STUDIES OF BIRTH WEIGHT AND INFANT MORTALITY IN INDIA

by

Rachel Margaret Whelan

BA, Yale University, 2002

MPH, Columbia University Mailman School of Public Health, MPH, 2005

Submitted to the Graduate Faculty of

Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2011

UNIVERSITY OF PITTSBURGH
GRADUATE SCHOOL OF PUBLIC HEALTH

This dissertation was presented

by

Rachel Margaret Whelan

It was defended on

May 4, 2011

and approved by

Dissertation Advisor:
Clareann H. Bunker, PhD
Associate Professor
Department of Epidemiology
Graduate School of Public Health
University of Pittsburgh

Committee Members:
Lisa M. Bodnar, PhD, MPH, RD
Assistant Professor
Department of Epidemiology
Graduate School of Public Health
University of Pittsburgh

Catherine L. Carlson Haggerty, PhD, MPH
Associate Professor
Department of Epidemiology
Graduate School of Public Health
University of Pittsburgh

Roslyn A. Stone, PhD
Associate Professor
Department of Biostatistics
Graduate School of Public Health
University of Pittsburgh

Gong Tang, PhD
Assistant Professor
Department of Biostatistics
Graduate School of Public Health
University of Pittsburgh

STUDIES OF BIRTH WEIGHT AND INFANT MORTALITY IN INDIA

Rachel Margaret Whelan, PhD

University of Pittsburgh, 2011

Background

Birth-weight and infant mortality are both important indicators of the health of populations. Unfortunately, these measures have not been studied to a great extent in India, where high rates of both birth-weight and infant mortality persist. This is, in part due to a dearth of quality data from India and lack of methods to adjust the existing data for digit preference ("heaping"). Beyond that, while there has been extensive study of the birth-weight to infant mortality relationship in the developed world, this topic has remained basically unexplored in India. In order to develop methods to reduce the high rates of IM in India and better understand the role of birth-weight as a determinant of infant mortality, population specific studies are needed.

Methodology

Using data from a cohort in rural South India, we developed a method of adjusting BW data to account for heaping. Using data from a nationally representative survey of all of India, and US vital statistics, we compared characteristics of the BW to IM relationship in India and the United States. Finally, we analyzed data from rural India to identify predictors of very small birth-weight and infant mortailty in that specific population.

Results

Our method of adjusting birth-weight data to account for heaping using modified statistical calibration and multiple imputation produced imputed birth-weight data sets that reduced heaping and preserved known associations. After comparing the US and India, we found that the relative contribution of birth-weight to infant mortality in India is reduced. We also found differences between the US and Indian birth-weight distributions and infant mortality curves. Finally, we determined that measures of sanitation and hygiene, acting as surrogates for infectious disease exposure, were significant predictors of both lower birth-weight and infant mortality in a rural population in South India.

Conclusions

While birth-weight and infant mortality have not been studied to a great extend in the developing world due to issues with data quality and sources, statistical methods can be used to address these issues. Being able to adjust birth-weight data and study it and infant mortality in an Indian population showed that characteristics of these measures are not the same across countries. Also, birth-weight may not be as important a factor in determining infant mortality in India as it is in the US. Therefore, finding other causes of infant mortality is extremely important to address this problem. To that end, we found that sanitation and hygiene are strongly associated with the high rates of infant mortality in a rural Indian population.

Public Health Significance

The methods developed here can be used and applied to study birth-weight data in other developing country populations. The comparison of the US and India highlighted the fact that current policies to reduce infant mortality in India may be misguided. Finally, our data suggest that an intervention to improve sanitation and hygiene in order to reduce infant mortality could be successful and could also be used in other populations with high rates of infant death.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF EQUATIONS

PREFACE

This document is the result of a wonderful collaboration over the last four years. I would be remiss if I did not acknowledge the contributions and support of Drew Feiner, Clareann H. Bunker, PS Reddy, Roslyn Stone, Lisa Bodnar, Catherine Haggerty, Jamie Eastman and my family. Thank you so much for all your contributions!

1.0 SPECIFIC AIMS

Birth-weight (BW) and infant mortality (IM) are two of the most studied variables in epidemiology and health research. A search of PubMed shows almost 70,000 references to BW, more than 80,000 references to IM and hundreds of articles are released each year that study these variables. IM is a key measure of a population's overall health, but it is a rare event and often difficult to study as an outcome. BW is the measure most closely associated with IM and is often used as a surrogate for this variable. However, for all that BW and IM have been studied so extensively, there are still holes in the literature, especially when it comes to studying these measures in the developing world. We will address three topics relating to BW and IM, including how to adjust heaping in BW data from India, how an Indian population compares to a US population in terms of characteristics of BW distributions and the BW-IM relationship, and predictors of IM in a rural Indian cohort.

BW data are an important source of information about the health of mothers and children in the developing world. However, few data sets exist that capture BWs that have been reliably measured and recorded from the developing world. Some of the biases seen in these types of data are from measurement error using crude tools like analog scales, operator bias when health care workers round the weight to the nearest 100 or 500 grams, or recall bias when mothers are asked to remember the BW of their baby.

Specific Aim 1: To develop a method to adjust the continuous variable BW to account for bias and heaping due to measurement error and operator error.

- *We hypothesize that, using a modified statistical calibration and multiple imputation, we will be able to adjust heaped BW data from Indian surveys and create new data-sets that are closer to the actual BW distribution in this population.*

While BW is often viewed as an outcome unto itself and correlates or causes of low BW (LBW) are the focus of research, one of the major values of BW lies in its extremely strong association with IM. BW is often treated as a categorical variable (Low BW <2.5kg and Normal BW ≥2.5kg) for purposes of analysis and policy. LBW babies have a reportedly 20 times higher risk of death than normal NBW babies. Unfortunately, due to many of the issues with collecting quality BW data from developing countries, our understanding of characteristics of BW distributions and the relationship between BW and IM comes primarily from studies that have been conducted in populations in developed countries. These studies have resulted in a set of accepted characteristics of both BW distributions and the BW – IM relationship. Our goal for this aim is to assess whether these characteristics are applicable to an Indian population.

Specific Aim 2: **To compare the BW distribution and the BW–IM relationship in an Indian population, from the Indian National Family Health Survey, to a US population, from US vital statistics.**

- *We hypothesize that the BW distributions from both India and the United States will be essentially normal distributions but with extended lower tails*
- *We hypothesize that the relationship between BW and IM in the Indian cohort will be different from the relationship seen in the US population and that, in particular, the relative contribution of BW to risk of infant death will be less in the Indian population*
- *We hypothesize that this analysis will provide a possible alternative cut-point for LBW and NBW than the arbitrarily derived 2.5kg cut-point for LBW*

Efforts in developed countries have succeeded in steeply reducing the rates of LBW and IM. These efforts focused on known correlates of BW and IM that were specific to these populations. With the methods described in Aims 1 and 2, we can now analyze data taken from a rural Indian population in Andhra Pradesh and assess correlates and potential causes of LBW and IM specific to this community. We can also determine whether these data support the placement of BW on the causal pathway to IM

Specific Aim 3: To identify correlates of BW and IM in this rural Indian population while applying the Modified Calibration-Multiple Imputation (MCMI) technique to adjust the BW **data in the REACH dataset**

- *We hypothesize that we will confirm the known association between BW and IM in this population*
- *We hypothesize that sanitation and hygiene measures, acting as surrogates for exposure to infectious disease, will be significantly associated with IM, independent of BW*
- *We also hypothesize that that these data and results will support the idea that BW, while related to IM, is not on the causal pathway to IM in this population in rural India*

2.0 BACKGROUND AND SIGNIFICANCE

2.1 SPECIFIC AIM 1: BACKGROUND AND SIGNIFICANCE

Birth-weight has become a very important variable for research and study, not only from a clinical or research perspective, but from a policy perspective as well. The United Nations has focused on BW in its "A World Fit for Children" program and the Millennium Development Goals.[1,2] Weight at birth is an indicator of a baby's chances for survival, growth, long-term health and development. Low BW, defined as <2500 grams, has been linked not only to infant death and childhood diseases, but to adult morbidity and mortality as well. [3-18]

Typically, BW data are very useful because BW is easily measured and reliably recorded.[19] However, this assertion is only true in developed countries where mothers deliver in institutions and a system is in place for collecting vital statistics. For developing countries where there is no nationwide system of data collection for BW or other vital statistics, BW data comes from various sources, many of them biased.[3,10] The main source of BW data is health care facilities. However, these data are subject to selection bias, because in many developing countries not all women give birth in an institution and those that do tend to be from a group that is not representative of the whole population.[20-22] Nationally representative household surveys are another source of information on BW. However, a paper by Boerma *et al.* from 1996 showed that mothers in developing countries are often unable to report numerical BWs for their babies, often because the baby was not weighed at delivery. Boerma *et al.* also found that rates of LBW were systematically under-estimated by these types of surveys.[21] A similar paper by Robles *et al.* supports these findings.[22]

A 2005 paper by Blanc and Wardlaw examined the quality of BW data from these types of surveys and identified biases in them. They also hoped to develop a method for adjusting BW data to account for some of these biases. They studied BW survey data from 62 sources in developing countries and found that numerical BW data exhibits "heaping" on digits that are multiples of 500 grams. "Heaping refers to a pattern of misreporting in which the distribution of

a number reported by respondents, such as age or BW, shows implausibly large frequencies of particular values, usually values ending in 0 or 5." This heaping effect was seen in all of the survey data and a typical example of it is shown Figure 2.1.[20]

2.1.1 Methods for adjusting BW data from developing countries

In response to the poor quality of BW data from developing countries, several methods have been proposed that could be used to adjust this data and glean some meaning from it. Boerma *et al.* proposed a method that took into account not only the reported numerical BW, but also the mother's answer to a question about the relative size of her baby. Many of the population surveys being used in developing countries included the following series of questions, or something similar:

- When (NAME) was born was he/she very large, larger than average, average, smaller than average, or very small?
- Was (NAME) weighed at birth?
- (IF YES) How much did (NAME) weigh?

Boerma and colleagues thought that combining the mother's response on relative weight with the reported numerical weight would allow them to better deduce the true proportion of LBW babies in a population. However, only the "very small" category had a high sensitivity and specificity for predicting LBW and many of the LBW babies were missed when looking at only the "very small" category.[21]

Blanc and Wardlaw also proposed an adjustment method to account for the heaping of babies at 500 gram intervals and, in particular, the heap at 2500 grams, which was throwing off the categorization of LBW and normal BW (NBW) babies. Their adjustment was based on 62 Department of Health surveys between 1990 and 2000 from 42 developing countries, which included altogether 433,967 recorded births. The end result was a recommendation that 25% of those babies heaped at 2500 grams be re-classified as LBW. This method involved counting the number of babies between 2000 and 2500 grams (non-inclusive) and the number of babies between 2500 and 3000 grams (non-inclusive). The authors suggested that one could make a ratio of 2000-2500 g babies over the total number of babies between 2000 and 3000 grams,

excluding those listed as exactly 2500 g. This ration is then multiplied by the number of babies listed as exactly 2500 grams. The resulting number of babies listed at exactly 2500 grams (classified as NBW) would be moved into the LBW category (see graphical representation of adjustment method in Figure 2.2).[20]

This method, while giving a correction for the under-reporting of LBW babies in a population, does not allow for any further analysis. Also, by focusing exclusively on the dichotomized version of LBW, they have lost much of the power of the continuous BW variable. Finally, when this method was applied to data from our cohort in India with a known gold standard, the 25% adjustment did not produce the correct proportion of LBW infants. The true proportion of LBW infants in our dataset, who were weighed using a digital scale by a trained research nurse, was 29.5%. When the Blanc and Wardlaw 25% adjustment was applied to a set of BW data with biases from measurement error and operator error, the proportion of LBW infants was only 26.4%. For this data set, the Blanc and Wardlaw method for adjusting biased BW data corrected only a portion of the mis-classified babies.[23]

Though not necessarily developed in relation to BW, other methods exist to adjust for bias. Fox and Lash developed an automated and elegant method referred to as probabalistic bias analysis. This is a multiple imputation method that uses known sensitivity and specificity of the data to impute new data sets that are corrected for misclassified categorical variables. While this method was appealing in that it produces imputed data sets that would be available for further analysis, the method of adjustment is designed for categorical variables and does not apply to a continuous variable that is biased or mis-measured.[24-26] Fox and Lash's use of multiple imputation after the adjustment is applicable to other data sets, including ours, because the reconstructed data set represents only one possible association that could have occurred after correcting for misclassification. The range of association coming from the combination of the imputed data sets comes closer to the true association, were there not error in the data.[24,25]

Various other methods exist to adjust data that have been mis-measured or are known to have biases. Carroll, Ruppert and Stefanski developed a method (the CRS method) whereby the "gold standard" measure is regressed on the mismeasured, "proxy" covariate as well as on

other covariates in the dataset that were not subject to measurement error. The regression model coefficients are then used to estimate the predicted value for the "gold standard" for each observation in the main study. Then, in the main study, the outcome is regressed on the estimated "gold standard" and on covariates measured without error. The measurement error model is then used to correct the coefficients and their variances for bias due to covariate measurement error. Standard errors are obtained from bootstrapping or other resampling methods, or asymptotic standard errors can be obtained using the sandwich method.[27-31]

Rosner, Willett and Spiegelman also developed a method (the RSW method) using statistical calibration to adjust for the bias resulting when one or more regression model covariates are measured with error. The RSW method regresses the outcome on the "proxy" covariates and on the covariates measured without error in the main study and then bias-corrects the regression coefficients using estimated coefficients from the validation data regression. They adjusted for measurement error by regressing the outcome on the "proxy" measure and then adjusting the variances. Standard errors from this method are found using the delta method.[30-35]

The CRS and RSW methods were shown to give identical adjustments for the coefficients and their asymptotic variance when the validation set was strictly internal validation data.[30] When looking at studies with "hybrid" validation data, where the validation set is a combination of internal and external validation data, the estimates are the same, but the RSW method is shown to have an asymptotically smaller variance than the other methods.[31]

There are several characteristics of these methods that are important to note. They assume that other covariates measured without error are available in both the main and validation studies. They use a classical type of statistical calibration, which assumes a normal error distribution. And, they have only been applied to covariates/predictors, not outcomes. These methods provided valuable direction for the development of our own adjustment method to correct for heaping in the outcome variable BW.

2.2 SPECIFIC AIM 2: BACKGROUND AND SIGNIFICANCE

2.2.1 The BW distribution

In large population studies conducted in Western countries, the frequency distribution of BW has been found to be Normal with an extended lower tail.[11,36-38] (See Figure 2.3) The "predominant" distribution has been defined as that part of the curve that falls within the Normal curve. The vast majority of births fall within this predominant distribution. In Figure 2.3 the predominant distribution is indicated by the area below the line with triangle markers. The "residual" distribution is the remainder of the BW distribution that falls outside of the Normal curve. In a typical population, 2 to 5% of births are in the residual distribution. In Figure 2.3, the residual distribution is indicated by square markers.

The births that fall in the predominant distribution correspond to the distribution of term births (≥37 weeks gestation). Logically then, almost all those births in the residual distribution are preterm. However, some of the preterm births still end up in the predominant distribution of BW. The preterm births that do end up in the residual distribution tend to be those that are the smallest and thus those with the highest risk of mortality.

The predominant and residual distributions of BW are independent of one another. Therefore, an exposure that affects fetal growth does not necessarily affect the risk of preterm delivery. Also, the mean of the predominant distribution can change without affecting the percent of births in the residual distribution. Conversely, a factor that increases the risk of preterm delivery would not necessarily change the average weight of babies that are delivered at term, greater than 37 weeks gestation. Thus, the percentage in the residual distribution can change without affecting the predominant distribution.[11,37-40]

This construct of looking at BW distributions from the perspective of the predominant and residual distributions is particularly useful because it provides indirect information about gestational age without actually requiring data on gestational age of each infant. Term births are found in the predominant distribution while the residual distribution estimates the

percentage of both small and preterm births.[11,37-41] Automated methods for calculating the predominant and residual distributions are now available through the NIH.[38,42]

2.2.2 BW and IM Relationship

BW predicts a newborn's survival better than any other characteristic that has been studied. In general, the lower weight at birth, the higher the risk of IM. This has also been observed on a population level where groups with lower mean BWs (i.e. smokers, African-Americans) often have higher rates of IM.[8,11,19,37,43-45]

Studies in large, developed countries typically show a very typical pattern of BW specific neonatal mortality. Risk approaches 100% for the smallest babies and declines to less than 1% in the middle of the range, which is close to the mean BW for that population. Risk increases slightly for the heaviest babies. The same pattern is seen within each gestational-age stratum, indicating that this is not simply a reflection of preterm births at the lowest weights. (Figure 2.4) [37,40,41]

While the association between BW and IM is certainly extremely strong (LBW babies are 20 times more likely to die than NBW babies[18]) BW is not the best indicator for risk of perinatal mortality. If it were, then variations in BW among populations would match changes in risk. So, an exposure that reduced fetal growth would be predicted to increase risk. However, this is not seen when comparing groups with and without these types of exposures. As the BW distribution shifts due to different population characteristics or exposures, the mortality curve moves with it. This phenomenon lead to the description of several "paradoxes" related to BW.[19]

2.2.3 Birth-weight Paradoxes

When studying BW in populations, several paradoxes have been described in the literature. An excess of LBW in a population does not necessarily mean higher rates of IM in that population. Another paradox of the BW – IM relationship is that if one compares LBW babies from two different populations, it is often the case that the LBW babies with a lower risk

of mortality are from the group that has higher mortality overall. This was noted among smoking mothers, who have babies with higher mortality overall, but the LBW babies of smoking mothers have better survival than the LBW babies of non-smoking mothers. [11,19,41,46-49]

Based on these paradoxes and other studies of the BW to IM relationship, Allen J. Wilcox and Ian T. Russell developed a hypothesis to explain this paradoxical relationship and the true nature of the interplay between BW and IM.[50] They debunked the usual assumption that a change in BW directly affects perinatal survival. Instead, they hypothesized that, on a population level, BW is not on the causal pathway to mortality. They acknowledged that a change in BW is often strongly associated with a change in infant health. However, they postulated that it is not through the change in BW that the change in health (or increase in risk of IM) occurs. Instead, they suggested that that BW and IM can change together because a single factor (or group of factors) affects them both.

To support this theory, they studied many examples of populations where IM rates are similar but the BW distributions are slightly different. For instance, at high altitudes, babies are smaller and the BW distribution of a high altitude population is shifted to the left (lower mean BW). They found that, when comparing a high altitude population to a normal altitude population, the weight-specific IM curves intersect. For babies weighing less than the optimum weight, this shift gives the appearance of lower mortality at any given BW. For babies heavier than the optimum weight, the shift gives the appearance of higher mortality. In other words, mortality rates are lower at high altitudes for small babies and higher for large babies. There is no obvious biological explanation for this.

However, using a method of adjusting both the BW distribution and the IM curve to a z-scale with mean of zero and standard deviation of one, the two BW distributions correspond nearly exactly, as do the two mortality curves. The explanation for this is that altitude affects BW, but not mortality. While the BW distribution and the accompanying mortality curve shifts for populations at higher altitudes, there is no change in the survival of individual babies. In this example, lower BW on the population level has no effect on IM.

This method of comparing BW and IM at the population level using the z-scale adjustment has been applied to many populations in the developed world, but not to

developing world populations to see if these theories are supported. In fact, Wilcox acknowledges the gap in the literature when it comes to this type of reproductive epidemiology and the developing world:

> "…my perspective (and the perspective of most researchers in this field) is overwhelmingly First World. One of the ironies of epidemiology is that our data are best where the problems are least. This is certainly true for reproductive epidemiology. Those of us privileged enough to work in developed countries tend to focus on the small problems close to us and neglect the huge problems elsewhere…"[50]

2.2.4 Low Birth-Weight and 2500 grams

At the First World Health Assembly in 1948, the World Health Organization endorsed the international definition of prematurity as being a BW less than 2500 grams. Prematurity and LBW were considered to be synonymous until 1961 when the WHO recommended that LBW no longer be used as the definition of prematurity. [10,11] However, this tendency to dichotomize BW has persisted with the understanding that babies born less than 2500 grams are a higher risk for mortality than babies 2500 grams and larger.

2.3 SPECIFIC AIM 3: BACKGROUND AND SIGNIFICANCE

Rates of IM are extremely high in India. In 2009, it was 50.3 per 1,000 live births, more than seven times the IM rate in the United States (6.8/1000) and more than 20 times the IM rate in Sweden (2.3/1000).[51] Infant deaths in India account for a very large portion of the total infant deaths world-wide and India contributes a quarter of the world total for newborn deaths.[52] Despite efforts to reduce the numbers of infant deaths through social welfare programs and other interventions,[53] IM rates remain unacceptably high in India.

IM has been studied in many different populations and methods have been developed that have successfully reduced the numbers of infant deaths in certain populations, primarily in developed countries.[54-56] Globally, rates of child mortality are half of what they once were several decades ago. However, the rate of IM remains high in India and, in recent years, there has been a slowing decline in IM rates.[57]

A few studies have attempted to identify causes of IM in India. A mortality survey by Bassani, *et al* identified three causes that accounted for 78% of all neonatal deaths: prematurity/LBW, neonatal infections and birth asphyxia/birth trauma. For children ages 1 to 59 months, they found that pneumonia and diarrheal diseases accounted for half the deaths.[58] A report from the Indian National Family Health Survey 2 (1992-1993) identifies SES and demographic factors like mother's age, birth order, literacy, toilet facilities, rural residence and the use of unclean cooking fuel as being associated with higher risk of infant death.[59] A recent literature review of causes of child death in India supports these findings.[60]

Any study of IM must focus to some extent on BW as it is the strongest recorded predictor of IM. Babies weighing less than 1,500 g have a mortality risk at least 100-fold higher than babies at the optimum weight (the weight associated with the lowest mortality). This association is seen at both the individual and population level. BW is extremely predictive of an individual baby's survival. Also, groups with lower mean BW often have higher IM.[11,41,48] However, the true nature of the relationship between BW and IM is controversial. Researchers are divided over whether or not IM is caused directly by BW and if so, what the biological mechanism is. This is a particularly important question in India where efforts to reduce IM have focused on increasing BW.[57,61,62] However, this approach, which assumes that risk of IM is programmed from birth (in the infant's BW), does not address insults after the baby is born that could be contributing significantly to IM. Researchers have suggested that the same causal factors are affecting both BW and IM, and that this explains the strong association between BW and IM.

LBW, defined as less than 2500 grams, is a complex phenomenon with multiple causes and correlates. Some of these have been studied and addressed, such as quality of pre-natal care,[63-67] maternal nutrition,[68,69] anemia,[70-73] maternal thyroid function,[74-77] and maternal

infections.[69,78-84] In India, studies have been completed that assess factors associated with BW as well. These studies have identified socioeconomic factors,[85] maternal nutrition,[6,85-91] maternal activity,[6,92] anemia,[85,86] and parity,[85] among others, as being associated with risk for having a LBW baby. Each year, approximately 11.7 million babies in developing countries are born with LBW.[93] In India, approximately 30% of babies are born with LBW.[94]

With so many possible causes of and contributors to IM, it is clear that a single solution is not going to fix this problem across the developing world or even across India. However, identifying population specific predictors of IM will provide an opportunity for interventions at a local level to combat these problems of infant health and survival in an effective way.

2.3.1 The Rural Effective Affordable Comprehensive Healthcare (REACH) Project

The REACH Project is a working model of healthcare delivery to rural populations that has been providing health education, immunizations, antenatal care, and primary and tertiary care since 1999 for a population of 43,270 people in 43 villages in the Ranga Reddy District on the northern outskirts of Hyderabad in the state of Andhra Pradesh in South India. In each of the villages in the mandal (county), a Community Health Volunteer (CHV) has been recruited to visit each family one time each month. These CHVs focus especially on fertile woman in the village to ascertain pregnancy (by interview) and to educate and encourage the women to seek regular antenatal care, and other health care services. REACH has enumerated all households and household members in these communities and mapped each dwelling by GIS. With each visit, CHVs conduct interviews to collect and update information on demographics, household details, whether a woman is pregnant and receiving ante-natal care and other variables. Since 2004, CHVs have been collecting data on infant deaths and BWs in the population. A detailed database is maintained that contains all these data.

2.4 TABLES AND FIGURES

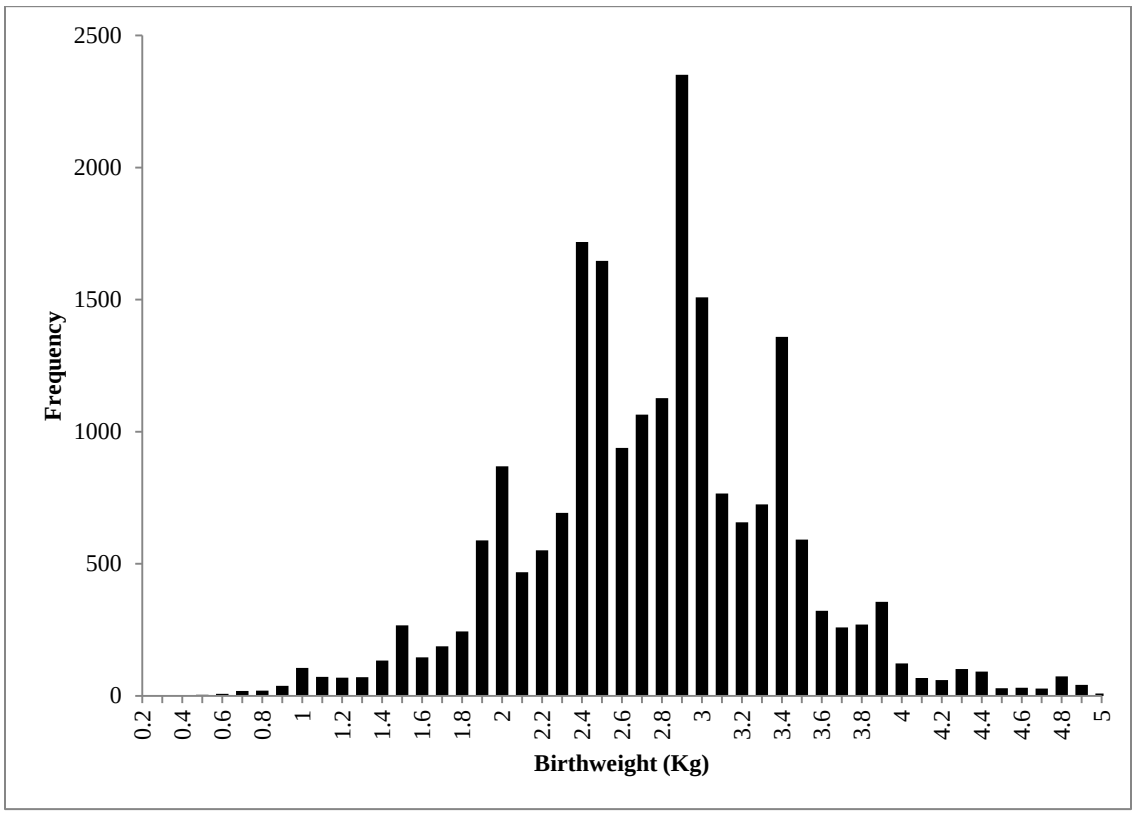

Figure 2.1: Heaping of reported birth-weights on multiples of .5 kg, Indian National Family Health Survey 3 2005-2006

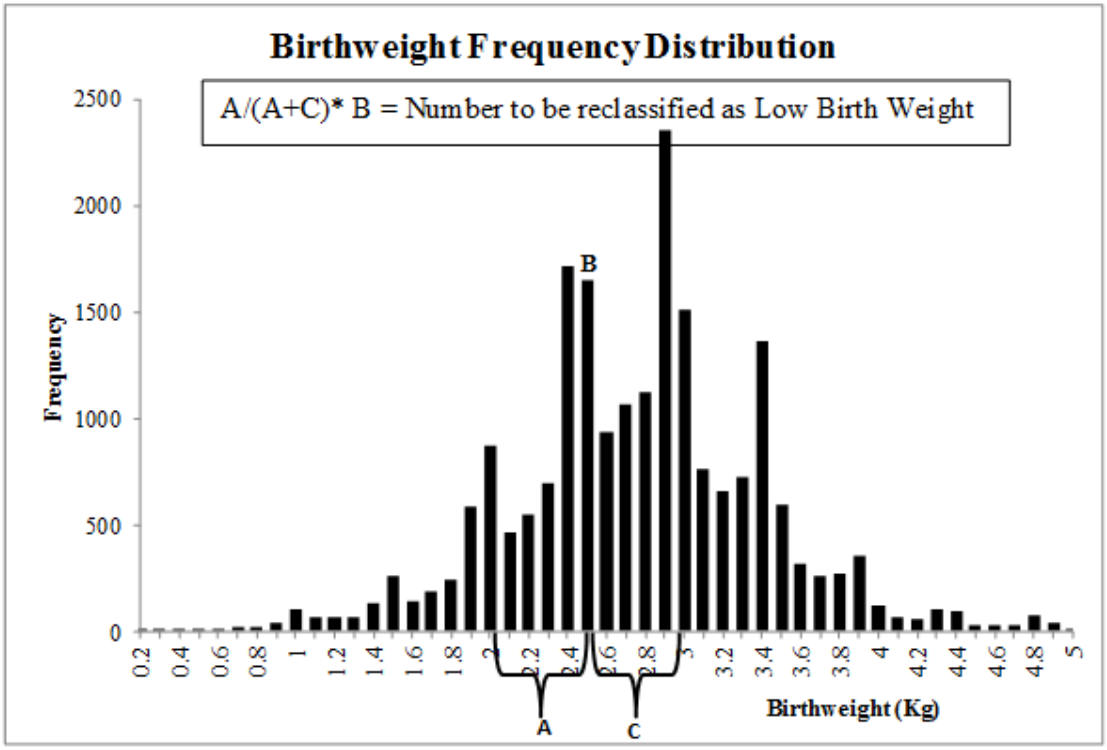

Figure 2.2: Application of Blanc and Wardlaw's method for adjustment of heaped birth-weights

