confirmatory model. A measurement model and a structural model are two parts of the full model [7]. The measurement model describes the relationships between the latent variables (i.e., hypothesized variables) and the observable indicator variables (i.e., directly observable variables). Using Confirmatory Factor Analysis (CFA), the hypothesized model is compared to a data driven model using several fit indices [2]. Specifically, model-data fit is assessed using Root Mean Square Error of Approximation (RMSEA) < .06, Comparative Fit Index (CFI) > .95, or Tucker-Lewis Index (TLI) > .95 as guides. If the measurement model fits the data via several tests of model fit indices, then the structural model will be interpreted to be valid. If the measurement model is rejected, then the structural model will be revised based upon improved theory, prior and research by other investigators, the error matrix, and modification indices [2].

For users that are not trained in SEM, such as nurses, a graphical model is prepared that shows latent variables as rectangles and observed variables as ovals. Lines are used to shown correlational relationships, and arrows are used to show the direction of correlation (i.e., hypothesized causation). A benefit of SEM is that it is a simple application tool and the path analysis provides a useful interface to translate user understanding into a testable statistical framework. It is expected that the raw data could contain missing values and outliers that could adversely affect the SEM results. It is important to explicitly state how problems with raw data are dealt with. For example, EQS, a type of software, allows the deletion from the analysis of apparent outliers as judged by their contribution to Mardia’s coefficient of multivariate kurtosis [6]. If large departures from normality are detected, then other methods for efficient parameter estimation must be considered for analysis. However, for these alternatives there are practical limitations. To satisfy the asymptotic theory underlying the SEM approach, relatively large samples, even larger samples than needed for normality, could be required as a solution to this problem [6]. Thus, while the current study is exploring SEM as an alternative to reduce the amount of data needed to be collected to evaluate gender inequity, there are substantial limitations that should to be considered.
Although the SEM methodology has become an important tool in applied multivariate analysis for theory testing and correlational modelling, it requires appropriate data analysis [8]. Even an appropriate use, however, does not necessarily imply that correct correlational inferences will be drawn from the results of an SEM-based analysis alone [2]. Such inferences typically require additional assumptions concerning the context of a study and its data [6].

3. Results

While a limited degree of randomization is being utilized, the samples of convenience being collected in the ongoing follow up study are likely prone to unidentified bias [2]. SEM was selected to help reduce respondent bias, yet recall and other forms of bias will persist. A major constraint is the selection of women in Mumbai who are visiting health clinics. This sample is anticipated to automatically produce a skewed result through self selection of women to visit a clinic rather than to receive care within the home or through an alternative means of homeopathic, traditional, or familial care.

Based upon prior experience with the target population reported at HumTech2014, a concern is that the lack of cesarean within India currently may require a much larger sample size than would be needed in a more developed country (i.e., the approximation of a sample size of 200 for SEM may need to be expanded due to low levels of cesarean) [2, 5]. Alternatively, because the sample is being collected at a Postnatal Clinic, it is anticipated that substantial bias towards higher socio economic status could result in a sufficient number of observed cesarean.

Initial results suggest that overall diet could be a confounder (i.e. sugar consumption may co-vary along with other negative behaviors or masking behaviors). In addition, maternal diet may vary during the 10 months of gestation, and it is unclear if there is a link between sugar update at a particular stage of pregnancy and cesarean. Also, sugar consumption could be confounded by differences in uptake including sugar present in liquid beverages, sugar present in prepared foods, and “hidden” sugar present as a substitute to increase flavor in “low fat” foods.

Acknowledgements

The author would like to thank Ms. Phalakshi Manjrekar of Hinduja Hospital, Mumbai for helpful discussions of this research.

References