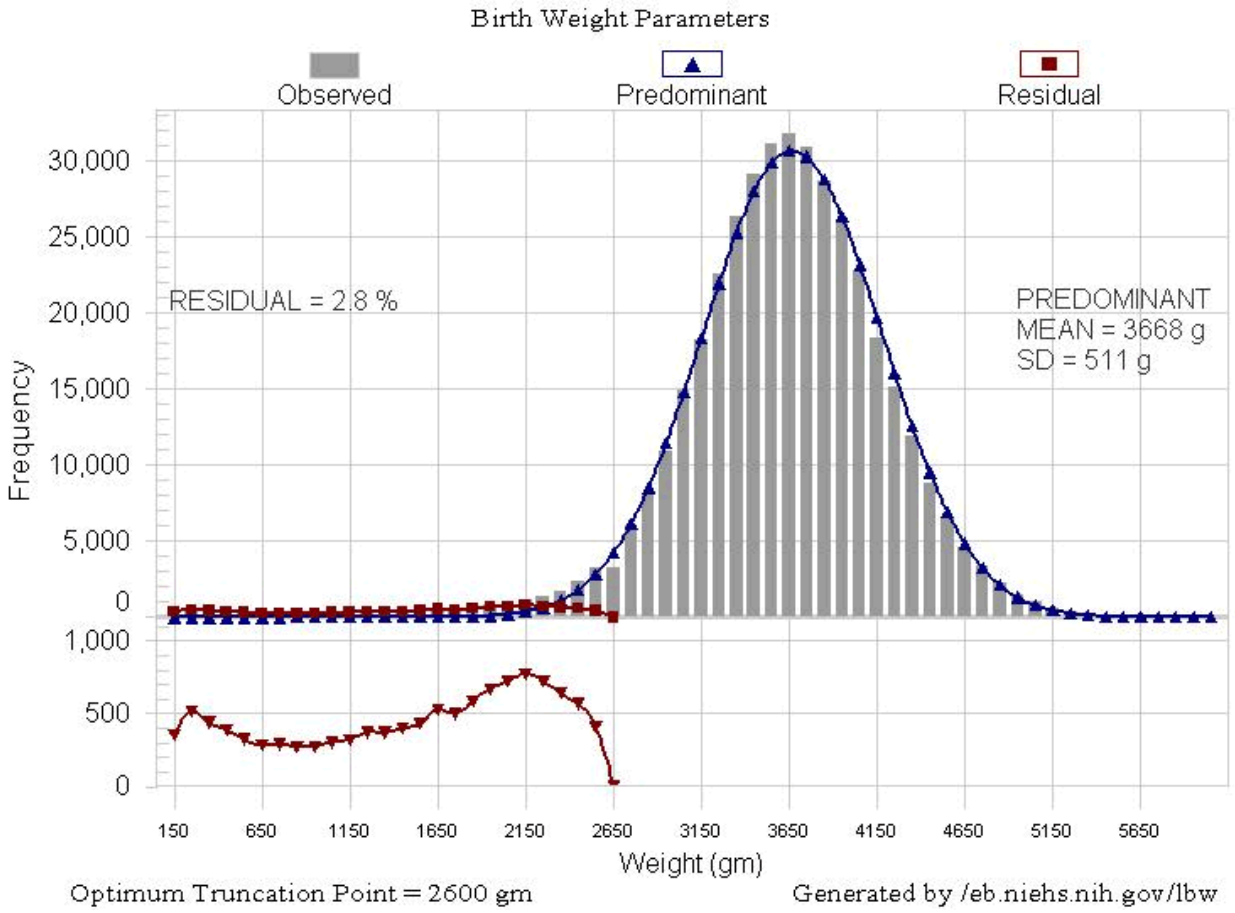

Figure 2.3 Birth-weight distribution with Predominant and Residual distributions

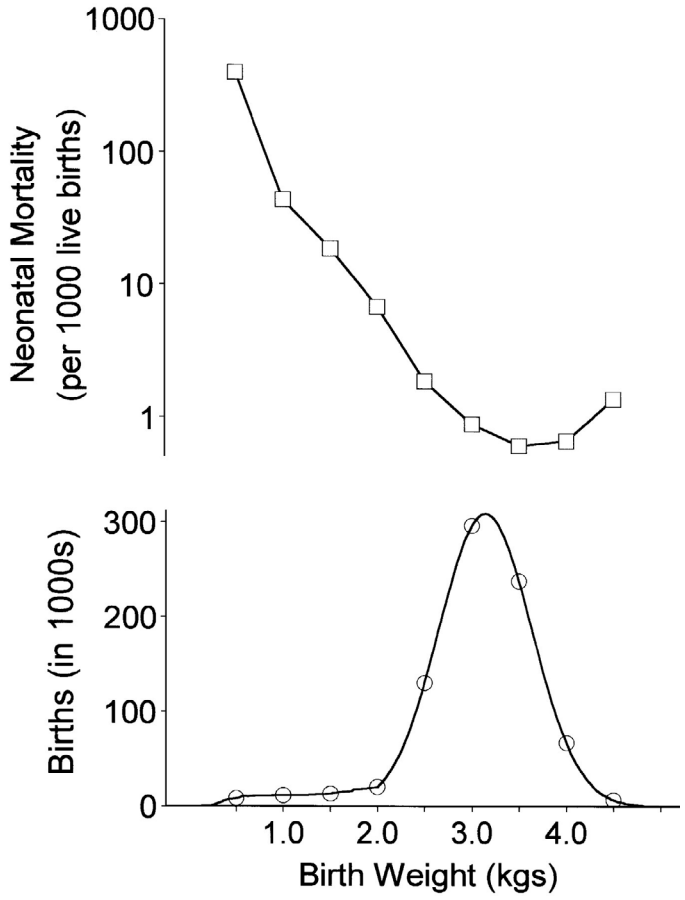

Figure 2.4: Weight-specific neonatal mortality and the distribution of weights for live births, USA, 1998[11]

3.0 MANUSCRIPT 1: ADJUSTING BIRTH-WEIGHT DATA FROM DEVELOPING COUNTRIES TO ACCOUNT FOR HEAPING: A NOVEL METHOD

Manuscript in Preparation

R. Margaret Whelan,* Drew Feiner,† C. Rekha,◊ Purushotham Reddy,◊ Clareann H. Bunker,* PS Reddy, ◊ Roslyn Stone‡
* University of Pittsburgh, Graduate School of Public Health, Department of Epidemiology
† Blade Diagnostics Corporation, Pittsburgh, PA
◊ MediCiti Institute of Medical Sciences, SHARE India, Andhra Pradesh, India
‡ University of Pittsburgh, Graduate School of Public Health, Department of Biostatistics

3.1 ABSTRACT

3.1.1 Background

Birth-weight (BW) is an important indicator of a baby's chances for survival, growth, and development. Few data sets capture BWs that have been reliably measured and recorded from the developing world. Biases in these data include measurement error, operator error, recall bias and heaping. No adequate method exists to correct these data and estimate accurate rates of low BW (LBW), defined as less than 2500 g. Our goal was to develop such a method.

3.1.2 Methodology

From October 2009 to May 2010, we weighed every baby born at MediCiti Hospital, Andhra Pradesh, India (n=859). Each baby was weighed by the usual method (analog scale, labor and delivery nurse), and then weighed by the gold standard (10-gram sensitive digital scale, trained research nurse). Head circumference (HC) and baby length (BL) were recorded.

We developed a method to correct the analog BW data using a modified statistical calibration and multiple imputation. Calibration is based on analog values from the calibration

data set. Multiple imputed data sets were combined using Rubin's Rules. Data were analyzed using Microsoft Excel, SAS 9.2 and MATLAB 7.11 R2010B.

3.1.3 Results

Rates of LBW for the original and calibrated/imputed data were analog (22%), digital (28%), and imputed (30% on average with a range of 28% - 33%).

Using logistic regression, we estimated the associations between BW and BL, and BW and HC. The association between the analog BW data and these variables differs from that of the digital BW data. The results from the calibrated/imputed data most closely matched the digital BW data (gold standard).

3.1.4 Conclusions

Regression calibration plus imputation produces adjusted analog weight data that accurately reflect the gold standard. This approach could be applied to other data sets in India, or other parts of the developing world, to more accurately estimate rates of LBW ascertained from analog weights.

3.2 INTRODUCTION

Birth-weight (BW) has become a very important variable for research and study, not only from a clinical or research perspective, but from a policy perspective as well. The United Nations has focused on BW in its "A World Fit for Children" program and the Millennium Development Goals.[1,2] Weight at birth is an indicator of a baby's chances for survival, growth, long-term health and development. Low BW (LBW) has been linked not only to infant death and childhood diseases, but to adult morbidity and mortality as well. [3-18]

Typically, BW data are very useful because BW is easily measured and reliably recorded.[19] However, this assertion is only true in developed countries where mothers deliver in institutions and there is a system in place for collecting vital statistics. In developing countries, where there is no nationwide system of data collection for BW or other vital

statistics, BW data come from various sources, many of them biased.[3,10] A common form of error seen in BW data from surveys or health facilities in developing countries is "heaping" or digit preference. "Heaping refers to a pattern of misreporting in which the distribution of a number reported by respondents, such as age or BW, shows implausibly large frequencies of particular values, usually values ending in 0 or 5."[20] In a study of 62 surveys from 41 developing countries, this heaping effect of BW data was seen throughout and a typical example of it is shown in Figure 3.1.[20]

3.2.1 Current methods for adjusting BW data from developing countries

In response to the poor quality of BW data from developing countries, several methods have been proposed to adjust the data. Boerma *et al.* proposed a method that took into account not only the reported numerical BW, but also the mother's answer to a question about the relative size of her baby. However, this method underestimates the proportion of LBW babies.[21]

Blanc and Wardlaw also proposed an adjustment method to account for heaping, in particular, the heap at 2500 grams, which alters the categorization of LBW and normal BW (NBW) babies, where NBW is defined as ≥ 2500 g. The end result was a recommendation that 25% of those babies heaped at 2500 grams be re-classified as LBW. [20] This method, while giving a correction for the under-reporting of LBW babies in a population, does not allow for any further analysis. Also, by focusing exclusively on the dichotomized version of LBW, much of the power of the continuous BW variable is lost. Finally, when this method was applied to data from our cohort in India with a known gold standard, the 25% adjustment did not produce an accurate estimate of the proportion of LBW infants.[23]

Though not necessarily developed in relation to BW, other methods exist to adjust for bias in a dataset. Fox and Lash developed probabalistic bias analysis, a multiple imputation method that uses known sensitivity and specificity to impute new data sets that are corrected for misclassified categorical variables. While this method is appealing in that it results in imputed data sets that would be available for further analysis, the method is designed for

categorical variables and does not apply to a continuous variable that is biased or mis-measured.[24-26]

Various other methods also exist to adjust data that have been mis-measured or are known to have biases. Carroll, Ruppert and Stefanski developed a method (CRS method) whereby the "gold standard" measure is regressed on the mismeasured, "proxy" covariate as well as on covariates that were not subject to measurement error. The regression coefficients are then used to estimate the predicted value for the "gold standard" for each observation in the main study. Then, in the main study, the outcome is regressed on the estimated "gold standard" and on covariates measured without error. The measurement error model is then used to correct the coefficients and their variances for bias due to covariate measurement error. Standard errors are obtained from bootstrapping or other resampling methods, or asymptotic standard errors can be obtained using a sandwich estimator.[27-31]

Rosner, Willett and Spiegelman also developed a method (the RSW method) using statistical calibration to adjust for the bias resulting when one or more regression model covariates are measured with error. The RSW method regresses the outcome on the "proxy" covariates and on the covariates measured without error in the main study, and then bias-corrects the regression coefficients using estimated coefficients from the validation data regression. They adjusted for measurement error by regressing the outcome on the "proxy" measure and then adjusting the variances. Standard errors from this method are found using the delta method.[30-35]

The CRS and RSW methods have been shown to give identical adjustments for the coefficients and their asymptotic variances when the validation set was strictly internal validation data.[30] In studies with "hybrid" validation data, where the validation set is a combination of internal and external validation data, the estimates are the same but the RSW method is shown to have an asymptotically smaller variance than the other methods.[31]

There are several features of these methods that are important to note. They assume that other covariates measured without error are available in both the main and validation studies. They use a classical type of statistical calibration, which assumes a normal error distribution. And, they have only been applied to covariates/predictors, not outcomes. These

methods provided valuable direction for the development of our own adjust method to correct for biases in the outcome variable BW.

3.3 EXPERIMENTAL DESIGN AND METHODS

3.3.1 The Rekha Data Set

From October 2009 through February 2010, data were collected from every birth that took place at MediCiti Hospital, a rural hospital in Andhra Pradesh, India. Each baby was weighed as usual by a labor and delivery nurse on a 50-gram graduated analog scale and their BWs were recorded. These "analog birth-weights" have measurement error from the inexact analog scale and also error from the nurses who weigh the babies and tend to round the weights to whole numbers. Second, each baby was weighed by a trained research nurse on a SECA 354 10-gram sensitive digital scale and their weights were recorded. These "digital birth-weights" are considered to be the true BWs. We now have a data set that includes 913 births and 859 sets of complete data where both the analog and digital weights were recorded.

MediCiti Hospital serves two counties (mandals) in rural Andhra Pradesh: Medchal Mandal and Shameerpet Mandal. The two counties have similar populations (about 50,000 people in each) and the births that take place in the hospital are divided between women from each of the two counties.

The analog BW data collected by Dr. Rekha per standard procedure by obstetric ward nurses on the analog scale is shown in Figure 3.2. This is a standard method of recording BW in hospitals in developing countries like India and is subject to measurement and information bias because of the use of analog scales and reader error/rounding that leads to heaping at round numbers. After being weighed by the standard method, each baby was weighed on a ten-gram sensitive digital scale by a trained research nurse. This is considered to be the gold standard. (Figure 3.3)

3.3.2 Method to adjust the analog birth-weight data using the digital birth-weight data as a gold standard

By plotting the digital versus the analog weights, one can begin to get a sense of the measurement and information error present in BW data collected by the standard method, as compared to the gold standard (Figure 3.4). The diagonal dotted line shows where all the points would be if the analog and digital weights were equal. Distance from the line indicates error. This figure also shows the misclassification that would occur from this type of measurement error if the weights were classified dichotomously as LBW and normal BW (NBW). The lower right quadrant holds the weights of all the babies that would be misclassified as LBW using the standard measuring technique when they were actually NBW by the gold standard. The upper left quadrant holds the weights of the babies that would be classified as NBW when they were actually LBW.

Figure 3.4 also illustrates the heaping that occurs at 500 g intervals from rounding and other operator error using the standard measuring technique with the analog scale. The number and spread of points on the 2.0, 2.5 and 3.0 marks from the analog scale show the number of BWs that are heaped on these round numbers as well as the spread of true digital BWs these rounded analog weights actually represent.

Using these data, our goal is to develop a model that relates the analog weights to the digital weights. This model can then be used to impute a dataset with statistics that more accurately represent the population. Once developed, this method can be applied to larger datasets from the same or similar populations to adjust for bias due to heaping.

The approach we have chosen to achieve this is to use a modified statistical calibration followed by multiple imputation. The calibration will allow us to adjust the analog data based on the known true BWs. Multiple imputation will allow us to appropriately impute new data sets and then perform regression analysis on these sets.

3.3.3 Statistical Calibration

Statistical calibration, often referred to as Inverse Regression, is a method used to estimate the value of one measurement (x) by some other measurement(s) (y) using a

regression model. This method is used when the quantity to be calibrated is harder, or more expensive to measure or when the value was not recorded and cannot be retrieved.[95] In our case, *x* represents the true digital weight that we wish to estimate based on the available analog weight, *y*. Here, all three reasons for using statistical calibration apply. The true BWs from this population are both harder and more expensive to measure, not to mention logistically very difficult to collect. Also, in national surveys from India, true BWs are often not recorded; instead, a mother may be asked the weight of her child at birth, or the weight may be copied off of a medical record that a mother has in the home.

Absolute calibration is a form of this method wherein *x* is assumed to be measured without error. In our study, we are assuming that the recorded weights that were obtained from a SECA model 354, ten-gram sensitive digital scale and recorded by a trained research nurse are without error. Therefore, this approach can be considered an Absolute Calibration.

There are two stages to statistical calibration. The first is the calibration stage, where one must estimate a regression function that establishes the relationship between *x* and *y*. The second stage is the inverse prediction stage, where the regression model is used to estimate the unknown x_0 in a new sample.

To begin the calibration stage, we consider a calibration data set consisting of *n* analog and digital measurement pairs. We will denote the analog and digital weights of the i^{th} pair as Y_i and X_i respectively. We assume that each analog weight record is equal to the true measurement according to the analog scale plus random error resulting from operator error, rounding and recording error

$$Y_i = ith\ analog\ weight$$
$$Y_i = y_i + \varepsilon_i \quad (1)$$

y_i = the true measurement from analog scale (approximately equal to average after multiple readings)

$\varepsilon_i = random\ error$

We assume that the error for each digital weight record is negligible. Therefore, each recorded digital record is taken to be the true record from the digital scale

$$X_i = x_i + negligible\ error \quad (2)$$

$X_i = ith\ digital\ weight$

x_i = true digital weight

Next, we account for the bias in the analog scale by assuming that the true measurements of the analog scale vary as a linear function of the digital weights:

$$y_i = \beta_0 + \beta_1 x_i \quad (3)$$

Substituting (3) into (1) yields the following calibration function:

$$Y_i = \beta_{0\,Y|X} + \beta_{1\,Y|X} X_i + \varepsilon_i \quad (4)$$

Where $\beta_{0\,Y|X}$ denotes the expected value of y given x. The calibration function shows that each recorded analog weight, Y_i, is equal to a linear function of the true digital weight, plus a random error, ε_i. The values of the coefficients, $\beta_{0\,Y|X}$ and $\beta_{1\,Y|X}$, and the error function, ε_i, can be estimated from the calibration data set. This process is discussed in detail in the following sections. Once $\beta_{0\,Y|X}$, $\beta_{1\,Y|X}$, and ε_i are estimated, equation (4) can then be applied to the Inverse Prediction Stage.

For the Inverse Prediction Stage, we invert the calibration function, (4), to obtain the following expression for the digital weight, X_i:

$$\tilde{X}_i = \frac{Y_i - \beta_{0\,Y|X} - \varepsilon_i}{\beta_{1\,Y|X}} \quad (5)$$

When x is not observed, we can use the coefficient from the calibration to estimate the digital weight ($\tilde{X}_i$). We can use this equation to estimate the digital weight of subsequent analog records when an actual digital measurement is unavailable.

3.3.4 Estimating the Regression Calibration Coefficients

We wish to estimate the linear coefficients, $\beta_{0\,Y|X}$ and $\beta_{1\,Y|X,}$ from (4). To do so, it is useful to take advantage of the principle of Mathematical Expectation, whereby the error term,

ε_i, can be taken out of equation (4) in expectation because the error is centered around a mean of zero. Therefore, equations (4) and (5) can be expressed as

$$\mathrm{E}(\mathrm{Y_i}) = \hat{\beta}_{0\,Y|X} + \hat{\beta}_{1\,Y|X} X_i \tag{6}$$

$$\bar{X}_i = \frac{E(Y_i) - \hat{\beta}_{0\,Y|X}}{\hat{\beta}_{1\,Y|X} X_i} \tag{7}$$

Equation (7) shows that each unique analog weight and the mean of the corresponding digital weights are linearly related by expectation. This forms the basis of a linear regression to estimate $\beta_{0\,Y|X}$ and $\beta_{1\,Y|X}$.

The weighting parameter for each sample is taken to be the number of records used to calculate the sample mean. Thus, means that are better defined are given greater weight. This weighted regression can be expressed based on equation (7) above for distinct values of *Y* where we weight n_j observations at a value of Y'_j.

$$\sum_{j=1}^{J} X_j = \sum_{j=1}^{J} \left[\frac{\left(Y'_j\right) - \hat{\beta}_{0\,Y|X}}{\hat{\beta}_{1\,Y|X} X_i} \right] \tag{8}$$

Regrouping the terms yields the following expression with a weighting factor of n_j

$$\sum_{j=1}^{J} n_j X_j = \sum_{j=1}^{J} \left\{ n_j \left[\frac{\left(Y'_j\right) - \hat{\beta}_{0\,Y|X}}{\hat{\beta}_{1\,Y|X} X_i} \right] \right\} \tag{9}$$

Consider Figure 3.5. The grey diamonds show the analog weight of each record versus its digital weight. For each unique analog weight, we selected all of the corresponding records and estimated the mean of their digital weights. These sample means are shown on Figure 3.5 as black circles. The relative size of each black circle relates to the number of records used to estimate that mean; thus, the regression is weighted. Larger black circles indicate that more data were available, and thus its value better reflects the true population mean than a small circle. Notice that the sample means vary linearly with analog weight. We then performed a weighted linear regression on the sample means and their corresponding analog weights. The

best-fit line is shown in black on Figure 3.5. Based on the weighted regression, β_0 = -0.111 and β_1 = 1.05.

3.3.5 Estimating the random error function (ε_i)

In certain cases, there is actually no need to estimate the random error function. Instead, we could use the actual residual distribution from the data in place of an estimated error function. The advantage of this approach is that it reduces overall error to use the true residual distribution as opposed to an estimated one. However, it is not appropriate to use the actual residual distribution in all cases. In order to generalize this approach to other populations, it is necessary to develop a method of modeling the error that could be applied to any calibration data set for any population. Also, as was the case here, we assumed that the true residual distribution in the population varied smoothly, as opposed to the residual distribution we see from this sample population. Therefore, depending on the data source, size of the calibration data set, representativeness of the sample and parameters of the actual residual distribution, one must decide whether to use a modeled error function or the actual residual distribution as the error function.

In standard statistical calibration, the random error (ε_i) is assumed to be a normal distribution with a mean of zero and a standard deviation that is constant across the range of records, in this case the range of BWs. However, as demonstrated below, the error in these data is not normally distributed. Furthermore, we do not wish to assume, a priori, that the distribution parameters are constant for all weight ranges. Therefore, a more flexible approach is used to estimate the appropriate error function based on the calibration data.

The analog weights are clustered into only 69 unique values, and each unique analog value appears in multiple data records. Consider figure 3.6, which shows the digital weight versus the analog weight of each record. The plot is zoomed in around 2.5kg, and the data points with an analog weight of 2.5kg are shown in black. The clustering of analog weights is clearly visible. Notice that the data with an analog weight of 2.5kg has digital values that range from 1.7 to 3.5 kg. Therefore, these data naturally lend themselves to considering the scatter distribution in the digital weights for a given analog weight.

However, the error function we wish to define, ε_i, characterizes the inverse quantity. As shown in equation (4), it describes the scatter in the recorded analog weight for a given digital weight. Therefore, we will establish a relationship between these two error quantities for the calibration data. Consider the inverse prediction equations (5) and (6):

$$\check{X}_i = \frac{Y_i - \beta_{0\,Y|X} - \varepsilon_i}{\beta_{1\,Y|X}} \tag{5}$$

$$\sum_{i=1}^{n} X_i = \sum_{i=1}^{n} \left[\frac{(Y_i) - \widehat{\beta}_{0\,Y|X} - \varepsilon_i}{\widehat{\beta}_{1\,Y|X} X_i} \right] \tag{6}$$

Regrouping the terms yields the following expressions for one observation and all the observations, respectively:

$$X_i = \left(\frac{-\beta_{0\,Y|X}}{\beta_{1\,Y|X}} \right) + \left(\frac{1}{\beta_{1\,Y|X}} \right) Y_i + \left(\frac{-\varepsilon_i}{\beta_{1\,Y|X}} \right) \tag{10}$$

$$\sum_{j=1}^{J} n_j X_j = \sum_{j=1}^{J} \left\{ nj \left[\frac{(Y_{jk}) - \widehat{\beta}_{0\,Y|X}}{\widehat{\beta}_{1\,Y|X} X_i} \right] \frac{\sum_{k=1}^{nj} \varepsilon_{jk}}{\widehat{\beta}_{1\,Y|X}} \right\} \tag{11}$$

This can be expressed in simplified form as

$$X_i = \beta_0' + \beta_1' Y_i + \varepsilon_i' \tag{12}$$

$$\text{Where } \beta_0' = \left(\frac{-\beta_{0\,Y|X}}{\beta_{1\,Y|X}} \right) \tag{12a}$$

$$\beta_1' = \left(\frac{1}{\beta_{1\,Y|X}} \right) \tag{12b}$$

$$\varepsilon_i' = \left(\frac{-\varepsilon_1}{\beta_{1\,Y|X}} \right) \tag{12c}$$

Here, ε_i' is the error function that describes the scatter in the digital weights which correspond to a given analog weight. This is the error quantity which can be most readily estimated from the data. Once the equation for ε_i' is established, we will relate it back to ε_i through equation (12c) to complete the calibration function.

Rearranging the terms in equation (12) yields the following equation for ε_i':

$$\varepsilon_i' = X_i - \beta_0' - \beta_1' Y_i \quad (13)$$

To begin the process of determining the function of Ɛ', we use (13) to estimate the value of $Ɛ_i'$ for each record. Physically, this error is the discrepancy of each digital weight from the estimated value based on the calibration model. Figure 3.7a shows the analog and digital weights of each record and the linear regression fit line. Subtracting the estimated weighted regression value from each digital weight results in the error, $Ɛ_i'$, as shown in figure 3.7b.

First, we looked at the errors associated with those analog weights listed as exactly 2.5 kg. There were enough data points at this specific analog weight to create a distribution and 2.5 kg is important because it marks the boundary between NBW and LBW. Figure 3.8 shows the distribution of error, ε_i', for the records with an analog weight of 2.5 kg.

Using the statistical software MATLAB (Natick, MA), a best fit curve was generated for the data. (Figure 3.9) The *t* location-scale distribution was found to match the data most closely. Figure 3.9 shows an overlay of the best fit t-distribution curve on the histogram. For comparison, we have included an overlay of the normal distribution curve, which does not accurately model the data.

The *t* location-scale distribution has the density function:

$$\frac{\Gamma\left(\frac{\upsilon+1}{2}\right)}{\sigma\sqrt{\upsilon\pi}\,\Gamma\left(\frac{\upsilon}{2}\right)}\left[\frac{\upsilon+\left(\frac{x-\mu}{\sigma}\right)^2}{\upsilon}\right]^{-\left(\frac{\upsilon+1}{2}\right)} \quad (14)$$

With location parameter μ, scale parameter σ > 0, and shape parameter ν > 0. If x has a *t* location-scale distribution, with parameters μ, σ, and ν, then

$$\frac{x-\mu}{\sigma} \quad (15)$$

has a Student's *t* distribution with ν degrees of freedom.

We see from this process that the error for the records with an analog weight of 2.5 kg is well modeled by a student's t distribution and that the model parameters for the best fit distribution can be estimated. In theory, the method described above could be applied to each unique analog weight, resulting in a custom fit error function for each analog weight. However, in practice many unique analog weights do not have enough records to generate a distribution. So, to apply the above concept across the whole range of BWs, we applied several methods to estimate the error functions across the full range of BWs without assuming that they would be the same. The conclusion of these analyses was that the distribution of digital weights is uniform for all analog values. This characteristic is beneficial because it means that to adjust the analog BW data, we will not have to rely on separate error distributions from different places in the range of BWs. Instead, a single error distribution can be derived which applies to all analog weights. For a full description of this process, please see Appendix A.

Having shown that a single error distribution was appropriate to use across the range of BWs, a single distribution was fit to all the error shown in Figure 3.10. The resulting best-fit Student's T distribution is shown in Figure 3.11. The distribution parameters are as follows:

location (μ) = 0.00279609
scale (σ) = 0.061233
shape (ν) = 1.96115

Now that that random error distribution (ε_i) has been estimated, we can return to the calibration function and inverse prediction phase. As shown in function (4) the appropriate calibration function for these data would be:

$$Y_i = \beta_{0\,Y|X} + \beta_{1\,Y|X} X_i + \varepsilon_i \qquad (4)$$

We have now established that the most appropriate function for error (ε_i') is the student's T-shape distribution, described above

$$\varepsilon_i{}' = \frac{\Gamma\left(\frac{\upsilon+1}{2}\right)}{\sigma\sqrt{\upsilon\pi}\Gamma\left(\frac{\upsilon}{2}\right)}\left[\frac{\upsilon+\left(\frac{x-\mu}{\sigma}\right)^2}{\upsilon}\right]^{-\left(\frac{\upsilon+1}{2}\right)} \quad (16)$$

Where the parameters for ε_i' are:

location (μ) = 0.00279609

scale (σ) = 0.061233

shape (ν) = 1.96115

Using the relationship between ε_i' and ε_i from equation (12c), we find that

$$\varepsilon_{\mathrm{i}} = \beta_1\left[\frac{\Gamma\left(\frac{\upsilon+1}{2}\right)}{\sigma\sqrt{\upsilon\pi}\Gamma\left(\frac{\upsilon}{2}\right)}\left[\frac{\upsilon+\left(\frac{x-\mu}{\sigma}\right)^2}{\upsilon}\right]^{-\left(\frac{\upsilon+1}{2}\right)}\right] \quad (17)$$

Therefore, the final calibration function would be

$$Y_i = \beta_{0\,Y|X} + \beta_{1\,Y|X}X_i + \beta_1\left[\frac{\Gamma\left(\frac{\upsilon+1}{2}\right)}{\sigma\sqrt{\upsilon\pi}\Gamma\left(\frac{\upsilon}{2}\right)}\left[\frac{\upsilon+\left(\frac{x-\mu}{\sigma}\right)^2}{\upsilon}\right]^{-\left(\frac{\upsilon+1}{2}\right)}\right] \quad (18)$$

With the function parameters of

location (μ) = 0.00279609

scale (σ) = 0.061233

shape (ν) = 1.96115

β_0 = -0.111

β_1 = 1.049

For the inverse prediction phase, discussed above and in equations (5) and (6), the final functions would be

$$X_i = \frac{Y_i - \beta_{0\,Y|X} - \beta_1 \left[\frac{\Gamma\left(\frac{\upsilon+1}{2}\right)}{\sigma\sqrt{\upsilon\pi}\,\Gamma\left(\frac{\upsilon}{2}\right)} \left[\upsilon + \left(\frac{x-\mu}{\sigma}\right)^2 \right]^{-\left(\frac{\upsilon+1}{2}\right)} \right]}{\beta_{1\,Y|X}} \quad (19)$$

$$\sum_{i=1}^{n} X_i = \sum_{i=1}^{n} \left[\frac{(Y_i) - \widehat{\beta}_{0\,Y|X} - \beta_1 \left[\frac{\Gamma\left(\frac{\upsilon+1}{2}\right)}{\sigma\sqrt{\upsilon\pi}\,\Gamma\left(\frac{\upsilon}{2}\right)} \left[\upsilon + \left(\frac{x-\mu}{\sigma}\right)^2 \right]^{-\left(\frac{\upsilon+1}{2}\right)} \right]}{\widehat{\beta}_{1\,Y|X} X_i} \right] \quad (20)$$

The surface in Figure 3.12 is a visual representation of the calibration/prediction functions described above. For each analog weight in the data set, the appropriate distribution is chosen from the surface distribution for that analog weight. In the example in Figure 3.13, the appropriate distribution is drawn from the surface for an analog weight record of 1.5 kg. The mean (μ) of the distribution is derived from the regression function described above.

3.3.6 Imputation of multiple data sets

Once the appropriate distribution is available for each analog BW record, a new dataset is imputed by drawing random numbers from within the Student's T distribution with the appropriate parameters for that analog data record. This will then create a new dataset that is adjusted and no longer contains the biases that were present in the original (in this case analog) BW data set. (Figure 3.14) The imputed data set can be compared to the digital weights in the calibration data set. Figure 3.14 shows that the imputed weights are quite similar to the digital weights, the known gold standard.

A single imputed data set using this method will be only one estimate of the "true" data set, had there not been measurement or information error. Repeated imputations would result in multiple, slightly different data sets. In order to combine the results from the analysis of multiple imputed data sets, one must apply Rubin's Rules.

To obtain a standard error, one must calculate the between-imputation variance

($B = (m-1)^{-1}\sum(\hat{Q}^l - \bar{Q})^2$) and the within-imputation variance ($U = m^{-1}\sum U^{(l)}$). Q denotes the odds ratio to be estimated and m denotes the number of number of simulated imputations. The estimated total variance is $T = (l + m^{-1})B + \bar{U}$. The total combined variance can be used to estimate confidence intervals or perform significance tests.[96] (Figure 3.15)

3.4 RESULTS

3.4.1 Comparing the proportion of LBW babies in original versus imputed data sets

The first measure used to assess the success of imputing data using this method was the percentage of babies that are LBW in the digital, analog and imputed data sets. The analog BW data shows that 22% of the babies are LBW. The digital BW data show that 28% are LBW and the imputed data from 10 imputed data sets shows that 30% are LBW, with a range of 28% - 33%.

3.4.2 Comparing the association of head circumference and baby length with BW in the original versus imputed data sets

The Rekha dataset also includes measurements for head circumference and length of the baby at birth. While these variables are not significantly associated with BW as a continuous variable, they are associated with BW as a categorical variable (NBW vs. LBW) by logistic regression analysis. As you can see in Figures 3.16 and 3.17, the imputed data (from 10 data sets), matches the curve for the digital data most closely and is different from the analog data.

3.4.3 Applying this method to an external data source – the Indian National Family Health Survey 3

The modified statistical calibration and multiple imputation method was applied to an external dataset for validation. We used the Indian National Family Health Survey 3 (2005-2006), a large representative sample from across India that includes data on 56,437 births and 19,237 recorded BWs. These data were collected through surveys where mothers across India were asked the BWs of all children born in the previous five years. Almost all recorded BWs

(85.7%) came from maternal memory, as opposed to a written record or health record. As a result, we expected significant heaping in these data.

As seen in the Normal Q-Q graph of the raw NFHS data (Figure 3.18a) the observed BW data does not produce a straight line that matches the reference line. Had the distribution of BWs been normal, as expected for a large population, the observed value would overlie the reference line. Here, the small segments indicate heaping or binning of data at certain intervals. After applying the modified statistical calibration method outlined above and imputing new data sets, the majority of the data overlies the reference line (Figure 3.18b) indicating that heaping has been reduced and the data now form a normal distribution

3.5 DISCUSSION

These results suggest that the adjustment method outlined here is successful at imputing data sets that match the gold standard and have reduced bias due to heaping compared to the analog BW data sets. Using this multiple imputation method, a pooled association coming from the combination of the imputed data sets comes closer to the true association, were there not error in the data. The imputed rate of LBW babies matches the "true" rate and the imputed association between BW and length or BW and head circumference matches the association between these variables and the "true" BWs in the calibration data set.

This adjustment method is quite conservative given the bias often found in BW data from the developing world. This method corrects for bias from measurement error with an analog scale and operator error, such as rounding to the nearest 100 g. This method does not correct for other types of biases often found in BW data from the developing world – namely recall bias when mothers are asked to remember the BW of their baby. However, even given this conservative approach, when applied to an external dataset with heaping due to biases besides measurement error, heaping is reduced and the method results in normally distributed BW data.

BW data are typically considered particularly useful because they are readily and accurately collected through systems of vital statistics. However, in the developing world this is not the case; BW data are only collected through survey data or health care facility data, both biased sources. Given the current lack of available methods to adjust these BW data and produce useable datasets for further analysis, the study of BW from developing countries has been sorely neglected.

With the method developed here, a significant portion of the bias of these survey-collected BW data can be adjusted and the use of multiple imputation makes it possible to study the adjusted BW data in imputed data sets and as a continuous variable. Researchers could use our methods directly in similar populations or replicate what we have done and collect a gold standard calibration set to develop function parameters and error distributions

that are specific to their own population of interest. Or, if that is not possible, our calibration function could be applied to other data sets from the developing world in order to reduce the heaping of BW data to some extent. However, this method might be applied and used, it provides a novel method for adjusting BW data and creating datasets that can be used to study this subject in the developing world countries where it is most needed.

3.6 TABLES AND FIGURES

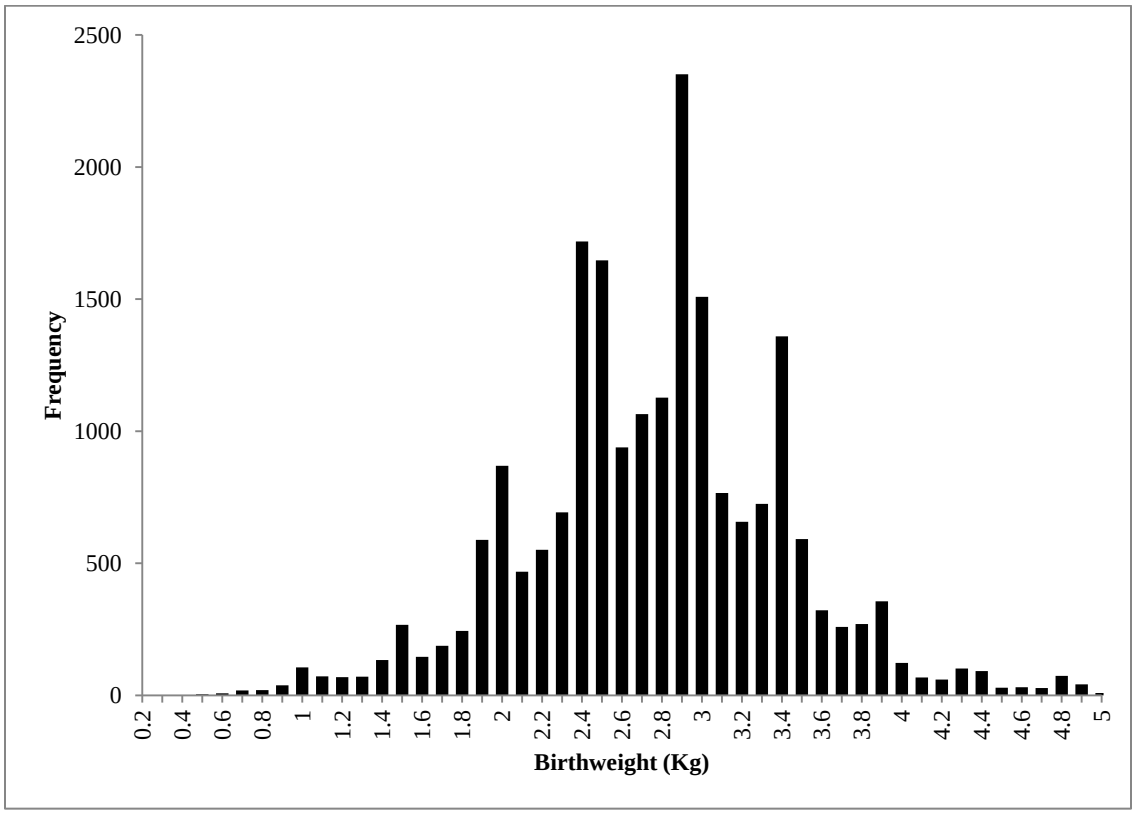

Figure 3.1: Heaping of reported birth-weights on multiples of .5 kg, Indian National Family Health Survey 3 2005-2006

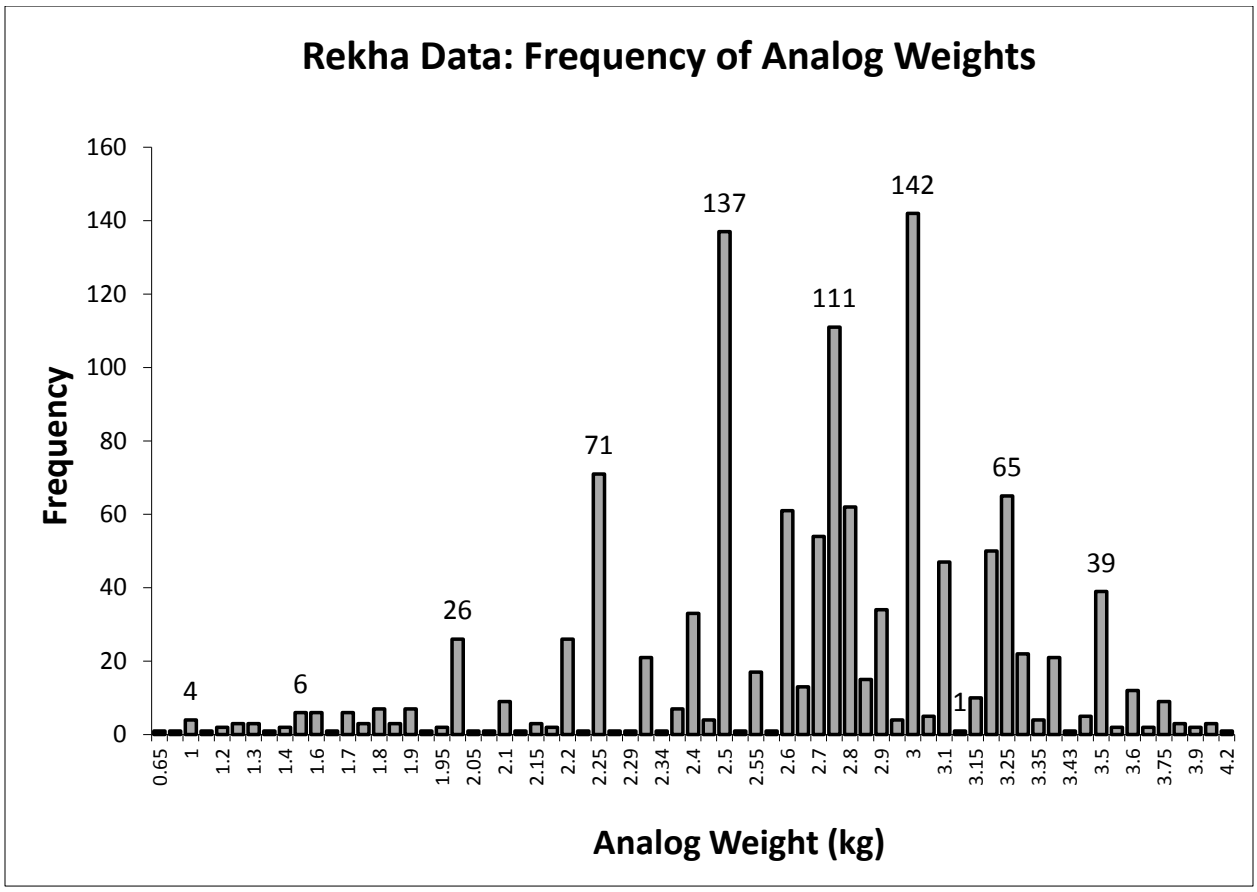

Figure 3.2: Rekha Data – Frequency of Analog Weights

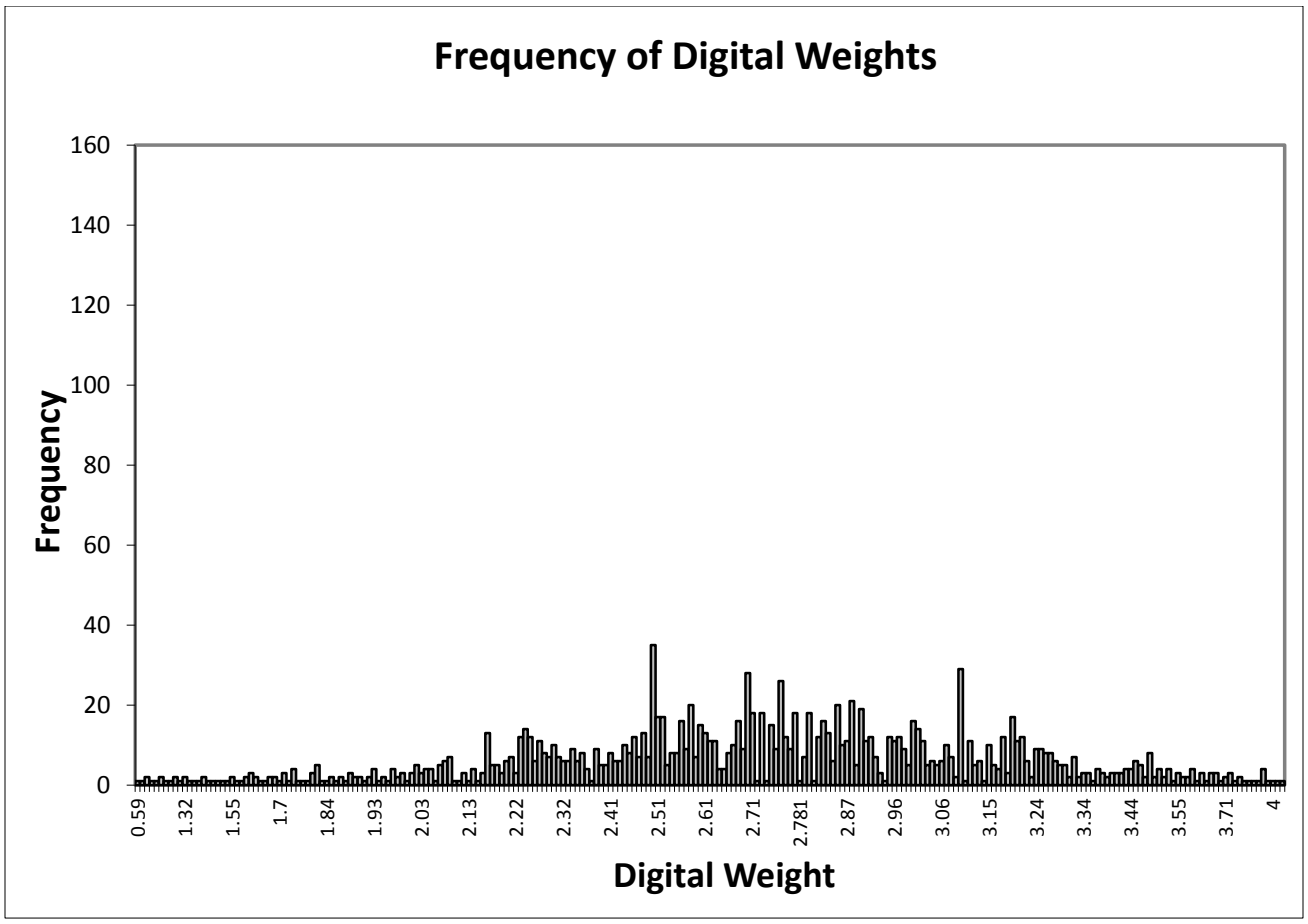

Figure 3.3: Rekha Data – Frequency of Digital Weights (gold standard)

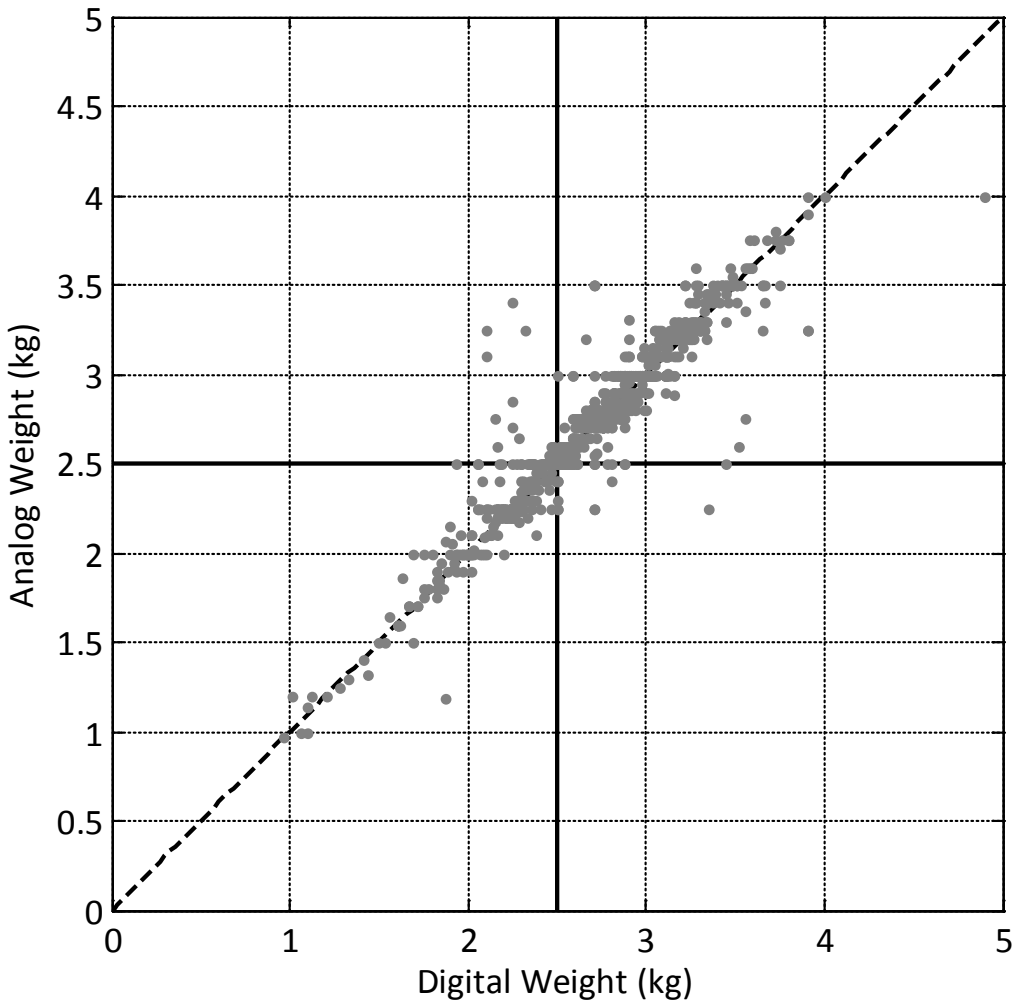

Figure 3.4: Digital versus Analog Weights

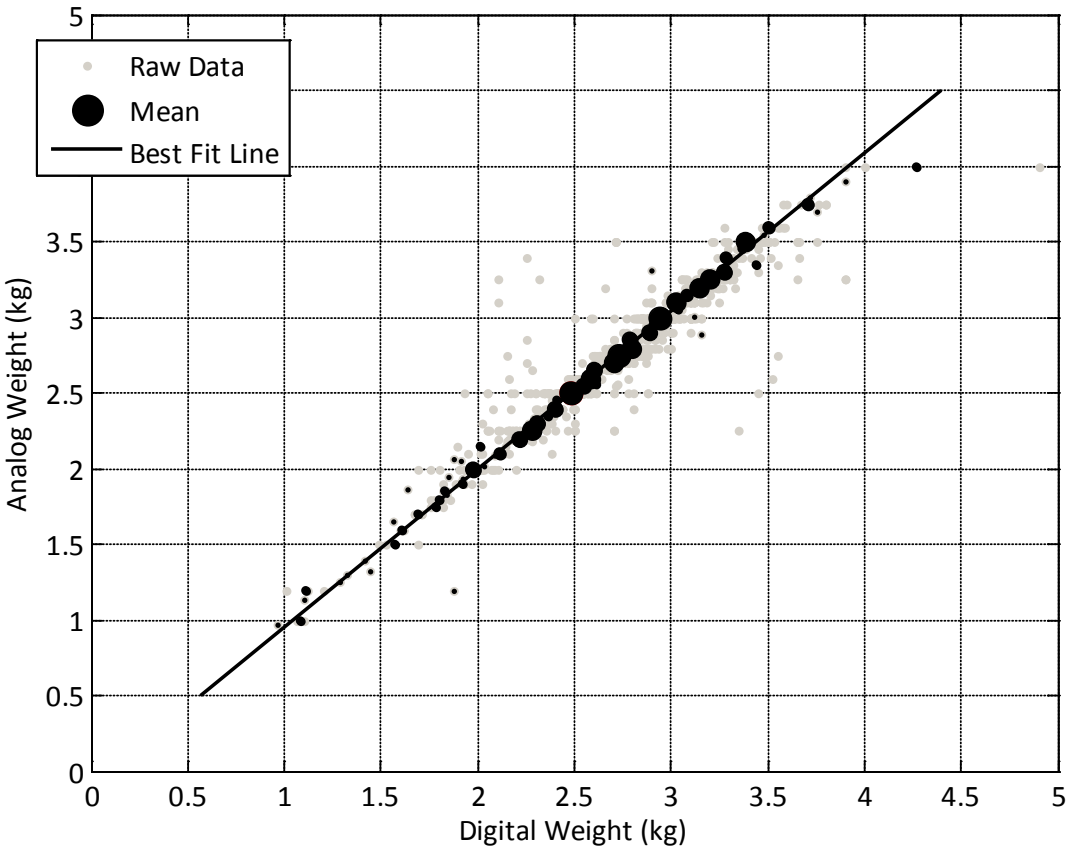

Figure 3.5: Weighted Calibration Regression

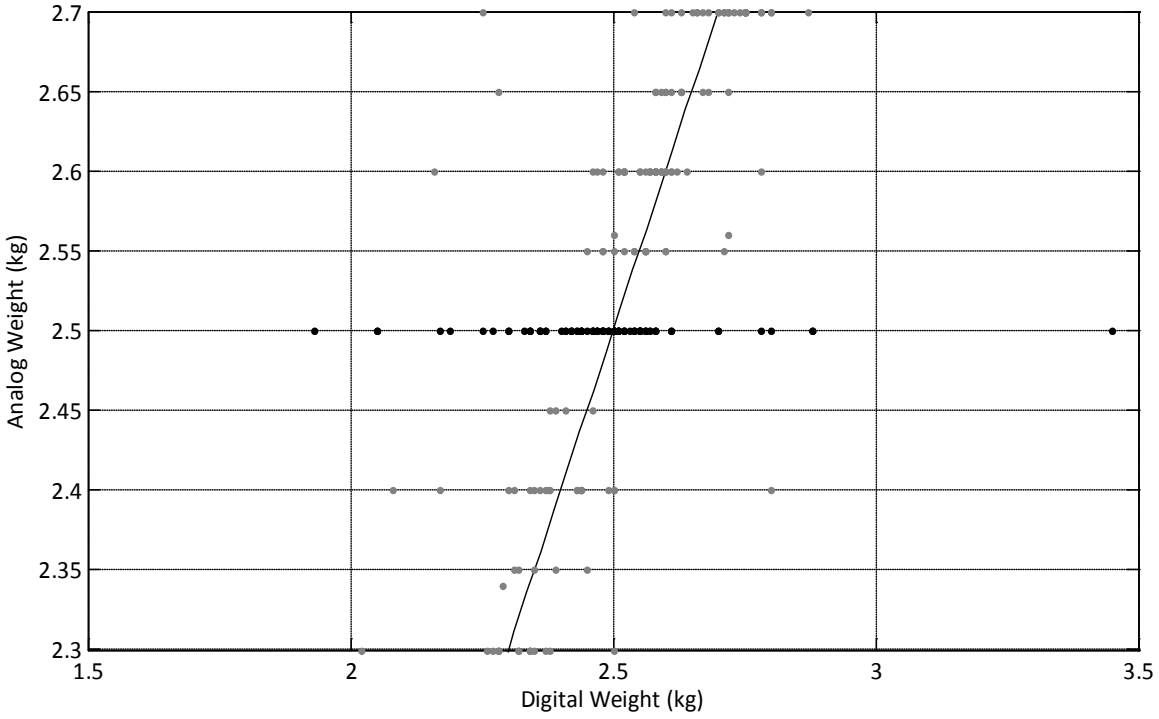

Figure 3.6: Digital versus Analog Weights, Zoom in at 2.5 kg

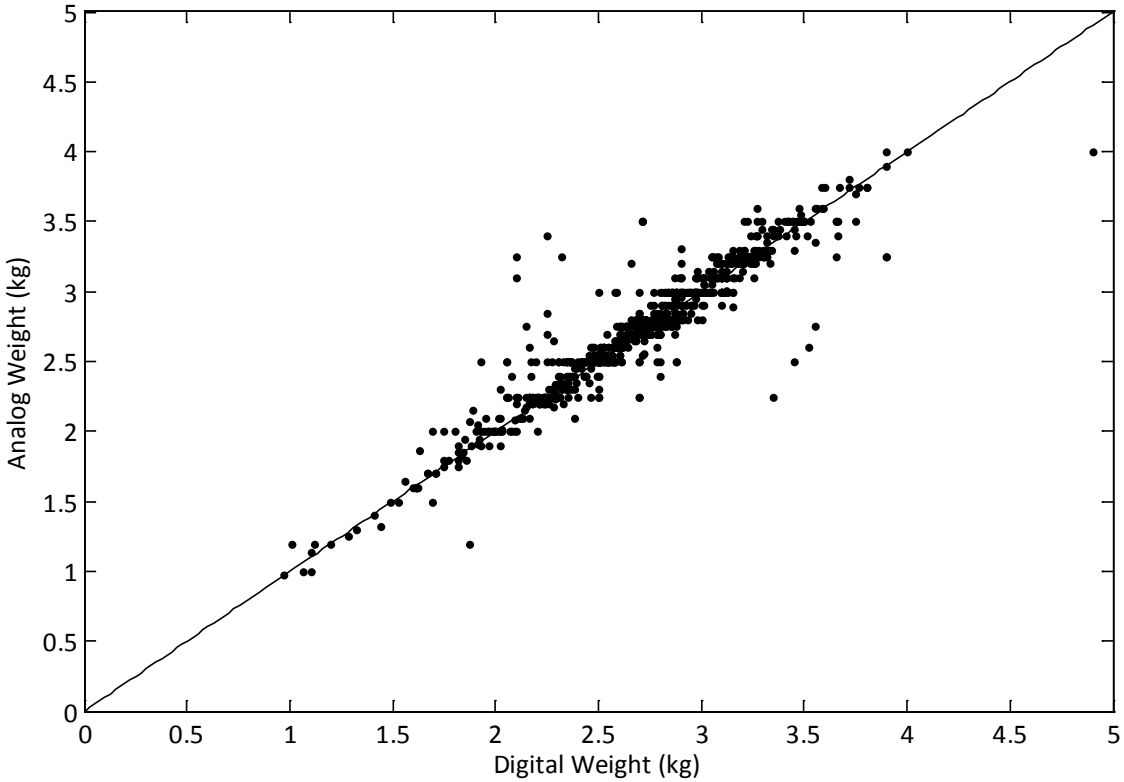

Figure 3.7a: Error in each digital record from the regression fit. Analog versus digital weight

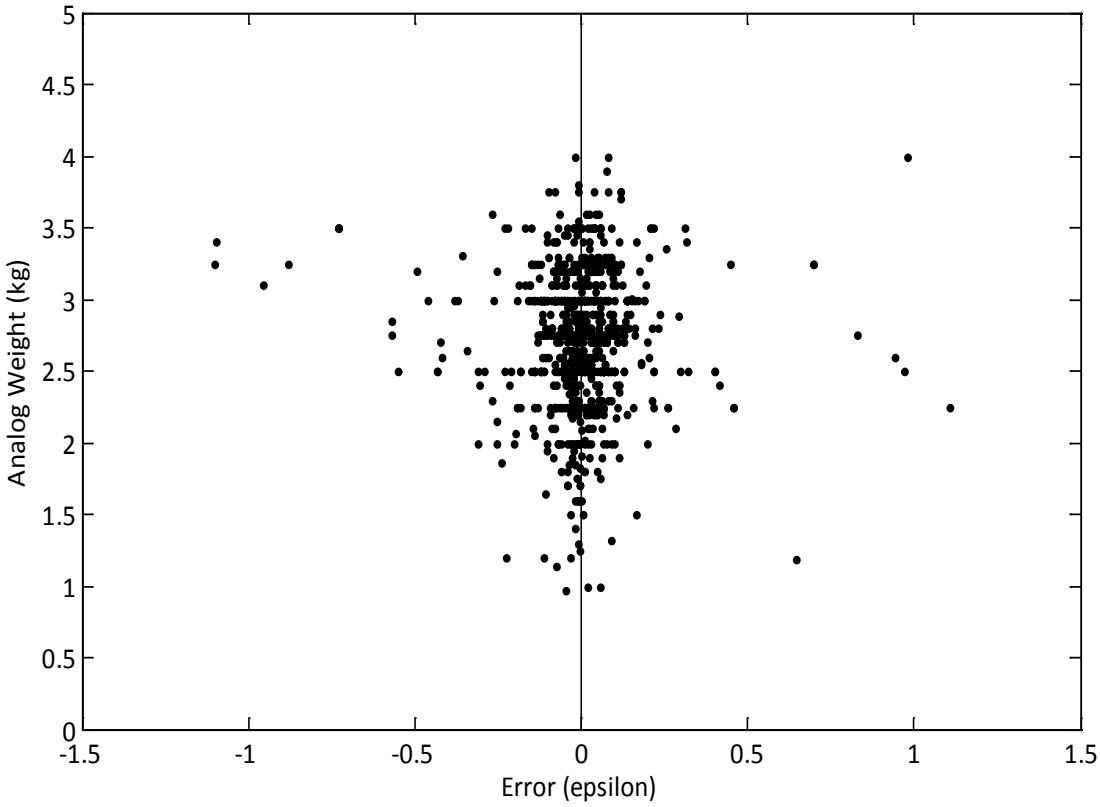

Figure 3.7b. Analog versus Error (Digital weight – best fit regression line)

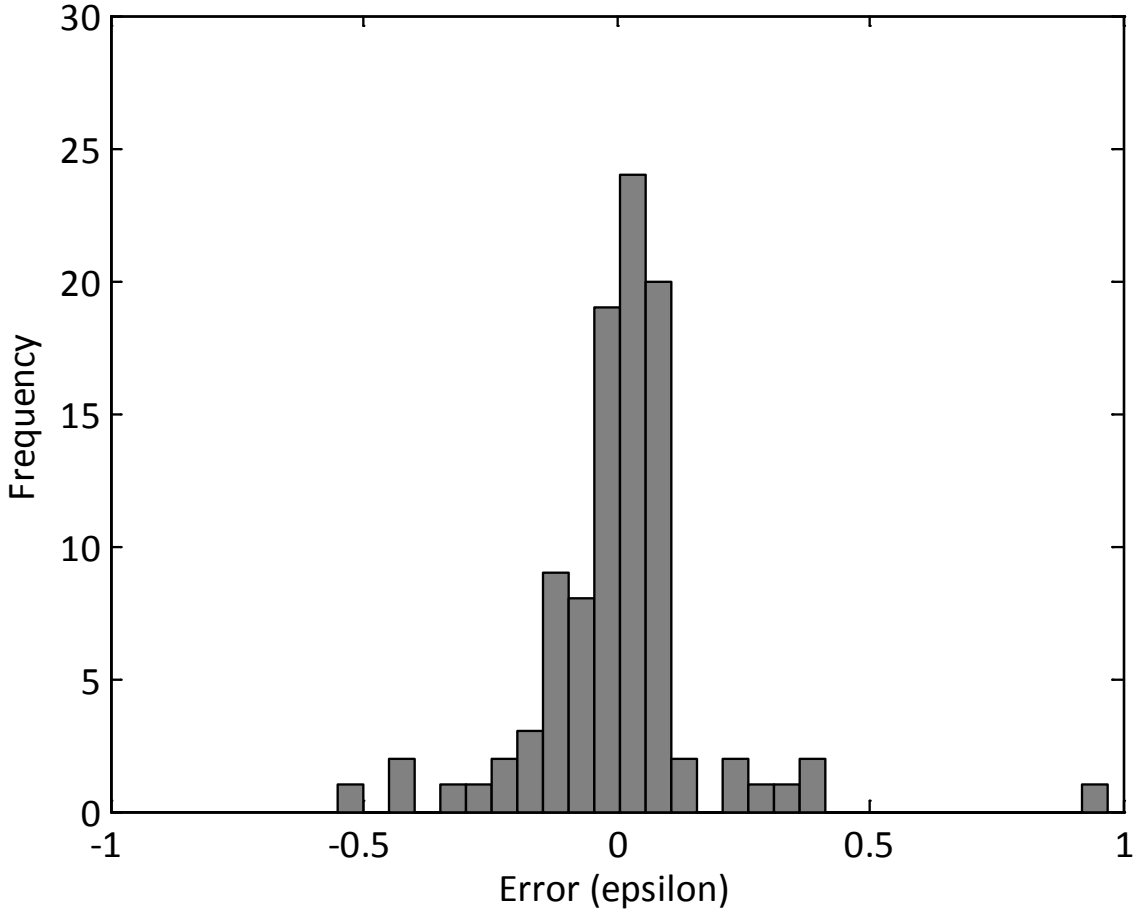

Figure 3.8: Distribution of error (ε_i') for the records that are listed as 2.5 kg by the analog scale

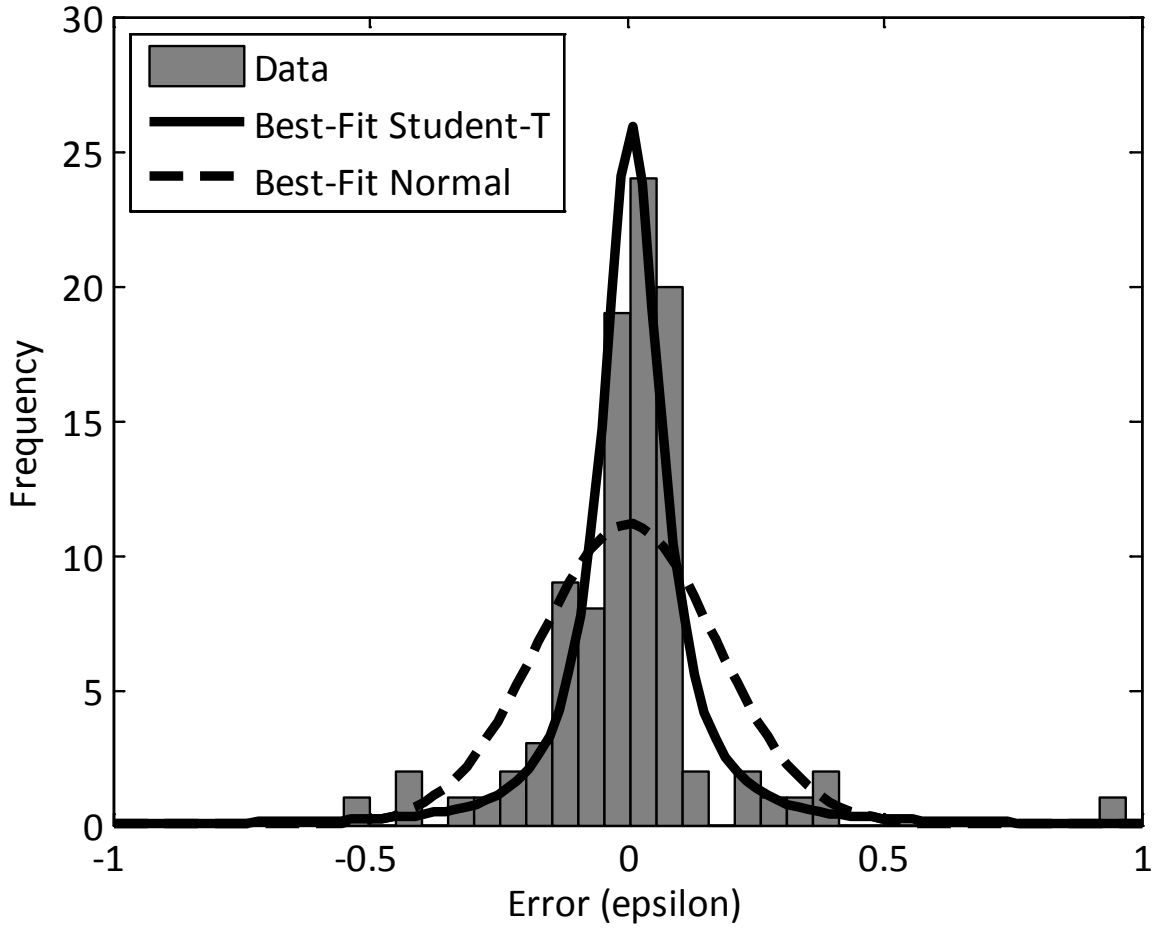

Figure 3.9: Distribution of error (ε_i') for weights at 2.5kg with best fit curve(s)

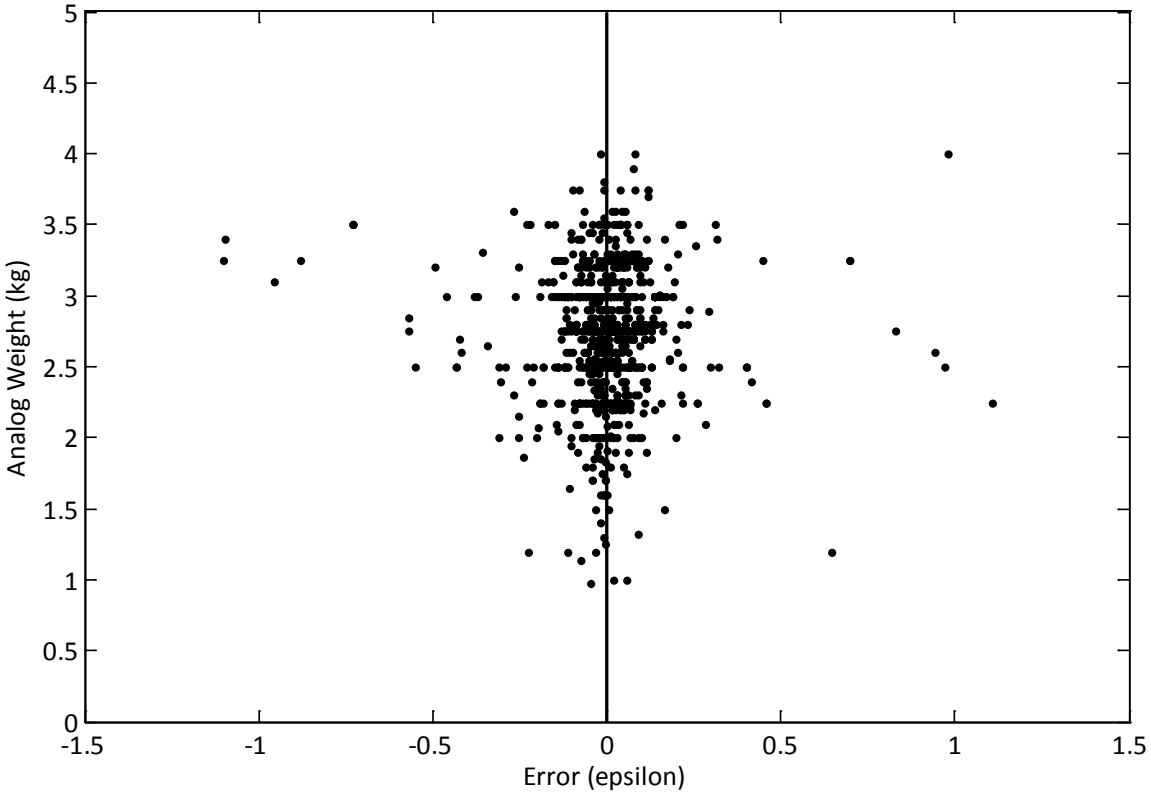

Figure 3.10: Analog weight versus Error (ε_i')

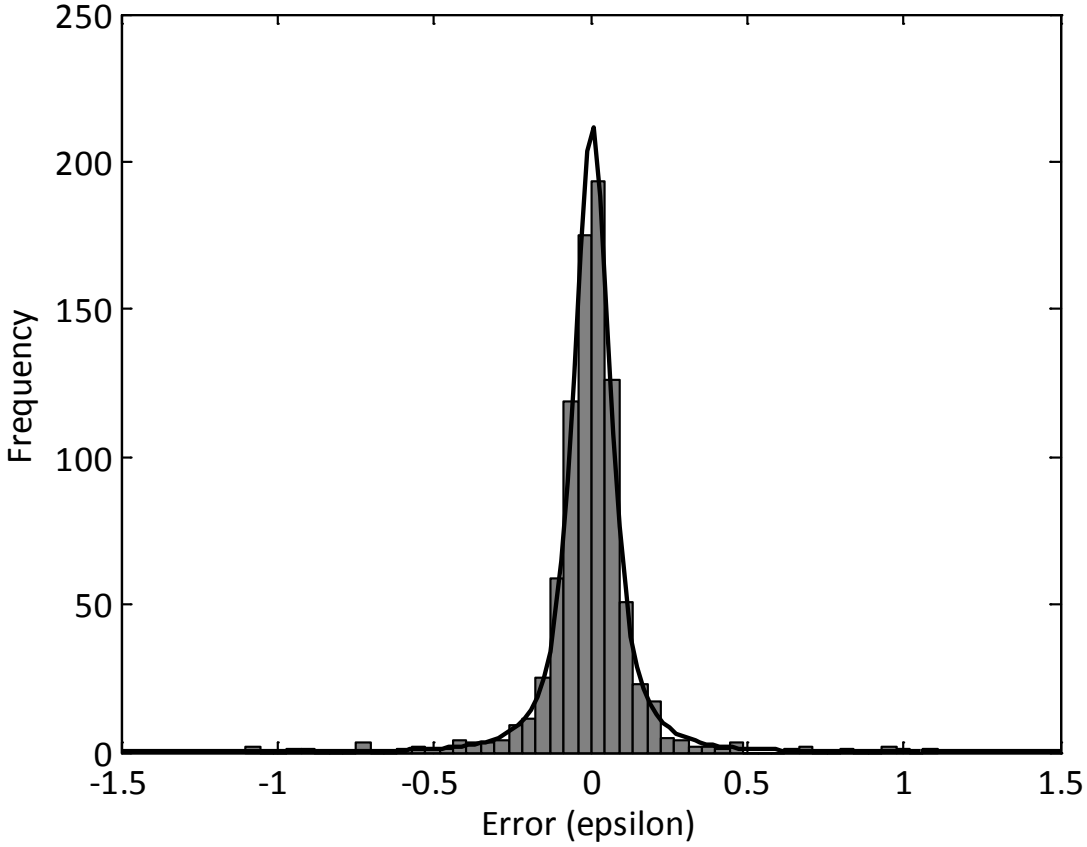

Figure 3.11: Error distribution

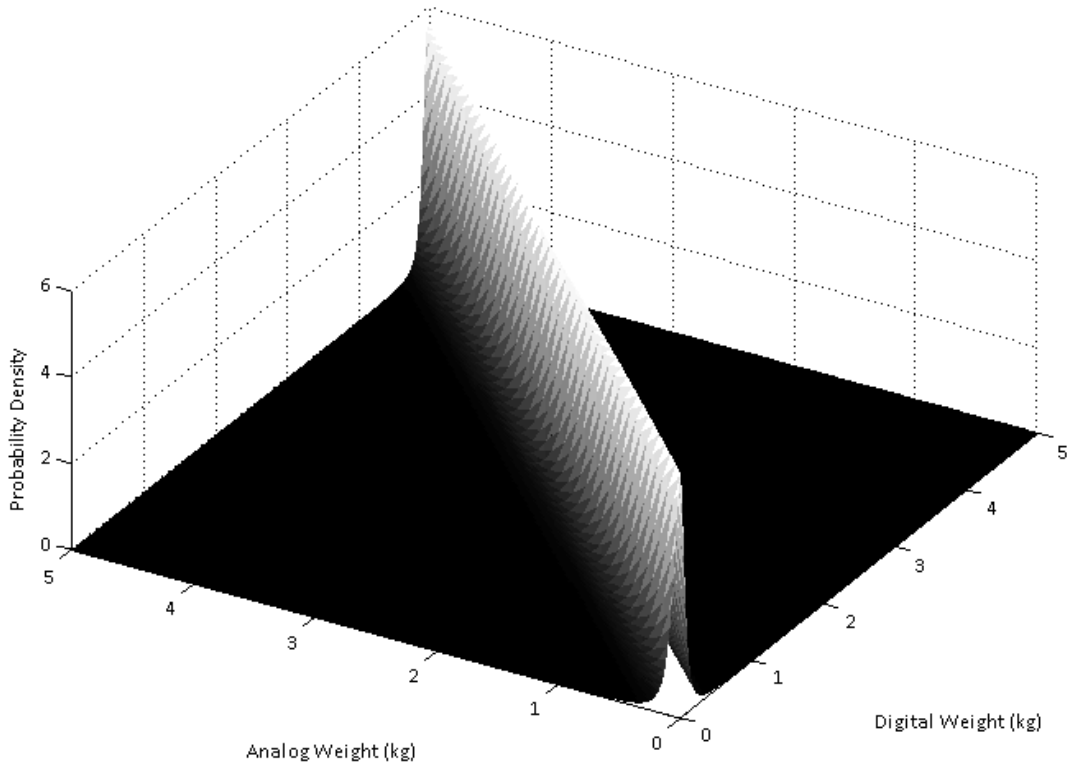

Figure 3.12: Surface representation of inverse prediction function

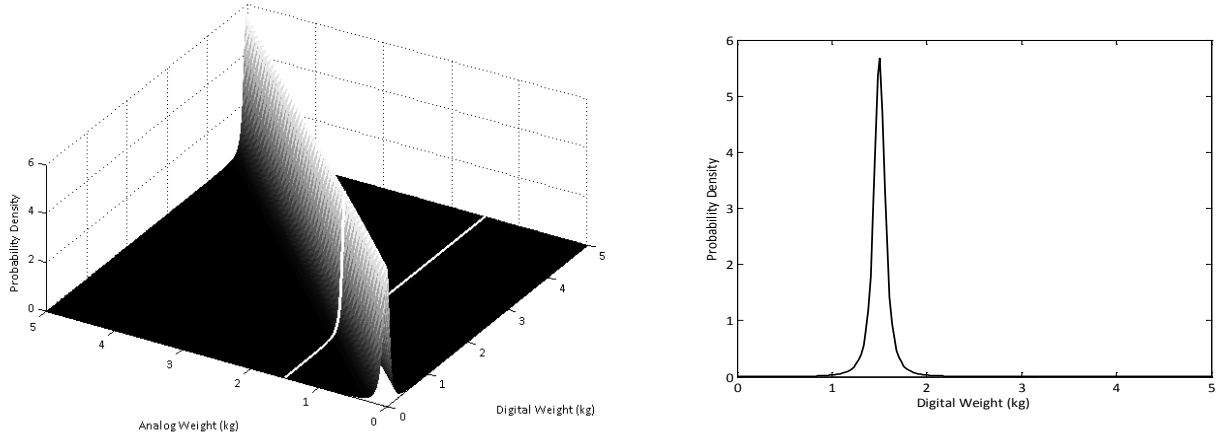

Figure 3.13: Derivation of unique distribution for adjustment of individual analog weight record (Analog weight = 1.5 kg)

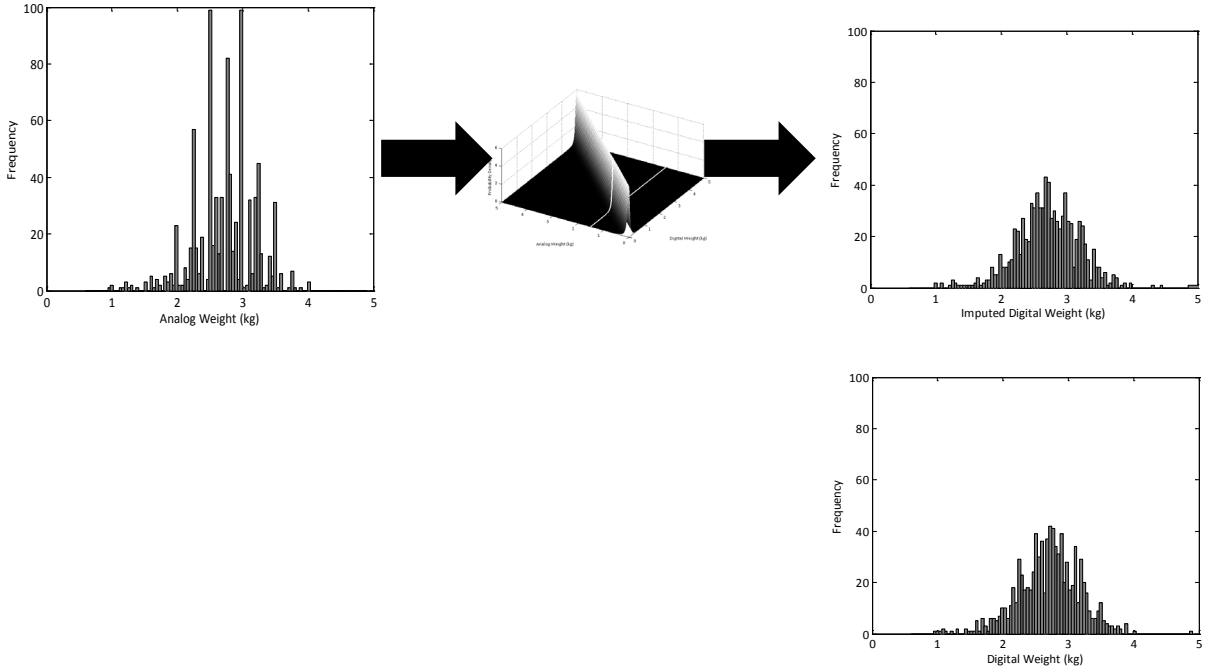

Figure 3.14: Imputation of adjusted data sets

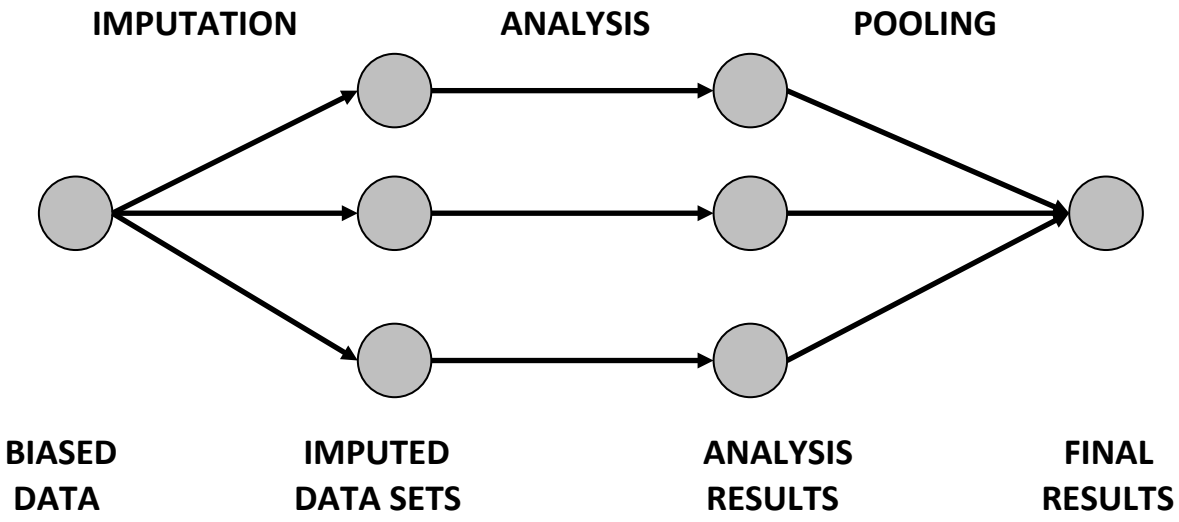

Figure 3.15: Analysis of multiple data sets from multiple imputation

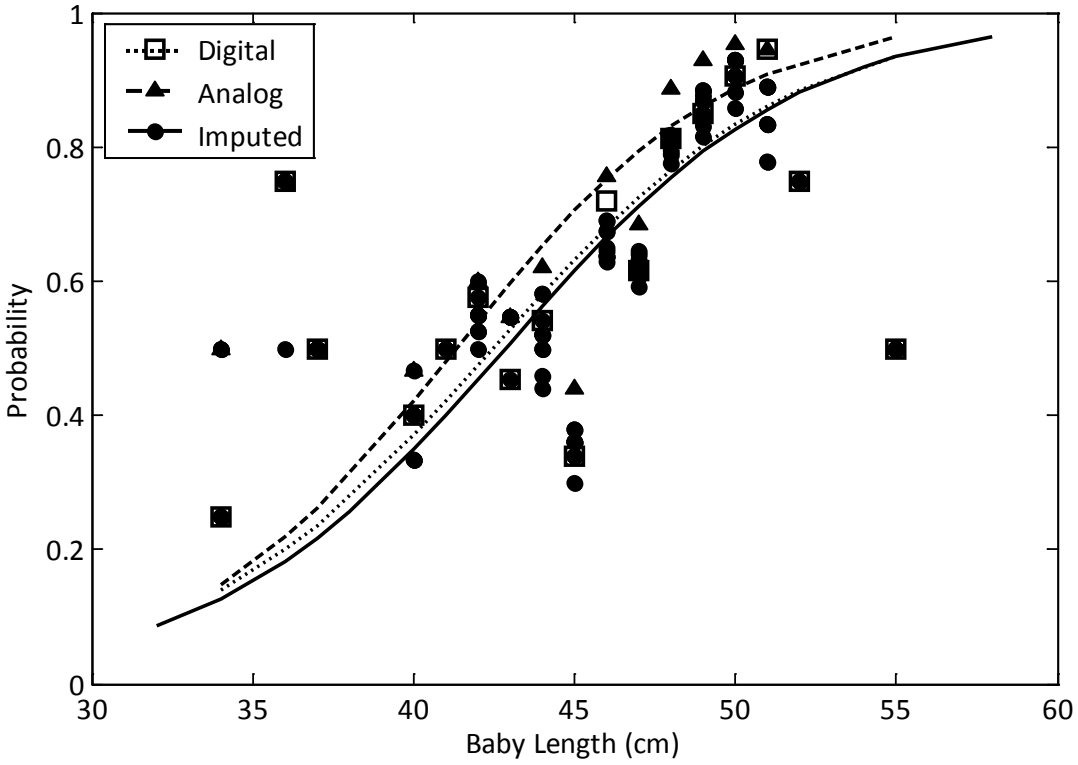

Figure 3.16: Baby's Length versus Probability of Normal Birth-Weight (NBW) – digital, analog and multiply imputed data

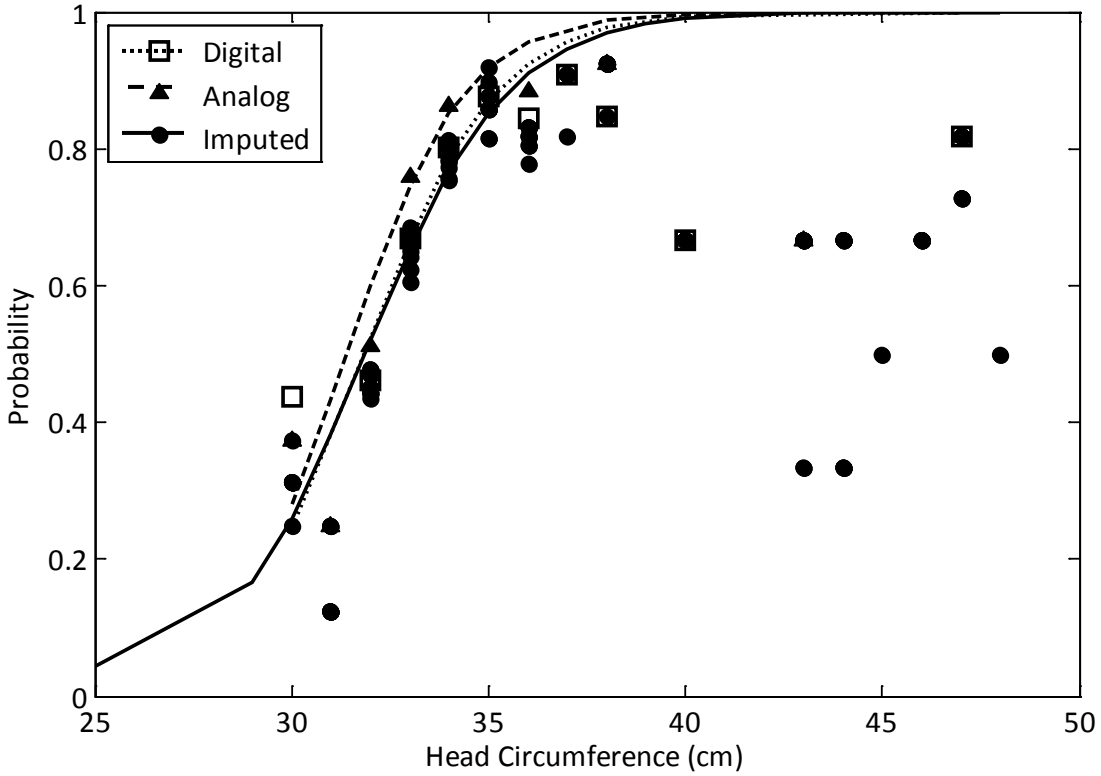

Figure 3.17: Baby's Head Circumference versus Probability of NBW – digital, analog and multiply imputed data

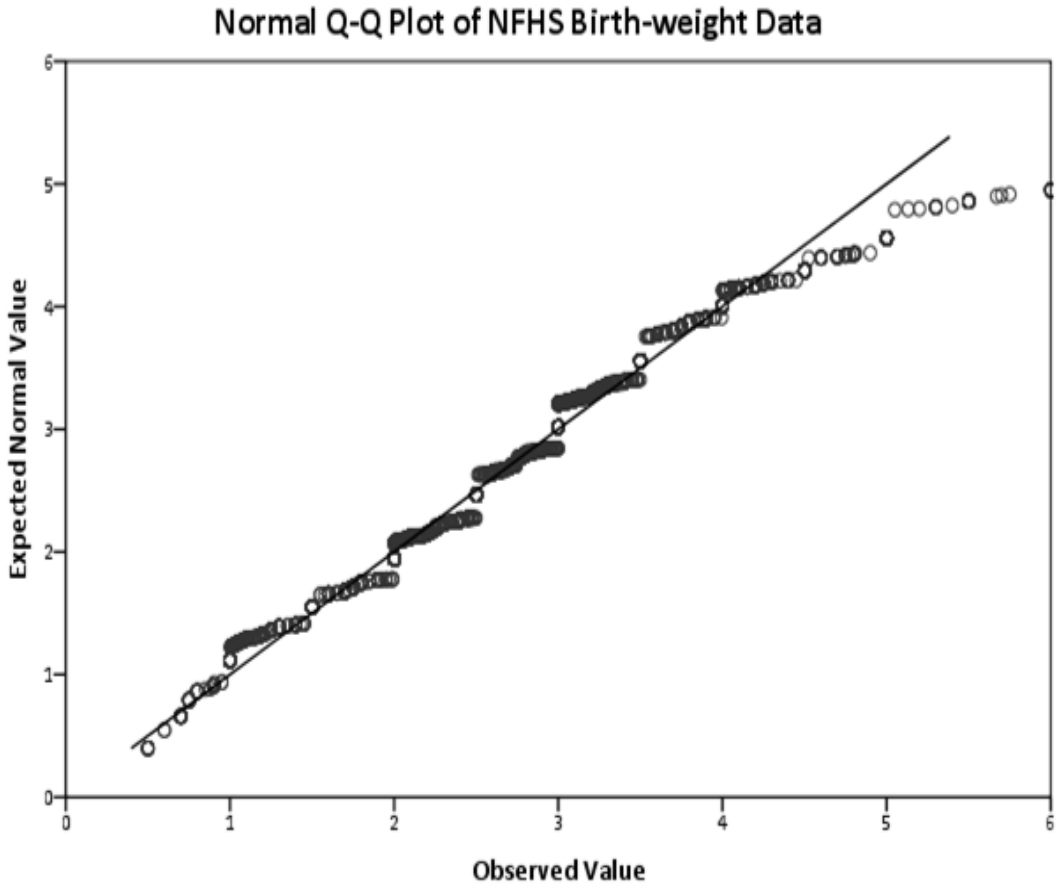

Figure 3.18a: Normal Q-Q Plot of NFHS Birth-weight data – Unadjusted data

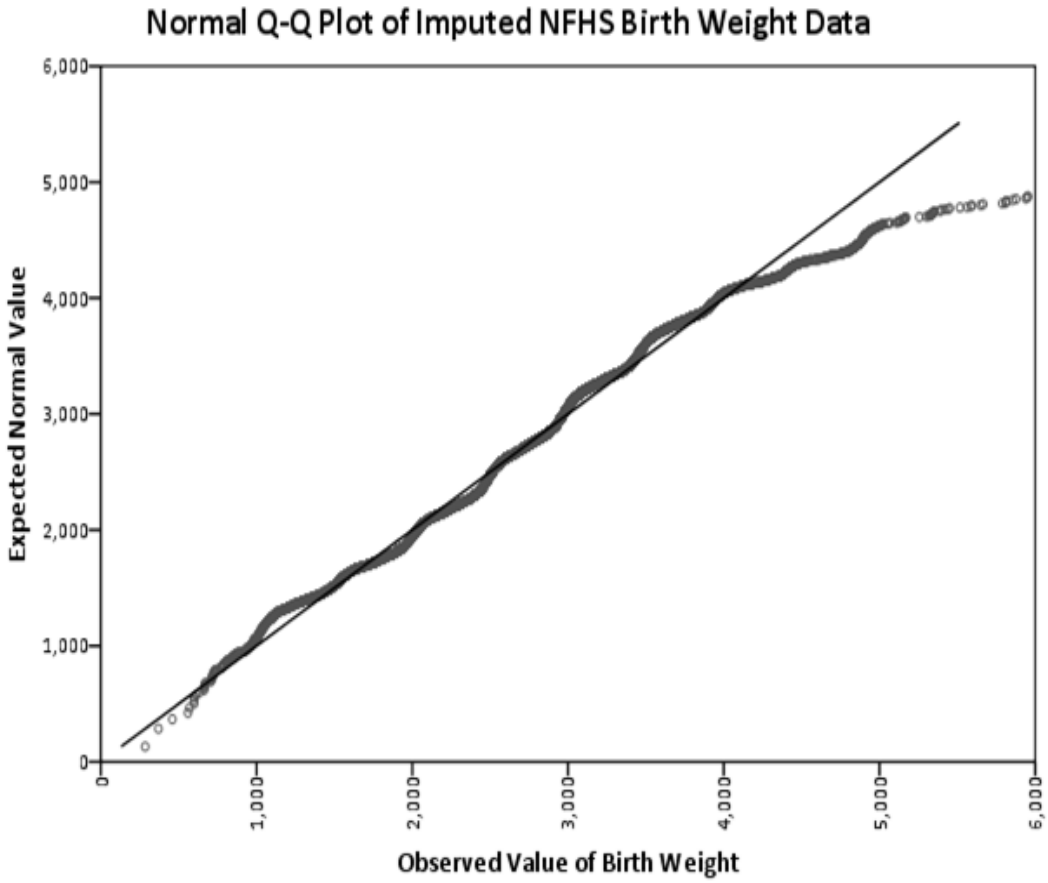

Figure 3.18b: Normal Q-Q Plot of NFHS Birth-weight data – Adjusted using modified statistical calibration and multiple imputation

4.0 MANUSCRIPT 2: CHARACTERISTICS OF BIRTH-WEIGHT DISTRIBUTIONS AND THE BIRTH-WEIGHT – INFANT MORTALITY RELATIONSHIP: A COMPARISON OF INDIA AND THE UNITED STATES

Manuscript in Preparation

R. Margaret Whelan,* Drew Feiner,†
* University of Pittsburgh, Graduate School of Public Health, Department of Epidemiology
† Blade Diagnostics Corporation, Pittsburgh, PA

4.1 ABSTRACT

4.1.1 Background

Birth-weight (BW) distributions possess certain characteristics across populations: a "predominant" Gaussian distribution & a "residual" distribution representing small, preterm babies in the lower tail. The BW-IM relationship is characterized by a J-shaped curve: risk of IM approaching 100% at lowest BWs, lowest risk in the middle of the range. We seek to confirm these characteristics in an Indian population.

4.1.2 Methodology

BW and IM data were taken from the Indian National Family Health Survey III (2005-2006). Indian BW data were statistically adjusted to account for heaping. US vital statistics include linked BW-IM data for the 2004 birth cohort.

4.1.3 Results

Among Indian births (n=20,947), the mean (SD) of the predominant distribution was 2842g (645g). 1.2% of births are in the residual distribution. Among US births (n=2,134,535), the mean (SD) is 3358g (496g) and 3.9% of births are in the residual distribution. Indian and US IM curves have nadirs at 3,152g BW (IM rate 19.8/1000) and 4,016g (IM rate 2.68/1000)

respectively. At 2500 grams (cut-off for low BW), Indian and US rates of IM were 23.8/1000 and 10.6/1000 respectively.

4.1.4 Conclusions

Much higher rates of IM exist in the Indian population, even in the best case (optimum BW, nadir of curve). The rates of IM at the 2500 grams LBW cutoff are not equivalent in India and the United States. Given these higher IM rates in India, the residual distribution should be larger in the Indian population, but it is not, suggesting the typical characteristics of the BW-IM relationship do not hold true. These results question the accepted meaning and use of BW distributions and the LBW cutoff when looking at data from India.

4.2 INTRODUCTION

BW is an exceptionally strong predictor of IM, morbidity and longer term health and development.[3-18] Whether BW is causally related to infant death is currently being debated.[11,42,47,48,50] However, these two variables have been studied extensively and have produced interesting insights into the defining characteristics of the BW distribution and the nature of the BW to IM relationship.

The BW distribution is normal with an extended lower tail.[11,36-38] It can be divided into two separate distributions: a predominant distribution, which represents the normally distributed births born at or after term (37 weeks). The residual distribution represents the small and pre-term babies in the distribution. These distributions are defined by specific parameters, namely mean, standard deviation, percentage in the residual distribution and the upper bound of the residual distribution. The predominant and residual distributions are independent of one another; the mean of the predominant distribution can change without affecting the percentage of babies in the residual distribution.[11,37-40]

Studies of BW and IM have shown an extremely strong association between the two measures. In general, the lower the weight at birth, the higher the risk of IM. This has also been observed on a population level where groups with lower mean BWs (i.e. smokers, African-Americans) often have higher rates of IM.[8,11,19,37,43-45] There is a typical pattern of BW specific

neonatal mortality. Risk approaches 100% for the smallest babies and declines to less than 1% in the middle of the range, which is close to the mean BW for that population. Then, risk increases slightly for the heaviest babies. The same pattern is seen within each gestational-age stratum, indicating that this is not simply a reflection of preterm births at the lowest weights. This pattern of risk invariably produces an IM curve in the shape of an inverse J. [37,40,41] The relationship of these two curves, the BW distribution and weight-specific IM curve, is at the heart of the debate over whether BW is actually on the causal pathway to IM.

In recent years, theoretical evidence has mounted to support the notion that BW is not a causal factor for IM, but that weight and mortality can change together because a common set of factors affects both. This has been supported by evidence that in many populations, as mean BW changes, the weight-specific IM curve changes by equivalent amounts.[11,41-43,47,48,97,98]

While BW and IM have been studied extensively and produced a body of literature defining the nature of the BW to IM relationship, there are few examples of these variables being studied in the developing world.[99-103] This is in large part due to the fact that many developing nations do not routinely record BW as part of a system of vital statistics. In these situations, BW data may be biased if it comes from health care facilities or survey data. In India, a country where BW and IM are issues of chief concern, the major source of BW data is the Indian National Family Health Survey, a representative demographic and health survey of the entire country. Using a statistical method to adjust for some of the biases in this survey data, we have attmepted to compare BW and IM in India and the United States to determine whether these characteristics and assumptions about the relationship of BW to IM hold true.

4.3 METHODS

We obtained linked BW and IM data for the United States through the CDC's National Center for Health Statistics, where they are publically available. The most recent dataset for linked births and deaths is the 2004 birth cohort. We downloaded these data from the CDC's National Center for Health Statistics website.[104] The denominator data included all births in the United States in 2004; these did not include US citizens born outside of the United States or

births to foreign citizens that occurred within the United States. The numerator data consisted of deaths of infants born in 2004 linked to their corresponding birth certificates, whether the death occurred in 2004 or 2005. 98.9% of all infant deaths were linked to the infant's birth certificate, even if birth and death occurred in different states. For the 2004 cohort, 1.1% of all infant death records could not be linked to their corresponding birth certificate. There are differences in the percentage linked by variables such as state and infant age at death. However, there is no way to know if there is a difference in percentage linked by BW and no method to correct for this potential bias.

In order to reduce potential bias in the computation of BW-specific IM rates, the researchers at the National Center for Health Statistics added an imputation for not-recorded BW. If BW was not recorded and the period of gestation was known, a value for BW was assigned taken from the previous record with the same period of gestation, race, sex, and plurality. This imputation reduced the percent of not-recorded responses, reducing (though not eliminating) the potential for underestimation when computing BW-specific IM rates.[105]

We analyzed these US data in comparison to data from the 2005-2006 Indian National Family Health Survey (NFHS-3), a large-scale survey conducted in a representative sample of households throughout India. Three rounds of the survey have been conducted since the first survey in 1992-93. The survey provides state and national information for India on fertility, infant and child mortality, maternal and child health, reproductive health, birth outcomes and other variables. The NFHS was coordinated by the Indian Ministry of Health and Family Welfare. The 2005-2006 Survey (NFHS-3) was the third such survey conducted in India. NFHS-3 asked all women age 15-49 to provide a complete history of their births including the sex, month and year of birth, BW, survival status, and age at the time of the survey for each live birth occurring in the previous five years, up to six births. Data on BW were collected by two methods – from mother's memory and from health records in the home. Data were classified based on their source. Survey researchers in India produced weighting variables using the sample selection probabilities of each household and the response rates, and we were able to apply these weighting variables to all the NFHS data. A significant portion of the BW data were missing in NFHS-3. So, we developed a further weighting variable based on the missing data within each

sampling unit and then applied the weighting variable to those recorded BWs to make this a representative sample of BWs from the period.

Due to heaping of BW at intervals of 500 grams, a phenomenon commonly seen in survey BW data from developing countries, the NFHS data was adjusted using a modified statistical calibration and multiple imputation method that has been developed for this purpose using gold-standard data from rural India.[106]

BW and IM data from US Vital Statistics were compared with NFHS BW and IM data in the following ways. Using software from the Wilcox group at the National Institute of Environmental Health Sciences,[38,42] parameters describing the BW frequency distribution were calculated and compared for both the US and India. The software identified the "predominant" and "residual" BW distributions (described further below) and calculated key parameters for each distribution. This technique is based on an underlying multinomial sampling distribution and involved estimating parameters in a mixture model for the multinomial bin probabilities after having chosen the support of the residual distribution with a model selection criterion.

There are two components of any BW distribution, with a total of four parameters. The "predominant" distribution is Gaussian (normal), with the parameters mean and standard deviation. The "residual" distribution is a portion lying outside the lower tail of the "predominant" distribution, which can be summarized as a percent of the whole population. One can also look at the optimum truncation point, or upper bound of the residual distribution. Births in the "predominant" distribution can be interpreted as term births while births in the "residual" distribution are both small and preterm and thus at highest risk of death. Therefore, mean and standard deviation of the "predominant" distribution as well as the percent in the "residual" distribution were compared for Indian and US BW distributions.

US and Indian BW and IM data were also be compared based on basic parameters of the BW distribution and IM curve including mean BW, standard deviation of the BW distribution, lowest rate of IM, optimal BW (BW at the lowest rate of IM), and difference between mean BW and optimal BW. Normality of both curves was assessed using a Q-Q normality plot and best-fit software available in the MATHLAB software. The two weight distributions were adjusted to a standard z-scale (with means set to zero and standard deviations to 1) and the weight-specific

mortality rates were adjusted to the z-scale as well. Weight distributions and IM curves were then further compared based on this adjusted scale.

Two thousand five hundred grams has been recognized as a relatively arbitrary cut-off for dichotomization of BW data, but it is still used very regularly in research and policy-making. We assessed the risk of IM at 2500 grams in these two populations and determined if the risk of death for Indian infants born below 2500 grams is different from the risk of death for LBW babies in the US cohort.

4.4 RESULTS

In the US birth cohort from 2004, 4,118,348 births were recorded. This represents the complete birth cohort of babies born in the United States to US Citizens in 2004. Births were linked to deaths via birth and death certificates. There were 27,642 infant deaths in the United States in the 2004 birth cohort, resulting in an IM rate of 6.71 per 1000 live births. The mean BW was 3,281 grams with a standard deviation of 602.9 grams. (Table 4.1)

In NFHS-3, 56,437 births were recorded across India in the five years preceding the survey. A BW was recorded for 34 percent of babies born in the five years preceding the survey; this BW came either from a BW recorded on a health card or from the mother's memory (recall). The great majority (85.7%) of BWs recorded came from maternal recall. The proportion of births with a reported BW is 60 percent in urban areas and 25 percent in rural areas. Response rates were not the same across all the sampling units. Therefore, a unique weighting factor was available within the NFHS dataset that we applied to each case based on the sampling unit and response rate. We also applied a second weighting factor to adjust for the missing BW data within each sampling unit. There was significant heaping in the raw NFHS BW data (Figure 4.1a), which was adjusted using a modified statistical calibration and a new data set was imputed (Figure 4.1b).

There were 615 infant deaths in the NFHS 3 data, resulting in a rate of 31.97 infant deaths per 1000 live births in India from 2000-2006. This rate is approximately four and a half

times higher than the IM rate in the United States for 2004. The mean of the BW distribution was 2,769 grams with a standard deviation of 739.1 grams. (Table 4.1)

The first characteristic of interest under discussion here are the US and Indian BW distributions. The normality of both these distributions was assessed using a Q-Q normality plot (Figure 4.2). Both plots demonstrate that the theoretical normal distribution fits the BW measurements. The US data from 2004 shows the measurements dipping below the reference line at the lowest BWs and rising above the reference line at the highest BWs. This can be interpreted to mean that there is a longer tail to the left (smaller BWs) and a shortened tail of the distribution to the right. This is to be expected of BW distributions. The data from India follow the reference line for the majority of the measurements. The largest BWs rise above the reference line, suggesting a shortened tail to the right of the distribution. In general, both plots suggest that assuming an overall normal distribution for both Indian and US BW data is appropriate.

The next analysis produced figures and parameters for the predominant and residual BW distributions. The US predominant distribution is bell shaped with a mean of 3,357g and a standard deviation of 489g. The residual distribution for the US holds 4.3% of the births. The "optimum truncation point" or upper bound of the residual distribution is 2,500 grams. The biological interpretation of the residual distribution is that it holds the small and preterm births, those babies at highest risk of death. (Figure 4.3, Table 4.2)

The Indian predominant and residual distributions are different from the pattern seen in the United States. Compared to the US, the Indian predominant distribution is left-shifted and wider due to a lower mean of 2,910 g and a larger standard deviation of 645 g. The portion of babies in the residual distribution is smaller with only 1.6% of births. The optimum truncation point for the residual distribution is 1,700 g. (Figure 4.4, Table 4.2).

Having analyzed the BW distributions alone, we next looked at the BW distribution as it compares to the weight specific IM curve in each population. In the US birth cohort, the IM curve has an inverse J-shape. The mean of the BW distribution is 3,281 g and the corresponding IM at the mean BW is 2.00 per 1,000 live births. The lowest rate of IM in the US was 1.42 deaths per 1,000 live births and this corresponded to a BW of 4,020 g (the optimum BW). The

difference, in grams, between the mean and optimum BW was 739 grams. (Table 4.1, Figure 4.5a)

For the Indian population of births, the BW distribution is again left-shifted and wider with a mean BW of 2,769 grams. The weight-specific IM curve has a different shape from the US cohort's J-shaped curve; it is more U-shaped. The rate of IM that corresponds to the mean BW in the Indian cohort is 23.69 per 1,000 live births. The lowest rate of IM is 23.58 per 1,000 live births and this corresponds to a BW of 2,934 grams (the optimum BW). The difference, in grams, between the mean and optimum BW was 147 grams. (Table 4.1, Figure 4.5b)

When both BW distributions (using frequency percent for the y-axis) and weight-specific IM curves are plotted together, the higher rates of IM become clear. The IM curves intersect at a BW of 1,304 grams suggesting that the risk of IM for US babies weighing less than this is higher than the risk of IM for these very small Indian babies. (Figure 4.6).

However, when both the BW distributions and weight-specific IM curves are adjusted to a z-scale, with a mean of zero and standard deviation of one, the relationship is clarified (Figure 4.7). Through this technique, when the BW distributions are forced to over-lie one another, the relationship of the IM curves is clear. The rate of IM in India is higher than in the United States across the whole range of BWs. However, the two IM curves have different shapes. The US IM curve is steeper in shape while the Indian curve has a much flatter slope across the majority of the range of BWs. This means that the difference in rates of IM between India and the US increases as one moves towards the middle of the range of BWs. While the rate of IM gets markedly lower for US babies near the mean BW, the rate of IM does not decrease nearly as much for the Indian cohort. (Figure 4.7)

4.5 DISCUSSION

From the outset, these data and analyses suggest that there are significant differences between the US and Indian cohorts in terms of the BW to IM relationship. While essentially normal in shape, the Indian BW distribution is left shifted and wider when compared with the

US cohort. The mean Indian BW is 500g less than the mean US BW. While this difference is not unexpected, the results from the analysis of the predominant and residual distributions are.

From previous studies of this type, one expects to see about 2-7% of babies in the residual distribution.[39] With only 1.6% of babies from the Indian cohort falling in this distribution, this is outside of the normal range. And, given that the residual distribution represents small, pre-term babies, those at highest risk for death, one would expect that a population with such high rates of IM would have a large proportion in the residual distribution. Especially in comparison to the 4.3% of babies in the US residual distribution, this result from the Indian cohort is counter-intuitive.

We must acknowledge that these results could be due to mechanical error of some sort. Perhaps because the Indian BW distribution is so left-shifted and encroaches so much more onto the residual distribution, the software, which was developed using data from the developed world, may not be able to accurately determine the parameters of interest. However, there is nothing in the statistics or assumptions of the analysis to suggest that this type of analysis must be restricted to BW distributions with particular characteristics.

Another interpretation of these results is that in India, BW does not play quite as large a role in the risk of IM. Rates of IM in the US have decreased substantially over the years, in large part due to a reduction in infectious and environmental insults that could kill a newborn. These insults, like vaccine preventable illnesses, diarrheal diseases and acute respiratory infections are still highly prevalent in India. So, these data may be suggesting that, in the face of so many other factors that increase the risk of IM, BW contributes less to the total risk of IM and produces results like those shown above.

Other results to note from this analysis are the values for the optimum truncation points. This parameter is the upper bound of the residual distribution. In the US cohort the optimum truncation point is 2,500 grams, the cut-off for the designation LBW. In India, the optimum truncation point is 1,700 grams, fully 700 grams lower. While these numbers do not relate directly to the cut-off point for LBW, they may lend some insight into what an equivalent cut-off point in India might be.

The second analysis of the relationship between the BW distribution and weight-specific IM rates also had some interesting results. These data show much lower rates of IM across the whole range of weights in the US cohort compared to India. However, the IM curve in India is U-shaped, which is different from the expected J-shaped curve that can be seen in the US data.

Another difference is apparent when looking at the optimum BW versus the mean BW in both populations. While the biologic mechanism underlying this phenomenon is not understood, it is common in most populations to see a difference between the optimum BW and the mean BW in a population. Researchers have postulated that this may indicate a biological struggle between the fetus, which is trying to grow, and the mother, who has to deliver the fetus and needs it to remain reasonably small. The difference between mean and optimum BWs in India is very small (147 grams). We do not believe that this means there is no fetal-maternal struggle for size of the baby. Rather, we interpret this result to show that, once again, BW is not contributing to risk of IM in this Indian population in the same way it does in the US population. In India, across the range of BWs from 2,000 grams to 3,500 grams, the rate of IM varies very little. Because the optimal BW is determined by the lowest rate of IM, the separation between mean and optimal BWs is very small. Around the mean, there simply is not much variation in rates of IM.

With the combined US and Indian data (Figure 4.6), one can clearly see the shift of the Indian BW distribution to the left. The intersection of the two IM curves is striking in these results. These types of crossing IM curves have confounded researchers for many years and led to what is commonly called the “Low Birth-weight Paradox,” wherein the intersection is interpreted to mean that for the smallest babies, they are better off being born to the group with the higher overall IM. In this case, the intersection could be interpreted to mean that, for babies born less than 1300 grams, it is better to be born in India because the rate of IM for those tiny babies is smaller than it is in the United States. This interpretation of intersecting IM lines has now been contradicted[46,107-109] and the method of using a z-scale adjustment gives us a way to look at two IM curves that are actually comparable when the BW distributions are forced to a fit where the mean is zero and the standard deviation is one.

This z-scale adjustment removes the intersection of the IM curves, but it does not produce results seen in other populations. In populations which typically have higher rates of IM and lower overall BW, like smokers, people at high altitude or the general US population from 50 years ago, the z-scale adjustment shows IM curves of the same shape that either overlap or are parallel, in the case of higher IM rates across the range of BWs. When this same technique is applied to the Indian and US data, the IM curves do not converge. Instead, this method further illustrates the different shapes of the IM curves and highlights the increasing difference in IM rate as the curves approach their respective nadirs and the fact that the change in IM across BWs is not as great in India as it is in the US. This again suggests that BW is not as great a contributor or indicator of IM in the Indian cohort as it is in the United States cohort.

Together, these results present an interesting picture of the nature of the relationship of BW to IM in India as compared to the United States. These results suggest the BW-IM relationship is different in India than the US. BW does not appear to contribute as much to the risk of IM in India, as shown through the relatively small changes in IM rate across most of the range of BWs. We have hypothesized that this may be the case because highly prevalent environmental and infectious insults to the infant are driving the high IM rates and that programmed characteristics, like BW, simply do not play as large a role in determining the infant's survival.

Wilcox and Russell have suggested that, on a population level, BW is not on the causal pathway to IM. They support this hypothesis with evidence that, while a change in BW may be associated with increased risk of IM or a change in perinatal health, it is not through the change in BW that the health effect occurs. They suggest that BW and IM change together and are highly associated because the same causal factors are affecting them both. They also show that BW can change without an effect on mortality. (Figure 4.8)

These data lend support to this hypothesis to some extent. The small variation in rate of IM across the middle of the range of BWs supports the notion that BW is not causing IM in this population. It is also appropriate to reason that similar causal factors may lead to the BW and IM found in this population. However, the decrease in IM variation across the BWs suggests

that IM is not responding to the common causal factor to the same extent that BW is responding. If we could measure that actual causal factor, the strength of the associations would be quite different to result in this differential response from the BW distribution and IM curve.

Most of the information we have as researchers about BW and IM comes from populations in developed countries, like the United States and Norway. To our knowledge, there is no study in the literature that discusses BW distributions and the BW-IM relationship in India. Likewise, to our knowledge there is no published study that seeks to confirm the published characteristics of BW and IM in an Indian population. Since there are significant differences between the US and Indian BW distributions and BW-IM relationship, this analysis provides an interesting addition to the literature and raises important questions about the importance of BW is as an indicator of IM risk and population health in India and perhaps in other developing countries.

From a larger, policy perspective, these data and results challenge the current practices and policies of the Indian government which seek to reduce rates of IM by increasing BW through maternal feeding campaigns. These data suggest that scarce resources should instead be devoted to interventions to help the infants after they are born as opposed to attempting to increase their BWs.

4.6 TABLES AND FIGURES

Table 4.1: Cohort Characteristics

	US Vital Statistics 2004	Indian NFHS 3‡
Number of births	4,118,348	56,437
Births with recorded BWs		19,237 (34.1%)
BWs from maternal recall n(%)		16,482 (85.7%)
Number of infant deaths	27,642	615
Overall IM rate (per 1000)	6.71	31.97
BW distribution mean (SD)	3,281 (602.9)	2,769 (739.1)
IM at mean BW (per 1000 births)	2.00	23.69
Lowest IM rate (per 1000 births)	1.42	23.58
Optimal BW (lowest IM rate)	4,020 g	2,934 g
Difference of BW distribution mean and optimal BW	739 g	147 g
IM rate at 2500 g BW (per 1000 births)	5.45	25.23

‡ Reported statistics are based on the weighted sample cohort

*Cases where BWs were not recorded in US Vital Statistics were imputed, resulting in a complete set of 4,118,348 births in the cohort

Table 4.2: Characteristics of BW distributions

	US Vital Statistics 2004	Indian NFHS 3‡
Predominant Distribution Mean (SD)	3,357 g (489 g)	2,910 g (645 g)
Residual Distribution Percentage	4.3%	1.6%
Optimum Truncation Point	2,500 g	1,700 g

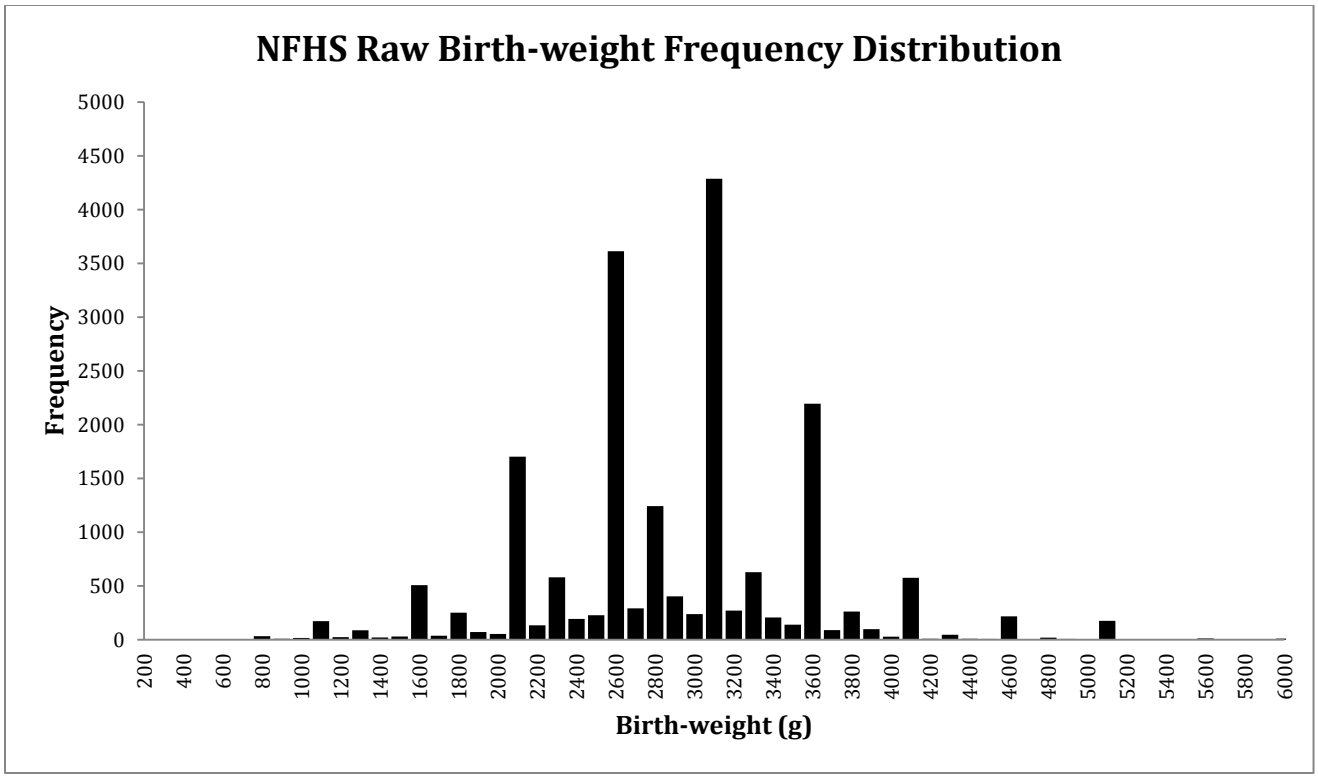

Figure 4.1a: BW Distribution for raw and imputed NFHS Data - Raw NFHS Data

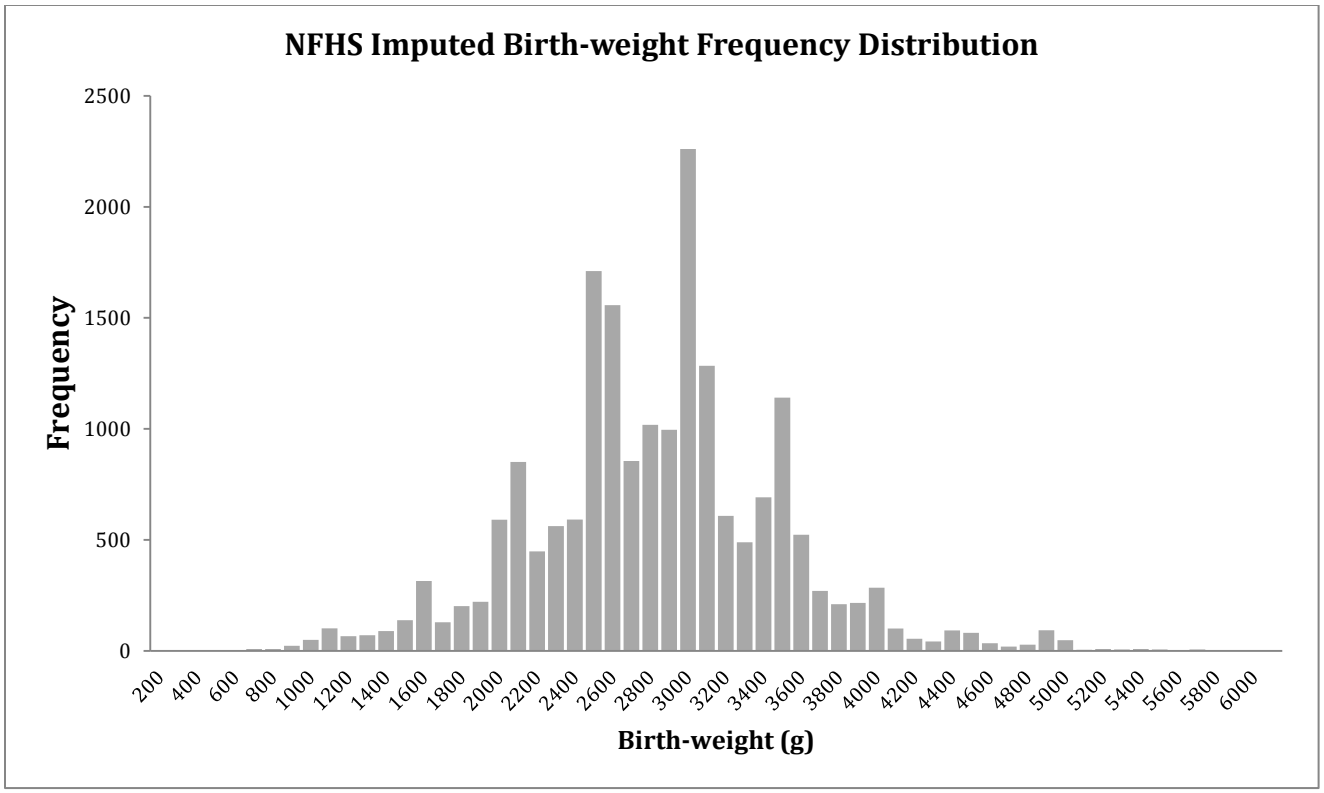

Figure 4.1b: BW Distribution for raw and imputed NFHS Data - Imputed NFHS Data

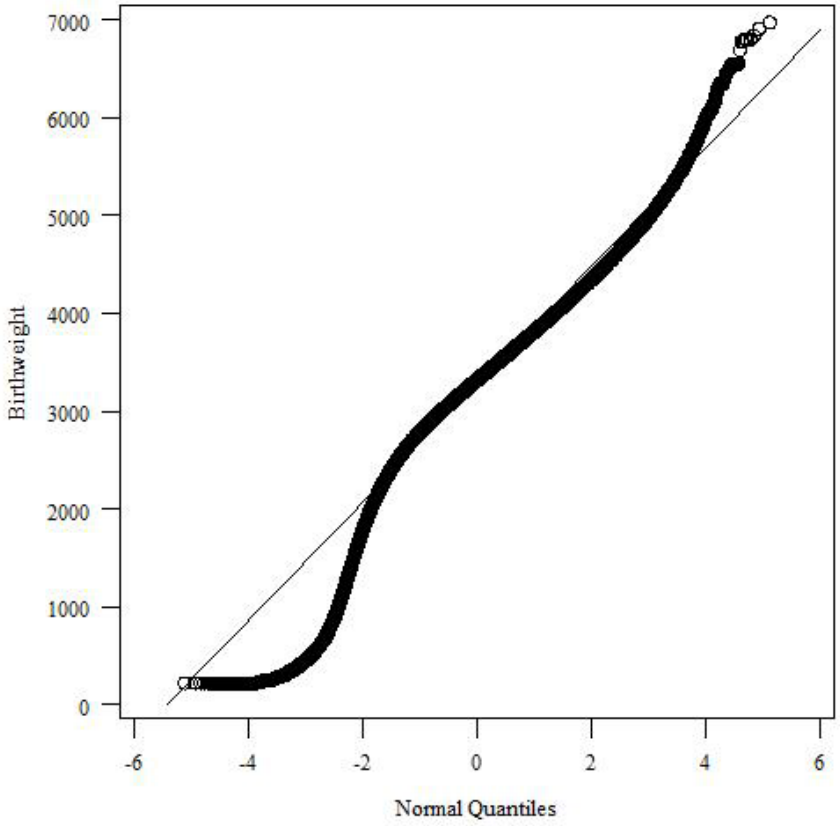

Figure 4.2a: Normal Q-Q Plots of Birth-weight for India (NFHS 3) and the United States (2004) - United States

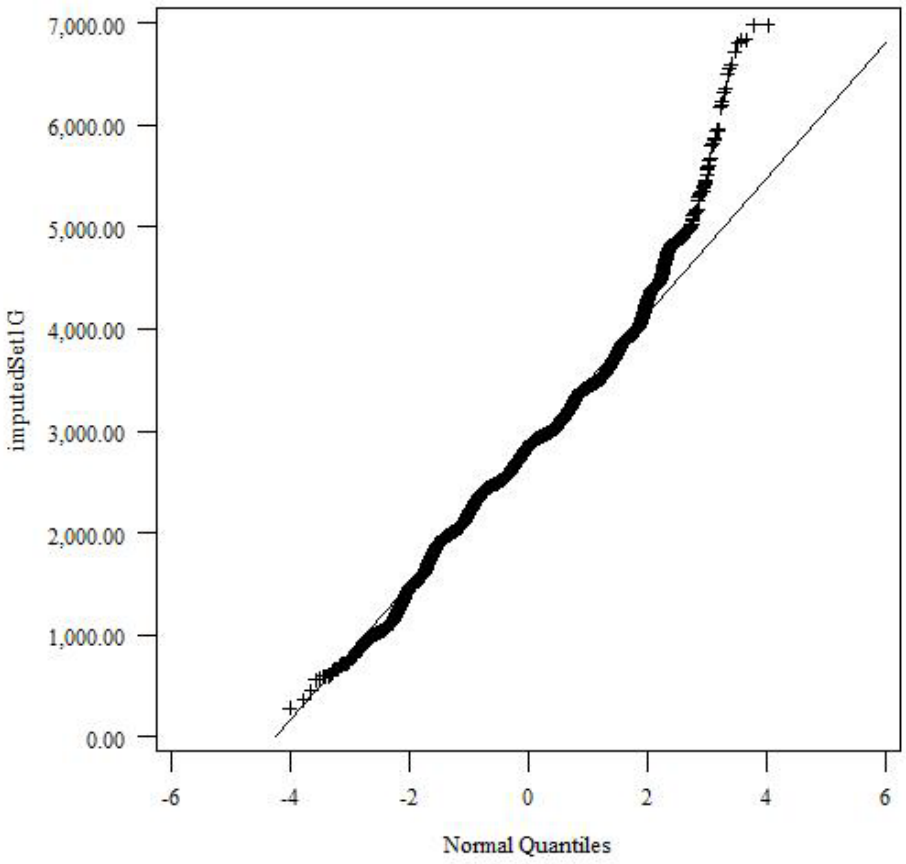

Figure 4.2b: Normal Q-Q Plots of Birth-weight for India (NFHS 3) and the United States (2004) - India, NFHS 3, 2005-2006

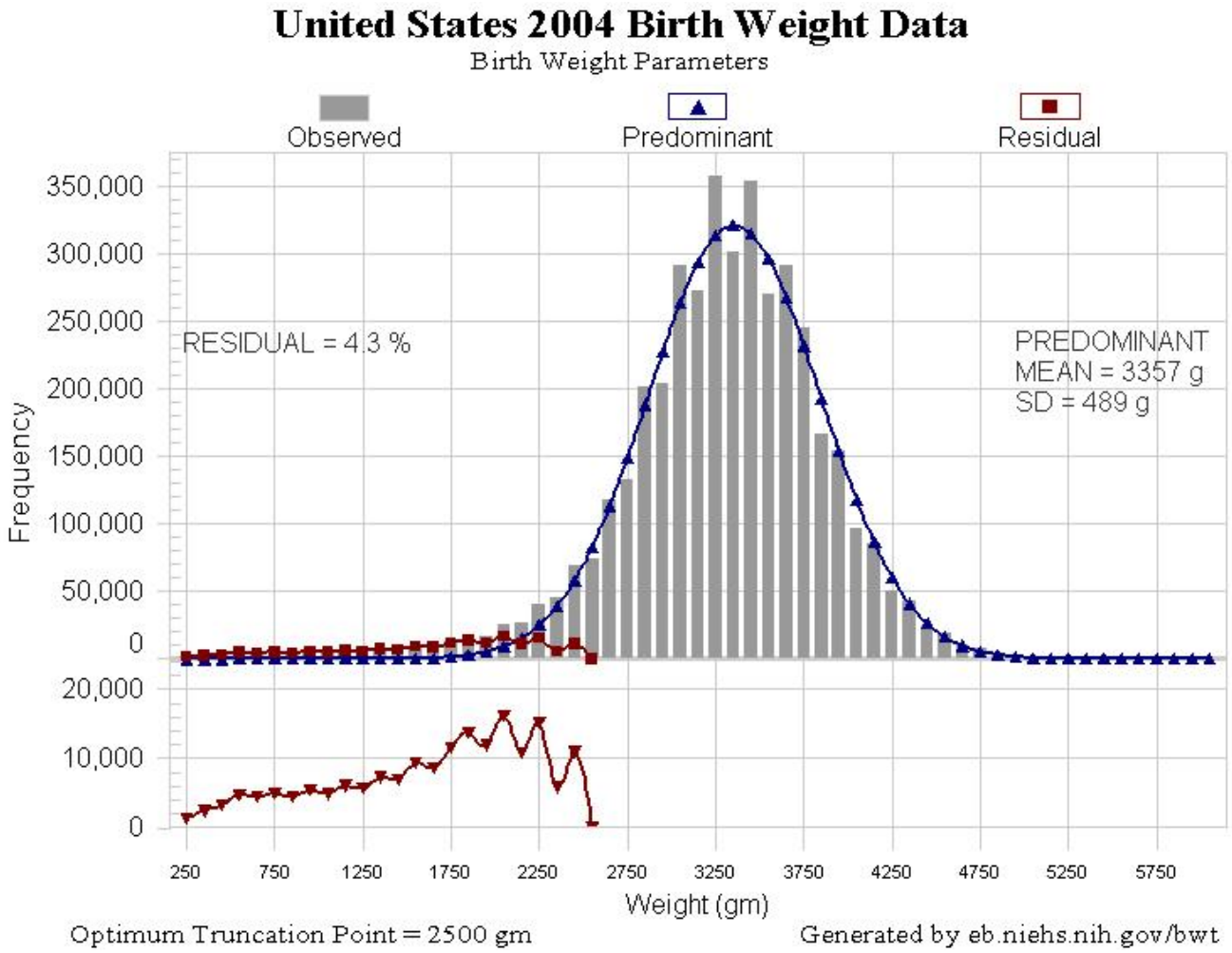

Figure 4.3 United States Birth-weight Distribution

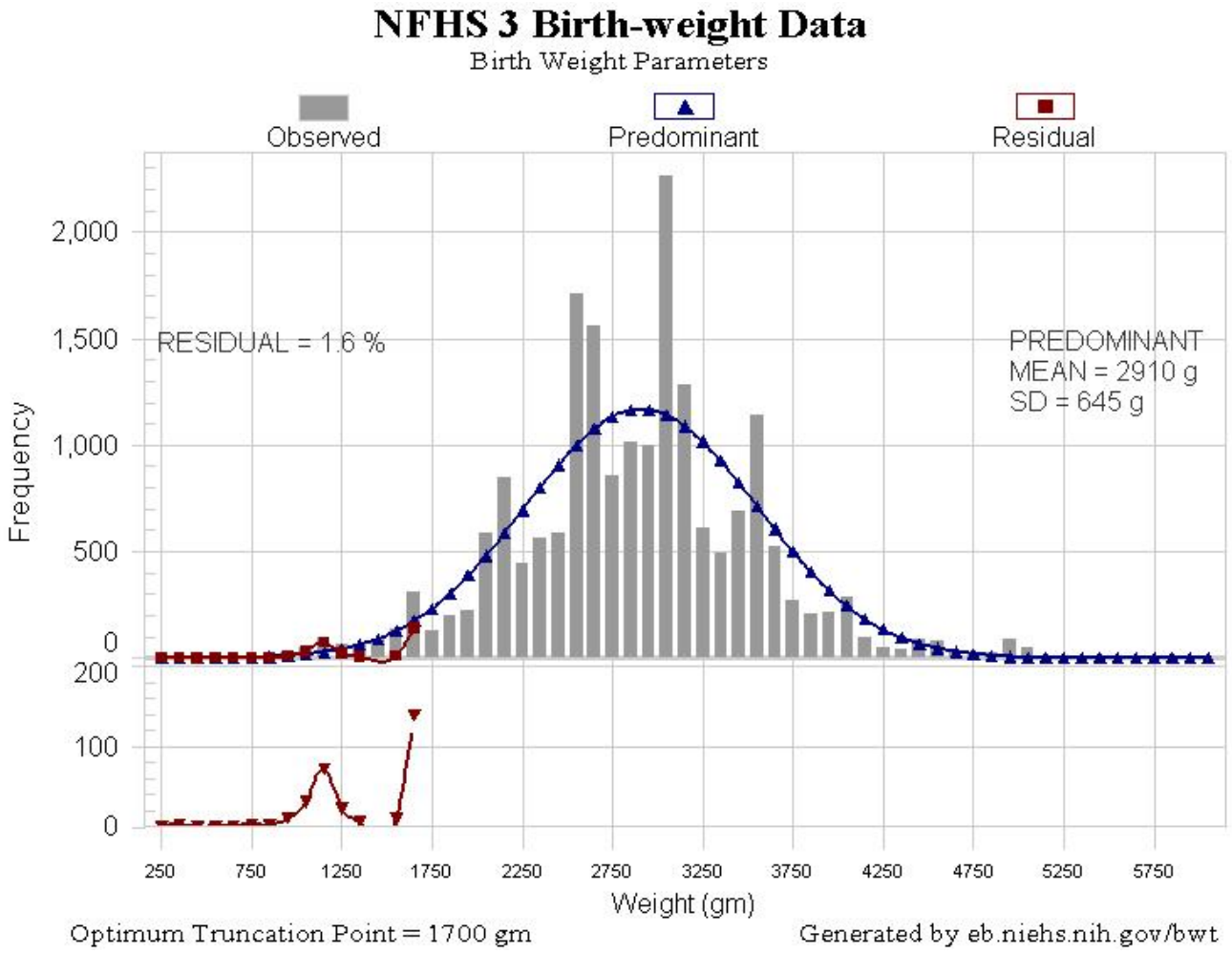

Figure 4.4 Indian Birth-weight distribution

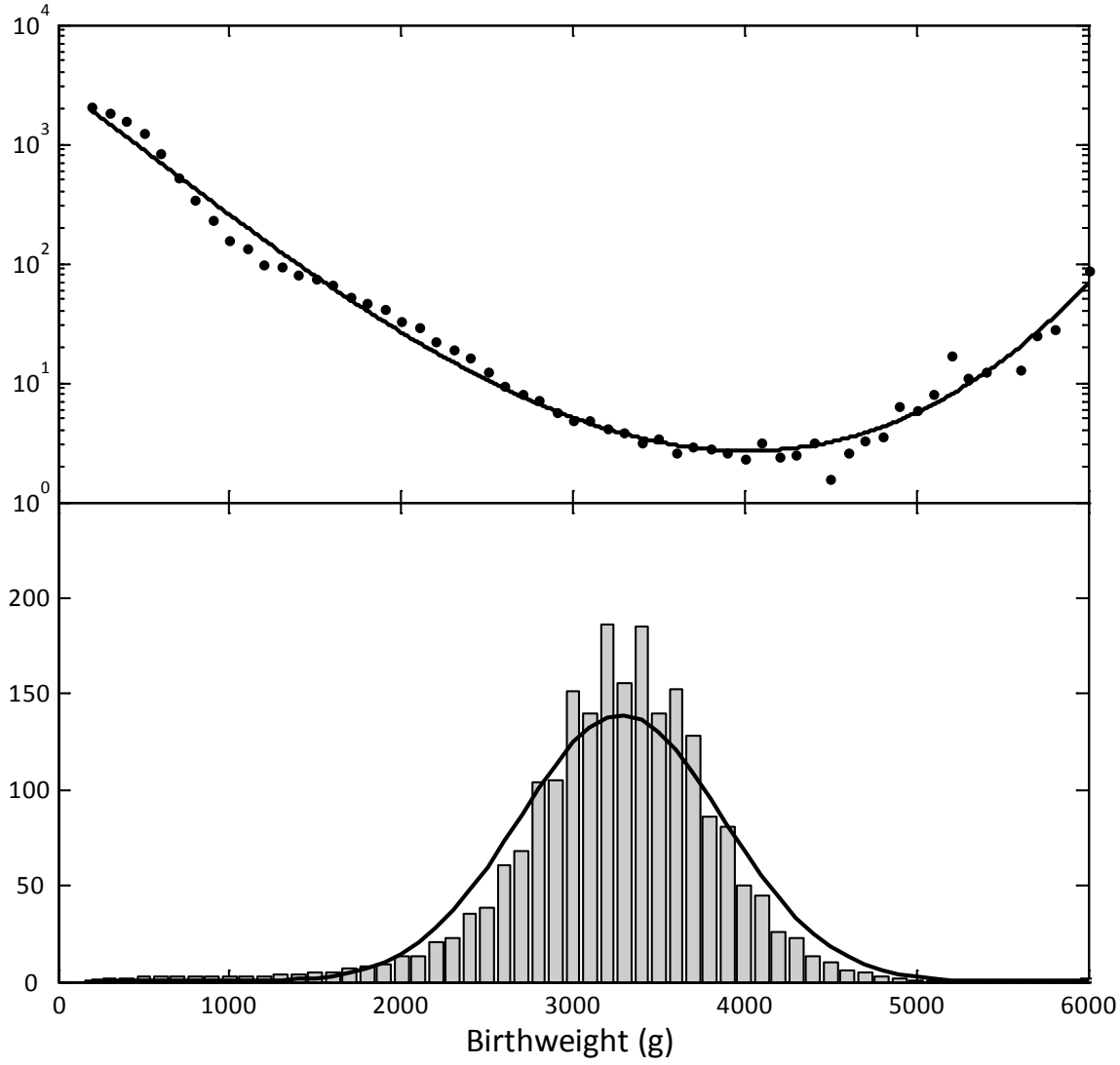

Figure 4.5a: Frequency distribution of birthweight and weight specific neonatal mortality for India and the United States – United States

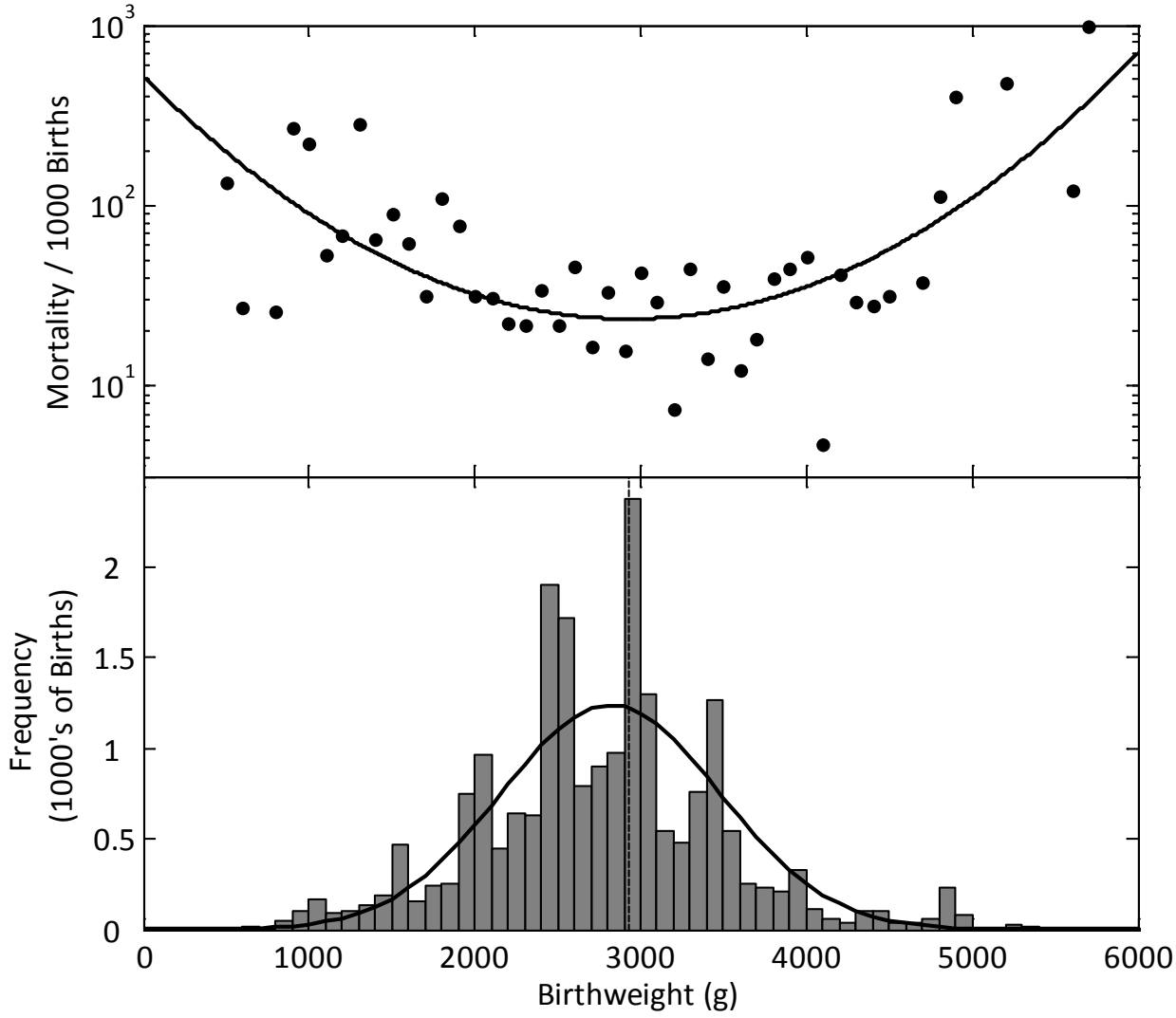

Figure 4.5b: Frequency distribution of birthweight and weight specific neonatal mortality for India and the United States – India

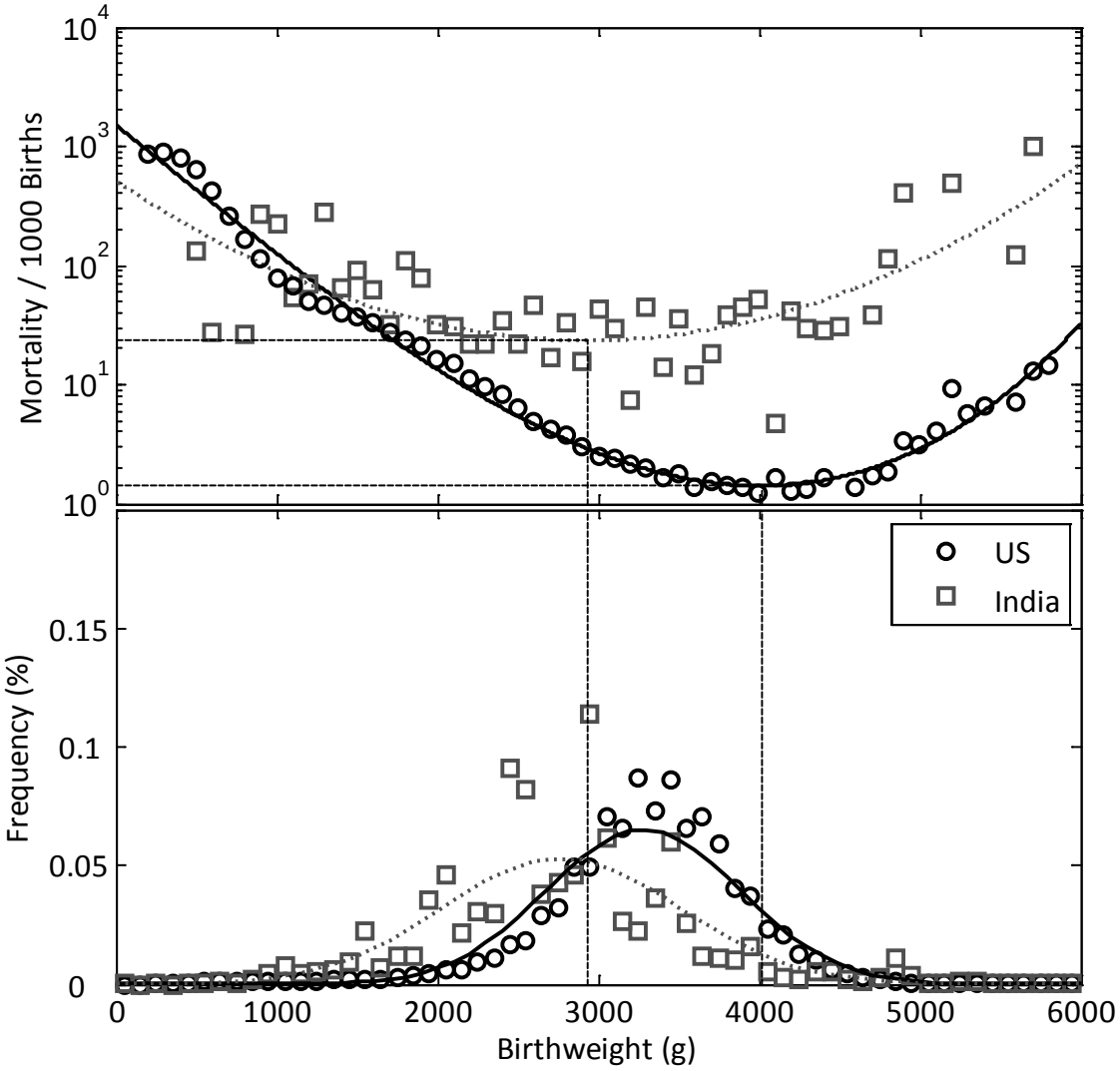

Figure 4.6: Frequency distribution of birthweight and weight specific neonatal mortality for India and the United States

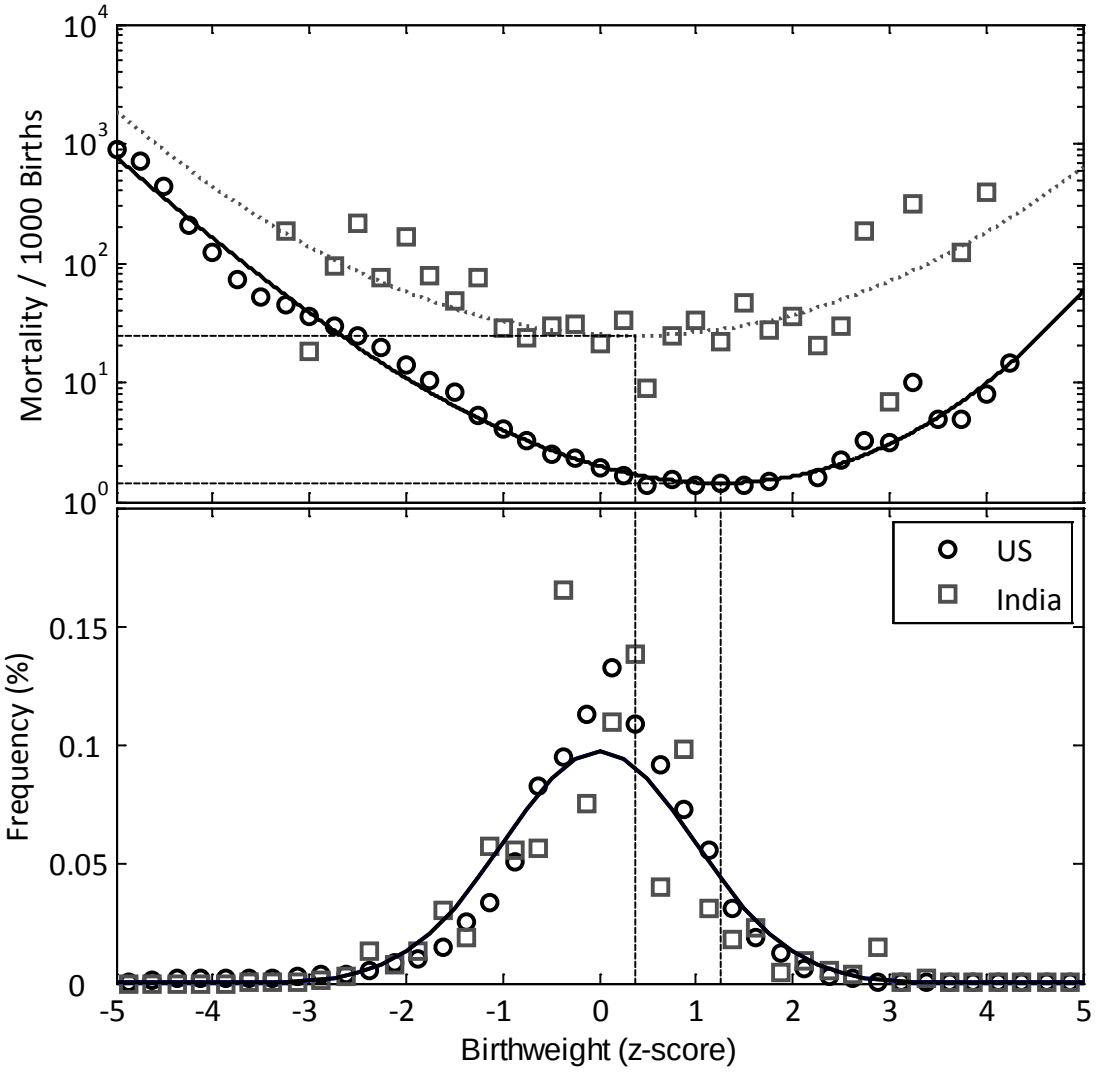

Figure 4.7: Frequency distribution of birthweight and weight specific neonatal mortality for India and the United States after adjustment to a z-scale of birthweight

Usual assumption: BW on the causal pathway

Wilcox-Russel Hypothesis: BW not on causal pathway, common causal factor responsible for changes in BW and IM

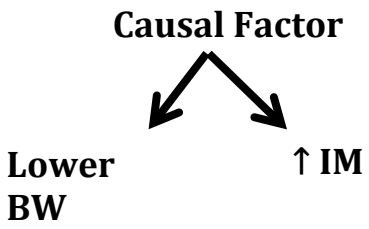

Figure 4.8: Directed Acyclic Graphs of the BW – IM relationship

5.0 MANUSCRIPT 3: HOUSEHOLD SANITATION AND HYGIENE STRONGLY ASSOCIATED WITH BIRTH-WEIGHT AND INFANT MORTALITY IN A RURAL INDIAN COHORT

Manuscript in Preparation

R. Margaret Whelan,* Drew Feiner,† Purushotham Reddy,◊ K. Balasubramanian,◊
Clareann H. Bunker,* PS Reddy◊
* University of Pittsburgh, Graduate School of Public Health, Department of Epidemiology
† Blade Diagnostics Corporation, Pittsburgh, PA
◊ MediCiti Institute of Medical Sciences, SHARE India, Andhra Pradesh, India

5.1 ABSTRACT

5.1.1 Background

Rates of infant mortality (IM) are extremely high in India (50.3 per 1,000 live births), seven times the rate in the United States and 20 times the rate in Sweden. Birth-weight (BW) has been shown to be an extremely strong predictor of IM in all populations, but it is also in question whether BW is on the causal pathway to IM. Studying a cohort in rural South India, we sought to identify factors associated with IM apart from BW that could provide opportunities for intervention to reduce the high rates of infant death.

5.1.2 Methodology

Since 2004, residents of Medchal Mandal (county) in the Ranga Reddy district north of Hyderabad, Andhra Pradesh, India have been followed and data collected on births, deaths, demographic factors and other measures related to general and reproductive health. In 2009 a more in-depth, cross-sectional survey of the families in the Mandal was completed. Using data from both these sources, we analyzed the associations of certain SES, demographic and sanitation/hygiene factors with the outcomes IM and very small BW (<1.7 kg).

5.1.3 Results

Some SES and demographic variables, including maternal age, sex of the child, place of birth, year of birth, family type, parents' education and caste, were significantly associated with very small BW and IM. Sanitation/hygiene factors, including facilities for human waste disposal, source of drinking water, sewage disposal, cooking fuel and garbage disposal were also significantly associated with very small BW and IM. After controlling for the significant SES and demographic factors, the associations between sanitation and BW and sanitation and IM persisted. The association of sanitation to IM persisted after controlling for BW.

5.1.4 Conclusions

These results suggest that sanitation/hygiene factors are associated with very small BW and IM in this rural Indian population. We believe that sanitation and hygiene factors are acting as surrogates for exposure to infectious disease in this case and that interventions to improve sanitation and hygiene in this population could reduce the high rates of IM.

5.2 INTRODUCTION

In 2009, the rate of IM in India was 50.3 per 1,000 live births, more than seven times the IM rate in the United States (6.8/1000) and more than 20 times the IM rate in Sweden (2.3/1000).[51] Despite efforts to reduce the numbers of infant deaths through social welfare programs and other interventions,[53] IM rates remain unacceptably high in India.

IM has been studied in many different populations and methods have been developed to reduce the numbers of infant deaths, primarily in developed countries.[54-56] However, given such high rates of infant deaths, surprisingly few studies have undertaken to determine what is responsible for these deaths in India, and particularly in rural India.

BW is the strongest recorded predictor of IM. Babies weighing less than 1,500 g have a mortality risk at least 100-fold higher than babies at the optimum weight (the weight associated with the lowest mortality). This association is seen at both the individual and population level. BW is extremely predictive of an individual baby's survival. Also, groups with lower mean BW

often have higher IM. [3-18,41,48] However, the true nature of the relationship between BW and IM is controversial. Researchers are divided over whether or not IM is caused directly by BW and if so, what the biological mechanism is. This is a particularly important question in India where efforts to reduce IM have focused on increasing BW.[57,61] However, this approach, which assumes that risk of IM is programmed from birth (in the infant's BW), does not address insults after the baby is born which could be contributing significantly to IM.

With this study, we are interested in the role of factors other than BW and their contribution to IM in a rural Indian population. It has been established that risk for IM is increased in lower SES and rural populations in India and has also been closely linked to nutrition.[57,59,61,89,110] Studies of cause specific IM in India identify infectious diseases as the leading direct cause of infant death.[58,60,110] Nutrition as a risk factor for lower BW in India has been studied at length.[58,85,90,111-113] Other determinants of BW in India are less well defined, but include lower SES,[85,114] maternal age,[85] pesticide exposure,[115] and maternal co-morbidities.[116-118]

In this study, we hypothesize that infectious disease plays a role in the high rates of IM and lower BW in this rural Indian cohort and that household sanitation and hygiene are surrogate markers of these infectious diseases. Household sanitation measures like toilet facilities and cooking fuel are viewed here as stand-ins for exposures to diarrheal diseases and pathogens causing upper respiratory infections. We also believe that studying these sanitation/hygiene factors and their associations with both IM and BW will provide insight into the BW-IM relationship in this population.

5.3 EXPERIMENTAL DESIGN AND METHODS

5.3.1 Study Setting and Design: The Rural Effective Affordable Comprehensive Healthcare (REACH) Project

The REACH Project is a working model of healthcare delivery to rural populations that provides health education, immunizations, access to antenatal care and access to primary and

tertiary care for a population of 43,270 people in 41 villages in the Medchal Mandal in the Ranga Reddy District on the northern outskirts of Hyderabad, Andhra Pradesh, India. REACH has enumerated all households and household members in these communities and mapped each dwelling by GIS.

In each of the villages in the catchment, a Community Health Volunteer (CHV) has been recruited to visit each family in the village and conduct monthly interviews. In this way, a detailed database is maintained, which contains information on family make-up, births, deaths, marriages, pregnancies, health, socioeconomic and demographic factors. (For a complete list of collected measures, see Table 1.) This database has been continually updated since 2004.

In 2009, a more in-depth survey was completed in all the households in the REACH catchment area. This Medchal Family Health Survey (MFHS) provides cross-sectional data that is more detailed in regards to health issues like birth histories, disease exposures and health care utilization. (For a complete list of collected measures, see Table 1.) The cross section data from the MFHS was linked to the REACH database using a common identifying variable.

For this study, we have used primarily data from the REACH database with a few additional measures taken from the MFHS. The outcomes of interest are BW and IM and the exposures of interest are sanitation and hygiene variables.

5.3.2 Methods

A previous study of BW distribution in this population and other survey collected BW data in India revealed heaping (also called digit preference), a pattern of misreporting in which the distribution of a number reported by respondents shows implausibly large frequencies of particular values.[20,23,106,119] In this case, the reported BWs showed large heaps at 100 and 500 gram intervals.

To adjust for heaping, the REACH BW data was adjusted using a modified calibration and multiple imputation (MCMI) technique developed specifically for this purpose.[106] This method is particularly appropriate because the validation data set that forms the basis for the MCMI technique is taken, in part, from births to women in the REACH catchment area. All calculations

involving BW were completed using three imputed data sets and the resulting statistics were combined using Rubin's rules for combining the results of multiple imputation.[96,120]

The key outcomes of interest were BW and IM. BW was considered in three forms: first, as a continuous variable, second as a dichotomous variable (LBW <2.5kg and NBW ≥ 2.5kg). Finally we divided BW into four categories based on parameters from the BW distribution previously described (Manuscript 2). The category Very Small BW (<1.7kg) was based on the truncation point of the residual BW distribution; the categories Small BW [1.7kg – 2.5kg), Optimum BW [2.5kg – 3.36kg) and large BW (≥ 3.36kg) were based on the optimum BW (lowest risk of IM: 2,934kg) and standard deviation (739.1g) of the BW distribution.

The covariates of interest, those measures related to sanitation and hygiene, were assessed to determine if they could be combined into relevant scales representing overall level of sanitation/hygiene. Exploratory factor analysis was used to determine if valid scales could be created and used. Chronbach's alpha was used to determine levels of reliability for any potential scales.

For categorical variables, a reference category was set for comparison in the statistical analysis. The demographer (K. Balasubramanian) who designed the data collection tools used for REACH and the MFHS ranked the response items from worst to best. The best option was set as the reference category and tested. In certain instances the reference category had to be changed when there were too few records to estimate an effect. All analyses of categorical variables were conducted in comparison to these "best response" reference categories.

First, the association between BW and IM was confirmed in this population using logistic regression. We then considered BW and IM as outcomes. We analyzed the outcome variables using univariate analysis with each of the demographic, socioeconomic (SES) and sanitation/hygiene variables of interest. In all cases, measures from the REACH database were given preference over measures from the cross-sectional Medchal Family Health Survey. We cross-tabulated the categorical independent variables using Pearson's χ^2 test or the Fisher exact test where appropriate.

Generalized linear models were used to analyze the univariate relationship between categorical dependent variables and the continuous outcome of BW. Logistic regression was used

to analyze associations between categorical dependent variables and the categorical BW and IM outcomes. Pearson's rho was used to assess co-linearity between continuous dependent variables and a correlation matrix was used to insure limited co-linearity between continuous or ordinal/scalar variables. We assessed overlap and confounding in the categorical variables using cross tabulation and logistic regression analysis. Special attention was paid to missing data, especially from the MFHS. Missing values are noted in the results tables and we also included the missing values as a separate response category (unknown missing response) in the models.

Using the above method, we identified sanitation/hygiene variables as potentially significant predictors and demographic/SES variables as potentially significant confounders based on a p-value of 0.20 or less. For each of the sanitation/hygiene predictors of interest we created separate models using linear and logistic regression where appropriate to predict BW and IM while controlling for demographic and SES variables. Finally, for those sanitation/hygiene-IM models showing significant associations, we controlled for BW to see if this addition to the models changed the significance, direction or strength of the associations.

Every analysis that involved BW (continuous, dichotomous or categorical) as a predictor or outcome was completed three times using each of three imputed data sets. The resulting odds ratios, confidence intervals and p-values resulting from the three data sets were combined using Rubin's Rules.[96,120] The overall estimate (odds ratio) is the average of the three computed estimates. Based on the standard errors and estimates from each imputation we calculated the within-imputation variance and the between-imputation variance to estimate a total variance. The overall standard error was derived from the total variance and was then used to calculate the overall confidence intervals. A significance test of the null hypothesis was performed using the overall estimates and total variance and resulted in p-values for each estimate.

5.4 RESULTS

5.4.1 Population Characteristics

A total of 5,270 births were recorded in the REACH catchment population between 2004 and 2010. There were 222 infant deaths during this time and the average age at death was 40.3

days. A numerical BW was recorded for 4,060 (77.0%) of the births and, after adjustment for heaping, showed a mean of 2.7kg (SD=0.4kg). 1,214 (29.9%) babies were categorized as LBW because imputed weight was less than 2.5kg at birth. 66 (1.6%) of babies were in the Very Small category, weighing less than 1.7 kg and 1,148 (283%) were in the Small category. The majority, 2,617 (64.5%) were in the Optimal BW category, weighing between 2.5 and 3.4kg. 229 (5.6%) of babies were in the Large BW category. (Table 5.2)

Data were collected from babies born in 41 villages in the catchment area. The average age of their mothers was 25.8 years old (SD=3.8 years) and 68.4% of them had only a middle school education or lower. The large majority, 92.1% were Hindus and 82.4% belonged to a caste of lower social status (Backward Caste, Scheduled Caste, Scheduled Tribe). 87.7% of participants held government ration cards and received government-subsidized food and social welfare. (Table 5.2)

Three thousand, seven hundred and sixty-six (71.5%) families reported having tap water in their homes. Just over half (52. 9%) reported closed drainage sewage disposal and the minority (27.4%) reported having a toilet facility in the house. (Table 5.2)

5.4.2 Birth-weight association with infant mortality

Using exploratory factor analysis and reliability testing, we assessed the covariates of interest (sanitation and hygiene, demographic and SES variables). While we were able to identify underlying factors in each of these three realms, none of the resulting scales was reliable enough to use.

We first confirmed the strong association between BW and IM (Table 5.3). This association held true for all the specifications of BW. In an unadjusted model we found that odds of IM increased by 4.9 times for every 1kg decrease in BW. In a model that controlled for the possible confounders maternal age, place of birth, delivery type, family type, mother's education, father's education and caste we found that the odds of IM increased to 5.0 times for every 1kg reduction in BW. The unadjusted odds of IM for babies born weighing less than 2.5kg was 2.6 times higher than for those babies weighing at least 2.5kg at birth. The categorical examination of BW demonstrated that those increased odds of IM for LBW babies are mostly

due to the increased odds of IM for babies weighing less than 1.7kg (adjusted OR=20.9, 95% CI=9.7-45.1). Meanwhile, the odds of IM for Small babies (1.7 to 2.5kg) was 2.1 times higher than for babies in the optimum range of BW (2.5-3.36kg). The association between BW and IM was significant for each BW variable and adjustment with relevant confounders increased the strength of the association in every case. (Table 5.3)

5.4.3 Demographics and SES association with IM and BW

IM was analyzed in relation to demographic and SES factors hypothesized to have significant associations. Those demographic factors that we found were not associated with IM included religion, dietary practices, the father's age and the mother's occupation (data not presented). Those SES factors that we found were not associated with IM included home ownership, land ownership, house type, welfare status (ration card), stated family income, goods ownership (car, computer, television, refrigerator, etc.), electricity in home, and literacy (data not presented). Some reproductive health variables that we tested and found to have no association included number of prenatal care visits and history of previous child death (data not presented).

Demographic variables that we found were significantly associated with IM included maternal age, place of birth, delivery type (vaginal delivery vs. caesarean section) and family type. Compared to babies born in a private hospital, odds of death for babies born at home were increased by 2.85 times (CI:1.86-4.37). Caesarean section appeared to be protective for IM (OR=0.68 CI: 0.49-0.92) (Table 5.4), though we believed this association was actually indicative of the role of access to health care and we found that the association was no longer significant when we controlled for place of birth (data not presented).

SES factors that were associated with IM included the parents' education levels, and their caste. When the father and mother had no school education the odds of IM increased by 1.94 and 1.81 times respectively when compared to secondary school educated parents. The odds of IM increased by 2.10 times for babies born to mothers in Scheduled Tribes, the caste of lowest social status. (Table 5.4).

Demographic variables that were associated with Very Small BW (<1.7kg) included sex of the baby, delivery type, and year of birth. Female babies were 1.29 (CI: 1.161-1.423) times more likely to be Very Small compared to Male babies. Year of birth was shown to be slightly protective against the odds of being born Very Small (OR=0.94 CI:0.90-0.99). Babies in this cohort were born between 2004 and 2010 and those born closer to the present had a decreased odds of Very Small BW. Caesarean section was shown to be protective against Very Small BW (OR=0.70 CI:0.60-0.82). However, as with the IM association, we reasoned that this association was indicative of access to health care and we found that when we controlled for place of birth that this association became non-significant. (Data not presented).

SES factors that were associated with Very Small BW were the parents' education levels and caste. Babies born to uneducated fathers and mothers were respectively 1.33 and 1.37 times more likely to be born less than 1.7kg. The odds of having a Very Small baby were increased by 1.77 times for members of Scheduled Tribes when compared to the caste of highest social status. (Table 5.4)

5.4.4 Sanitation/Hygiene association with IM and BW

Variables for poor sanitation/hygiene including human waste disposal, drinking water source, cooking fuel and trash disposal were associated with increased odds of IM. Risk of infant death for babies born in households using a shared community toilet was 5.42 times higher (CI: 1.77-16.59) when compared to households with their own toilet facility. Risk of infant death was increased by 2.22 times for families that used a shared community tap and increased by 20.8 times when the water source was a tanker trunk as opposed to using bottled water for drinking. Risk of IM was increased by 1.37 times in homes that used an unclean cooking fuel like dung cakes or fire wood. Risk of IM was increased for less desirable forms of garbage disposal such as field disposal (OR=1.86 CI:1.04-3.31) and dumping (OR=2.06 CI:1.17-3.62). (Table 5.5).

Poor sanitation and hygiene were also associated with Very Small BW. Sewage disposal and unclean drinking water were associated with BW less than 1.7kg. The odds of Very Small BW for babies born in houses with a soakage pit for sewage disposal were increased by 1.938

times (CI: 1.331-2.545). The odds of being Very Small BW for babies born to families whose drinking water came from a tanker truck were 9.7 times higher (CI: 1.14-46.43).

5.4.5 Adjusted associations with IM and BW

For the sanitation-IM relationship, we determined the potential demographic and SES confounders were maternal age, place of birth, delivery type, family type, mother's education, father's education and caste. We did not find any significant interaction among variables hypothesized to be related such as place of birth and delivery type and the parents' education level and caste. With both place of birth and delivery type in the model, delivery type became insignificant and was removed.

The adjusted models showed that human waste disposal, drinking water source and trash disposal maintained a significant association with IM. Risk of IM for babies born to households using a community toilet was increased by 8.18 times (CI: 2.46-27.13) Risk of IM for babies where drinking water came from a community tap or tanker truck were increased by 1.78 and 31.87 times respectively. All forms of trash disposal that were less sanitary than the reference option (composting) increased the odds of IM. (Table 5.6)

For the sanitation-BW relationship, we included the possible confounders maternal age, sex of the child, place of birth, delivery type, year of birth, mother's education, father's education and caste in the model. Again, we tested for possible interaction among the confounders and found none. We did find that, with the inclusion of both place of birth and delivery type that delivery type became insignificant and that variable was removed from the model.

The adjusted models showed that human waste disposal, sewage disposal and drinking water source were associated with an increased odds of Very Small BW. Using a community toilet increased the odds of Very Small BW by 2.96 times (CI:0.86-10.25), though this association was not statistically significant (p=0.086). Open drainage sewage disposal increased the odds of Very Small BW by 1.312 times (CI:1.02-1.68) and using a tanker truck for drinking water increased the odds by 12.86 times (CI:1.12-148.3). (Table 5.8)

5.4.6 Sanitation-IM associations and the effect of BW

After adding BW to the model, we found that the significant associations between poor sanitation and odds of IM held true except for one case. The association between water from a tanker truck and IM was no longer significant when we included BW in the model. However, the associations between human waste disposal in a community toilet, drinking water from a community tap and less sanitary forms of garbage disposal all remained significant. (Table 5.7)

5.4.7 Model Precision

Several of the associations seen in our final predictive models had very wide confidence intervals, especially the association of human waste disposal in a community toilet with IM, with and without BW included in the model. (Table 5.6, 5.7, 5.8). These wide confidence intervals indicate a lack of precision that is due to the high number of missing values in this data set as well as the very small numbers of people who are practicing the behavior like using a community toilet (n=21), using open drainage for sewage disposal (n=106) and burning garbage (n=54). (Table 5.5)

5.5 DISCUSSION

These results speak to several important phenomena in this rural Indian population. First, these results confirm the strong association between BW and IM. The highest odds of IM were found in the smallest BW category; the IM rate for the Very Small BW category was 318 per 1,000 live births. These results confirm that BW is strongly associated with IM, though they do not speak to the question of BW causing IM.

Second, these results clearly identify factors that are significantly associated with both IM and small BW: poor sanitation and hygiene. We believe that, in this population, sanitation and hygiene measures are surrogates for exposure to infectious disease. A shared community toilet is not the direct cause of IM, rather it is the increased opportunity for fecal-oral transfer of pathogens that comes from a community toilet that is contributing to IM. Dirty cooking fuel like dung cakes are surrogates for exposure to pathogens that cause upper respiratory infections, a leading cause of infant death.[51,52] The significant associations of these measures of

sanitation with IM, independent of BW, suggest that it is these infectious insults after the baby is born that are contributing in large part to IM in this rural Indian population.

Third, these results address the belief held by many researchers[121-125] that increased risk of IM is primarily the result of socioeconomic status and that, as wealth increases, so too will IM decrease. These results suggest that certain measures of SES are significantly associated with IM. However, if SES were the primary driving factor behind IM, then the strength of the associations with sanitation/hygiene factors would decrease, if not disappear completely, when SES is added to the models as a confounder. We have shown here that the association between sanitation/hygiene remains significant, and in fact becomes stronger, after adding SES variables to the models. Therefore, while SES may play a role in IM and even be associated with poor sanitation and hygiene in its own right, these results show that low SES does not fully explain the high rates of IM in this population.

Finally, these results challenge the contention that small BW causes IM. If BW is not on the causal pathway to IM, its relationship with IM would have to be explained by factors that both decrease BW and increase mortality. These sanitation measures are significantly associated with both a decrease in BW and an increase in IM. Human waste disposal, sewage disposal and drinking water were all significantly associated with increased odds of having a very small baby (<1.7kg). These associations shed light on whether BW falls on the causal pathway to IM. Our results suggest BW is not on the causal pathway to IM and that instead, both BW and IM are affected by a similar set of causal factors. (Figure 1) This has been previously hypothesized by Wilcox and others.[11,42,47,48,126] Sanitation and hygiene, or infectious exposures that directly affect the risk of IM are probably also affecting maternal health and pregnancy. Diarrheal diseases and the resulting anemia and malnutrition and other types of infections during pregnancy have been shown to have ill effects on pregnancy outcomes, including BW.[69,78-81] Our data support that notion that sanitation and hygiene are strongly associated with both lower BW and increased IM, though the associations are not exactly the same. These results lend themselves more to the second interpretation of the BW-IM relationship shown in Figure 5.1. Namely, while both BW and IM are affected by the same

causal factors, BW does not relate directly to IM. The strong association between BW and IM is then explained by the strong associations of BW and IM to the same causal factors.

The major weakness of this analysis is the missing data. Unfortunately, those variables taken from the Medchal Family Health Survey had a very high percentage of missing data because for many of the families in the MFHS we were not able to link them back to the REACH database. The missing data limited the types of analysis we were able to perform with this data set. For example, we were unable to test a regression model that included all the sanitation/hygiene variables as predictors of IM and BW because the cases were spread too thin amidst the many variables and missing data. (Table 5.9) However, for all that there are missing data, this data set is unique in that it represents a successful longitudinal data collection in rural India. And, while there are missing data from the MFHS, the REACH data are very complete and the analysis of these datasets shows interesting and statistically significant associations with IM and BW in this population.

These findings have important implications for public health and health policy. The current government approach to reducing IM in India involves social welfare and feeding programs for mothers with the explicit goal of increasing BW and thereby reducing IM. These results support the notion that a focus on increasing BW is not appropriate and that resources should be devoted to exposures to babies during infancy, such as diarrheal and respiratory diseases. These results also contradict the idea that IM is simply a result of socio-economic status. Our findings show that IM is linked to sanitary conditions independent of their association with SES factors.

Finally, these findings provide an opportunity for tailored, appropriate interventions to combat the determinants of IM in this rural Indian population. Because of our low numbers and missing data for some variables, these data should be confirmed to determine the scope of the public health impact of potential interventions. However, improving sanitation and hygiene and reducing exposure to infectious disease is a feasible goal for an intervention at the population level and, based on this study, would help reduce the rates of IM in this population.

5.6 TABLES AND FIGURES

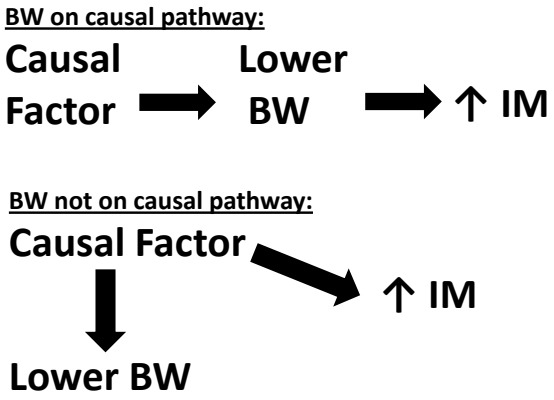

Figure 5.1: Directed Acyclic Graphs representing possible constructions of the BW-IM relationship

Table 5.1: Measures collected through REACH and the Medchal Family Health Survey

REACH Database			
Household	Village	Family members	Media sources
Demographics	Religion	Family type	Occupation
	Dietary practice	House type	Possessions
	Age	Electricity	
Reproduction	Births	Date of Birth of baby	Place of birth
	Sex of child	Gestation Length	Delivery Type
	Age of child now	Infant Death	# of Prenatal visits
	Year of birth	Child Death	Sterilization
	Birth-weight	Age at Death	
Socioeconomic	Caste	Family Income	Education level
Sanitation/	Cooking place	Trash Disposal	Water source
Hygiene	Human waste disposal	Sewage Disposal	
Medchal Family Health Survey			
Household	Religion	Social services used	Income
Demographics	Caste	Possessions in house	Welfare recipient
	# Members	Type of house	Animals owned
	Ages	Number of rooms	Cosangiuinous
	Relationships	Land ownership	marriage
	Marital Status	Home ownership	Occupation
	Education level	Nutrition	
Health	Asthma	Malaria	Tobacco
	Goiter	Tuberculosis	Health facility used
	Thyroid	Jaundice	Vaccination records
	Diabetes	Alcohol	Current illness
Sanitation /	Drinking water	Water treatment	Cooking fuel
Hygiene	Cooking water	Cooking facility	Hand washing
	Toilet facility	Chimney	
Reproductive	Births history	Prenatal vitamins	Family planning
health	Birth-weight	Current pregnancy	Pre-natal care
	Child/infant deaths	Contraception method	Pregnancy
	Place of delivery	Breastfeeding	complications
	Complications	Child feeding	Vaccinations
	Health care for baby	Child illnesses	HIV knowledge
	Previous child death	Repro health care	

Table 5.2: Population Characteristics

Participant Characteristics	n	(%)
Births recorded	5,270	
Infant Deaths	222	(4.21)
Days old at death (mean, SD)	40.30	(80.77)
BW recorded	4,060	(77.04)
Mean, SD of BW distribution	2.71	(0.42)
Low BW (<2.5 kg)	1214	(29.89)
BW in four categories		
Very small BW (1.7 kg)	66	(1.62)
Small BW [1.7-2.5kg)	1148	(28.27)
Optimal BW [2.5-3.36 kg)	2617	(64.47)
Large BW (>3.36 kg)	229	(5.64)
Demographic and Socioeconomic		
Number of villages	41	
Mother's age at time of birth (mean, SD)	25.76	(3.81)
Father's age at time of birth (mean, SD)	31.37	(4.99)
Mother's Education, less than high school	3194	(68.39)
Father's Education, less than high school	2,577	(51.02)
Religion		
Hindu	4852	(92.07)
Muslim	297	(5.64)
Christian	119	(2.26)
Other	2	(0.04)
Caste		
Forward Caste	479	(9.26)
Backward Caste	973	(18.82)
Scheduled Caste	2708	(52.38)
Scheduled Tribe	579	(11.20)
Other	43	(0.83)
Don't know	388	(7.50)
Ration card holders (social welfare)	1926	(87.70)
Sanitation and Hygiene		
Water, tap water in home	3,766	(71.49)
Sewage disposal, closed drainage	2,786	(52.89)
Human waste disposal, household facilities	1,441	(27.35)

All data presented as No. (%) unless otherwise indicated

All statistics based on models that included BW as an outcome or predictor variable were calculated for each of three imputed data sets and the results combined using Rubin's Rules to produce an estimated combined OR, confidence interval and p-value

Table 5.3: Birth-weight association with Infant Mortality

Birth-weight – infant mortality association								
	n	**n of IM**	**OR**	**95% CI**	**p**	**Adj. OR**	**Adj. 95% CI**	**Adj. p**
BW, continuous (1kg increase)	4060	142	4.91	3.35-7.20	<0.0001	4.99	3.24-7.72	<0.0001
LBW (<2.5kg vs. ≥2.5kg)	1214	73	2.64	1.69-4.13	0.0003	2.67	1.63-4.38	0.0006
BW, categorical								
Very small (<1.7kg)	66	21	19.16	9.39-39.12	<0.0001	20.89	9.68-45.10	<0.0001
Small [1.7kg, 2.5kg)	1148	53	1.99	1.23-3.22	0.0073	2.07	1.24-3.47	0.0075
Optimal [2.5kg, 3.36)	2617	62	*Reference*					
Large (≥3.36)	229	6	1.09	0.40-2.92	0.8646	1.28	0.41-4.07	0.6610

IM = Infant Mortality BW= Birth-weight OR = Odds Ratio Adj. = Adjusted

For Adjusted OR, all variables adjusted for maternal age, place of birth, delivery type, family type, mother's education, father's education and caste

All statistics based on models that included BW as an outcome or predictor variable were calculated for each of three imputed data sets and the results combined using Rubin's Rules to produce an estimated combined OR, confidence interval and p-value

Table 5.4: Demographic/Socioeconomic variables' association with Infant Mortality (IM) and Very Small (VS) birth-weight babies (less than 1.7kg)

			INFANT MORTALITY			BIRTH-WEIGHT		
	n	n of M	n of IM	OR (CI) or χ^2 value of Infant Mortality	p-value	n of VS	OR (CI) or χ^2 value of VS	p-value
DEMOGRAPHIC VARIABLES								
Maternal age	5270	0	222	1.04 (1.01-1.08)	0.0150	65	0.987 (0.971-1.004)	0.1252
Sex of Child		0						
Male	2665		122	*Reference*		26	*Reference*	
Female	2605		100	1.20(0.92-1.57)	0.1823	39	1.292 (1.136-1.471)	0.0001
Place of birth		0						
Private Hospital	2927		118	*Reference*		32	*Reference*	
Government Hospital	579		25	1.07(0.69-1.67)	0.7503	5	1.254 (0.739-2.128)	0.3208
Rural Health Center	1493		50	0.83(0.59-1.16)	0.2624	26	1.161 (0.550-2.451)	0.5694
Home	271		29	2.85(1.86-4.37)	<.0001	2	1.306 (0.575-2.969)	0.3767
Delivery Type		0						
Vaginal delivery	3617		169	*Reference*		46	*Reference*	
Ceasarean Section	1653		53	0.68(0.49-0.92)	0.0145	19	0.701 (0.603-0.815)	<.0001
Year of Birth		0		0.96(0.88-1.03)	0.2459		0.941 (0.896-0.988)	0.0160
Family type		0						
Nuclear Family	771		21	*Reference*		9	*Reference*	
Joined Hindu Family	4499		201	1.67 (1.06-2.64)	<.0001	56	1.084 (0.845-1.390)	0.5088
SOCIOECONOMIC FACTORS								
Father's education		219						
Post-graduate	11		2	6.80(1.44-32.02)	0.0154	0	1.418 (0.255-7.869)	0.6747
Graduate	189		11	1.89(0.98-3.63)	0.0559	2	0.761 (0.500-1.157)	0.1955
High School	2274		72	*Reference*		23	*Reference*	
Middle School	814		30	1.17(0.759-1.81)	0.4773	14	1.111 (0.893-1.382)	0.3364
Primary School	859		40	1.49(1.01-2.22)	0.0463	13	1.185 (0.901-1.559)	0.1970
Uneducated	904		54	1.94(1.35-2.79)	0.0003	9	1.333 (1.091-1.630)	0.0053
Mother's education		600						
Graduate	93		4	1.21(0.43-3.42)	0.7232	2	0.726 (0.439-1.199)	0.2105
High School	1365		49	*Reference*		15	*Reference*	
Middle School	1777		79	1.25(0.87-1.80)	0.2298	22	1.214 (1.017-1.450)	0.0323
Primary School	594		23	1.08(0.65-1.79)	0.7602	10	1.351 (1.070-1.706)	0.0116
Uneducated	823		52	1.81(1.21-2.70)	0.0036	6	1.366 (1.099-1.697)	0.0049
Caste		488						
Forward Caste	579		21	*Reference*		5	*Reference*	
Backward Caste	2708		94	0.96(0.59-1.55)	0.8531	39	1.573 (1.191-2.077)	0.0028
Scheduled Caste	973		53	1.53(0.91-2.56)	0.1060	9	1.469 (0.979-2.203)	0.0604
Scheduled Tribe	479		35	2.10(1.20-3.65)	0.0090	7	1.769 (1.210-2.587)	0.0055

M=Missing Data IM= Infant Mortality VS= Very small birth-weight category (<1.7kg) OR= Odds Ratio CI= Confidence Interval χ^2 = Chi-squared
Categories with case numbers too small to estimate an effect were removed (Mother's Education: Post Graduate)
Dichotomous categorical independent variables were cross-tabulated with the categorical outcome variables using Pearson's χ^2 test or the Fisher exact test. Categorical variables with >2 categories were assessed using logistic regression analysis. Continuous dependent variables were assessed using logistic regression analysis.
All statistics based on models that included birth-weight as an outcome or predictor variable were calculated for each of three imputed data sets and the results combined using Rubin's Rules to produce an estimated combined OR, confidence interval and p-value

Table 5.5: Sanitation/Hygiene variables' association with Infant Mortality (IM) and Very Small (VS) birth-weight babies (less than 1.7kg)

			INFANT MORTALITY			BIRTH-WEIGHT		
	n	n of M	n of IM	OR (CI) or χ^2 value of IM	p-value	n of VS	OR (CI) or χ^2 value of VS	p-value
Human waste disposal		4						
Household toilet	1441		60	*Reference*		14	*Reference*	
No facility, open field	3802		158	1.00 (0.74-1.35)	0.990	51	1.134 (0.979-1.312)	0.0934
Community toilet	21		4	5.42 (1.77-16.59)	0.003	0	1.794 (0.547-5.882)	0.3252
Sewage Disposal		4						
Closed Drainage	2786		119	*Reference*		40	*Reference*	
Kitchen Garden	180		11	1.46(0.77-2.76)	0.245	2	1.065 (0.725-1.566)	0.7425
Open Drainage	2194		91	0.97(0.73-1.28)	0.829	22	1.196 (0.973-1.471)	0.0813
Soakage Pit	106		1	0.21(0.03-1.54)	0.126	1	1.934 (1.221-3.062)	0.0055
*Drinking water		3078						
Bottled water	747		26	*Reference*		7	*Reference*	
Piped into home	421		19	1.31 (0.72-2.40)	0.3800	4	1.063 (0.739-1.530)	0.7274
Piped into yard	428		17	1.15 (0.62-2.14)	0.6662	1	1.121 (0.819-1.534)	0.4708
Shared community tap	459		34	2.22 (1.31-3.75)	0.0029	7	1.175 (0.788-1.752)	0.3947
Tube well	121		5	1.20 (0.45-3.18)	0.7204	4	1.173 (0.661-2.082)	0.5739
*Cooking Fuel		3086						
"Clean"** Fuel	1134		60	*Reference*		15	*Reference*	
"Unclean" Fuel	1050		44	1.37 (1.01-1.86)	0.042	10	0.970 (0.250-3.763)	0.9314
Trash Disposal		2						
Composting	651		16	*Reference*		10	*Reference*	
Field Disposal	1050		47	1.86 (1.04-3.31)	0.0347	17	0.918 (0.680-1.240)	0.5546
Dumping	1115		55	2.06 (1.17-3.62)	0.0123	10	0.846 (0.656-1.090)	0.1924
Burning	54		3	2.34(0.66-8.28)	0.1892	1	0.884 (0.395-1.978)	0.7495
Municipal trash collect.	2398		101	1.74 (1.02-2.98)	0.0412	27	0.862 (0.650-1.142)	0.2726

M=Missing Data IM= Infant Mortality VS= Very small birth-weight category (<1.7kg) OR= Odds Ratio CI= Confidence Interval χ^2 = Chi-squared

* indicates the variable is taken from the MFHS, not the REACH database. % missing data is much greater in MFHS.

Categories with case numbers too small to estimate an effect were removed (Drinking water: tanker truck, protected spring, protected dug well)

Dichotomous categorical independent variables were cross-tabulated with the categorical outcome variables using Pearson's χ^2 test or the Fisher exact test. Categorical variables with >2 categories were assessed using logistic regression analysis. Continuous dependent variables were assessed using logistic regression analysis.

All statistics based on models that included birth-weight as an outcome or predictor variable were calculated for each of three imputed data sets and the results combined using Rubin's Rules to produce an estimated combined OR, confidence interval and p-value

**Clean fuel sources include kerosene, charcoal, coal/Lignite, biogas, natural gas, electricity. Unclean fuel sources include dung cakes, fire wood, straw/shrubs/grass, agricultural crop waste.

Table 5.6: Sanitation/Hygiene variables' association with Infant Mortality while adjusting for potential confounders (separate models for each predictor)

Model	Adjusted OR	95% CI	p
1. Human waste disposal, community toilet	8.18	2.46-27.13	0.0006
2. Drinking water, community tap	1.780	1.02-3.17	0.0482
3. Trash Disposal, dumping	2.109	1.16-3.85	0.0151
4. Trash Disposal, field disposal	2.452	1.30-4.61	0.0054
5. Trash Disposal, municipal trash collect.	2.306	1.30-4.09	0.0043

M=Missing Data IM= Infant Mortality OR= Odds Ratio CI= Confidence Interval
All variables adjusted for maternal age, place of birth, family type, mother's education, father's education and caste.
All categorical variables compared against reference variables listed in tables 5.4 and 5.5

Table 5.7: Sanitation/Hygiene variables association with Infant mortality, while adjusting for potential confounders, including birth-weight (separate models for each predictor)

Model	Adjusted OR	95% CI	p
1. Human waste disposal, community toilet	10.631	2.488-45.423	0.0014
2. Drinking water, community tap	1.935	1.245-3.008	0.0036
3. Trash Disposal, dumping	2.605	1.229-5.521	0.0125
4. Trash Disposal, field disposal	2.277	1.110-4.673	0.0248
5. Trash Disposal, municipal trash collect.	2.041	0.912-4.569	0.0827

OR= Odds Ratio CI= Confidence Interval
All variables adjusted for maternal age, place of birth, family type, mother's education, father's education, caste and BW.
All categorical variables compared against reference variables listed in tables 5.4 and 5.5
BW included as a continuous variable, adjusted for functional form of BW
All statistics based on models that included BW as an outcome or predictor variable were calculated for each of three imputed data sets and the results combined using Rubin's Rules to produce an estimated combined OR, confidence interval and p-value

Table 5.8: Sanitation/Hygiene variables association with Very Small Birth-weight, while adjusting for potential confounders (separate models for each predictor)

Model	Adjusted OR	95% CI	p
1. Human waste disposal, community toilet	2.964	0.857-10.255	0.0858
2. Sewage disposal, open drainage	1.312	1.025-1.679	0.0315

OR= Odds Ratio CI= Confidence Interval

All variables adjusted for sex of the child, place of birth, delivery type, year of birth, mother's education, father's education and caste

All categorical variables compared against reference variables listed in tables 5.4 and 5.5

All statistics based on models that included BW as an outcome or predictor variable were calculated for each of three imputed data sets and the results combined using Rubin's Rules to produce an estimated combined OR, confidence interval and p-value.

Table 5.9: Overlap of significant IM determinants when compared against one another using cross-tabulation

		Human Waste Disposal			
		Household Toilet		Community Toilet	
Source of Drinking Water	Bottled water	Composting: 15	Dumping: 40	Composting: 0	Dumping: 2
		Field: 89	Municipal: 100	Field: 1	Municipal: 2
	Community Tap	Composting: 8	Dumping: 16	Composting: 0	Dumping: 0
		Field: 46	Municipal: 47	Field: 0	Municipal: 1
		Trash Disposal			

6.0 CONCLUSION / PUBLIC HEALTH IMPACT

While BW data has been extensively studied over the years, it remains one of the most powerful and interesting variables for providing insight and information about the health and well-being of babies as they are born and grow. While there is controversy about this variables and its meaning, particularly in relation to IM, it is still worthwhile to continue to try to discover what BW means and predicts for the newborn.

From our perspective, with a focus on maternal and child health in rural India, there are two major holes in the literature and knowledge related to BW. The first is that there are no published BW data from India that are representative of the population and accurately measured. While India has one of the highest reported rates of LBW,[127] representative data are not available to study or address this phenomenon. Developing a method to correct for two of the most common biases found in survey-collected BW data is an important step towards being able to truly study this indicator of child health and development.

It is also important to note that this method is generalizable to other populations in several ways. Ideally, researchers would collect a subset of gold standard data that can be used as the calibration set to develop a calibration function specific to the population of interest, as we have done here. However, for cases where collecting a calibration set is impossible, our method could be used directly to adjust BW data from other developing country populations. Our method can help make survey BW data from the developing world accessible for study when it has not been useable before.

Secondly, while plenty of analyses have been completed in developing countries regarding the nature of BW distributions and characteristics of the relationship between BW and IM, for the reasons listed above, very few of these studies includes BW data collected in the developing world and we have not been able to find any published accounts of data from India. Therefore, these characteristics, while accepted as fact, have not actually been verified in a large, representative sample taken from a developing country. As we have shown, there are

differences between the US and India in terms of the relationship between BW and IM and these differences could have policy implications. It appears that BW does not contribute to IM to the same degree in India as it does in the US population and the relationship we demonstrate contradicts some of the current focus to combat IM in India by increasing BW through programs to feed mothers. The results of our comparison between the US and India beg the question of what is really causing IM in India and what can be done to effectively reduce infant death in this population.

Through studying a small population in rural South India, we have highlighted some of the determinants of IM that are independent of BW. We have shown that sanitation and hygiene, which we believe to be surrogates of infectious diseases, are significantly associated with IM and very small BW in this population. These results provide an opportunity for designing interventions to combat IM that would be specific to the community and feasible.

Altogether, these papers provide new methods, insight and conclusions regarding BW and IM in India. The deaths of babies in their first year of life is a tragedy that afflicts South Asia and India more than any other part of the world. In studying these topics we have had the great opportunity to gain new understanding and knowledge that will hopefully aid in reducing the tragedy of IM in India in the future.

APPENDIX A: DETAILS OF DEVELOPMENT OF MODIFIED STATISTICAL CALIBRATION AND MULTIPLE IMPUTATION METHOD

At the start of this analysis, we did not want to assume, a priori that the distribution of error in the calibration set was either the same across the range of BWs or normally distributed (as is the assumption in standard calibration methods). To discover the nature of the error function, we first divided the full range of BWs into "bands," whereby weights within 250 grams of one another are grouped in the same band.

The 250g range was selected for convenience because it is relatively narrow, yet large enough that the majority of the resulting bands contain enough samples to find a best fit distribution. For each band that contained at least 10 samples, a best fit distribution was estimated. For each band, the best-fit software estimated that the best-fit distribution follows the Student's T distribution with adjustment for centering. The band distributions parameters and confidence intervals are shown in Table 7.1.

Figures 7.2 and 7.3 show a graphical representation of the range of the scale (σ) and shape (v) parameters, respectively. These figures support our contention that, within the confidence intervals, these parameters are equal across all bands.

From this point, one approach would be to define the error function to be a Student's T distribution whose parameters are a function of the analog weight. Thus, one would first estimate which band a particular analog record falls within, and then that band's corresponding distribution parameters would be applied to the error function for that record.

However, as seen in Table 7.1 and Figures 7.2 and 7.3, the scale and shape parameters of the best-fit distributions for the different bands are actually the same within the confidence

intervals. Therefore, the scale and shape of the distribution remain constant across the bands. We can make use of this consistency to improve the model parameter estimation.

The implication of the banded analysis described above is that the distribution of digital weights is uniform for all analog values. This characteristic is beneficial because it means that to adjust the analog BW data, we will not have to rely on separate error distributions from the bands. Instead, a single error distribution can be derived which applies to all analog weights.

This has several advantages. First, it allows us to define probability distributions to adjust data that lies in the bands with too few records to estimate a best-fit distribution. Next, it establishes a separate probability distribution for each unique analog weight, and thus avoids the artificial clustering introduced by the bands. Furthermore, we can fit a single distribution to all 759 records, with the advantage of tighter confidence intervals on the distribution parameters.

APPENDIX A TABLES AND FIGURES

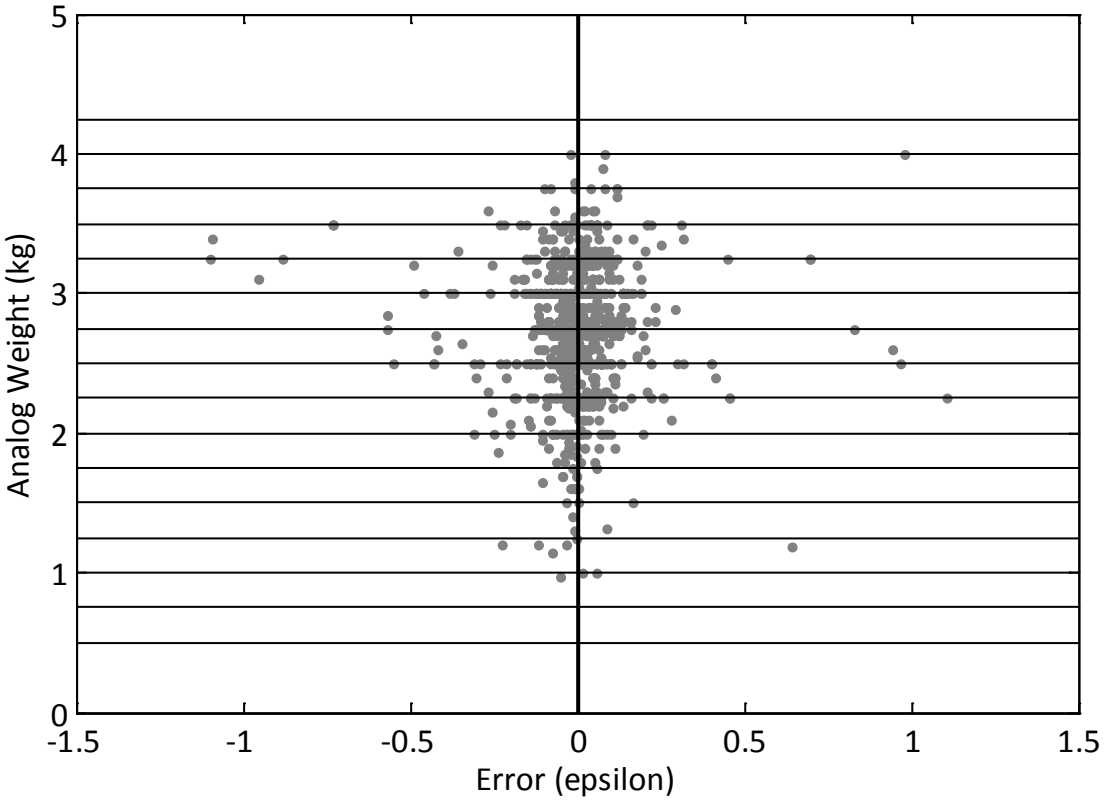

Figure 7.1: Analog weight versus Error (ε_i') with "bands" (horizontal lines = "bands")

Table 7.1: Parameters for the distribution within each band

Band Bounds (kg)	n samples in Band	Location Parameter (μ_bar)	μ_bar CI		Scale Parameter (σ_bar)	σ_bar CI		Shape Parameter (ν_bar)	ν_bar CI	
			min	max		min	max		min	max
(1.50 - 1.75]	13	0.007	0.002	-0.0196	-0.0249	1.000	0.036	1.000	0.010	1.000
(1.75 – 2.00]	20	0.045	0.022	-0.0202	-0.0425	5.000	0.089	5.000	0.000	23.00
(2.00 - 2.25]	42	0.060	0.034	0.0104	-0.0140	1.514	0.106	1.514	0.314	2.713
(2.25 - 2.50]	66	0.055	0.041	0.0113	-0.0059	2.181	0.073	2.181	0.929	3.433
(2.50 - 2.75]	147	0.052	0.042	0.0153	0.0045	2.227	0.065	2.227	1.276	3.178
(2.75 – 3.00]	113	0.063	0.052	0.0097	-0.0048	3.131	0.078	3.131	1.471	4.791
(3.00 - 3.25]	111	0.064	0.051	-0.0009	-0.0160	2.945	0.080	2.945	1.334	4.557
(3.25 - 3.50]	53	0.076	0.054	0.0189	-0.0084	2.304	0.107	2.304	0.710	3.898
(3.50 - 3.75]	16	0.037	0.017	0.0318	0.0070	1.339	0.081	1.339	0.010	2.716

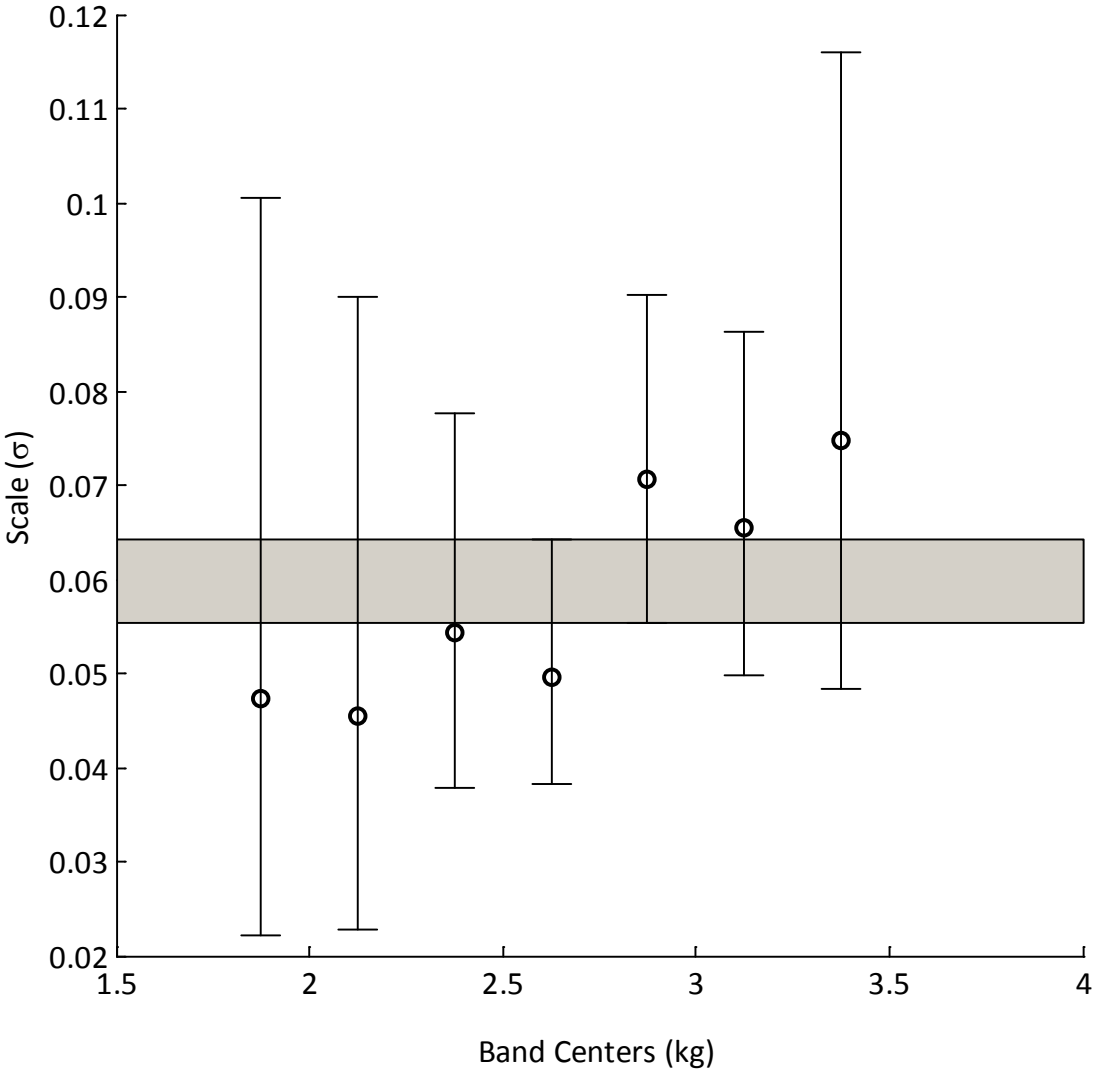

Figure 7.2: Band center versus Scale Parameter (σ)

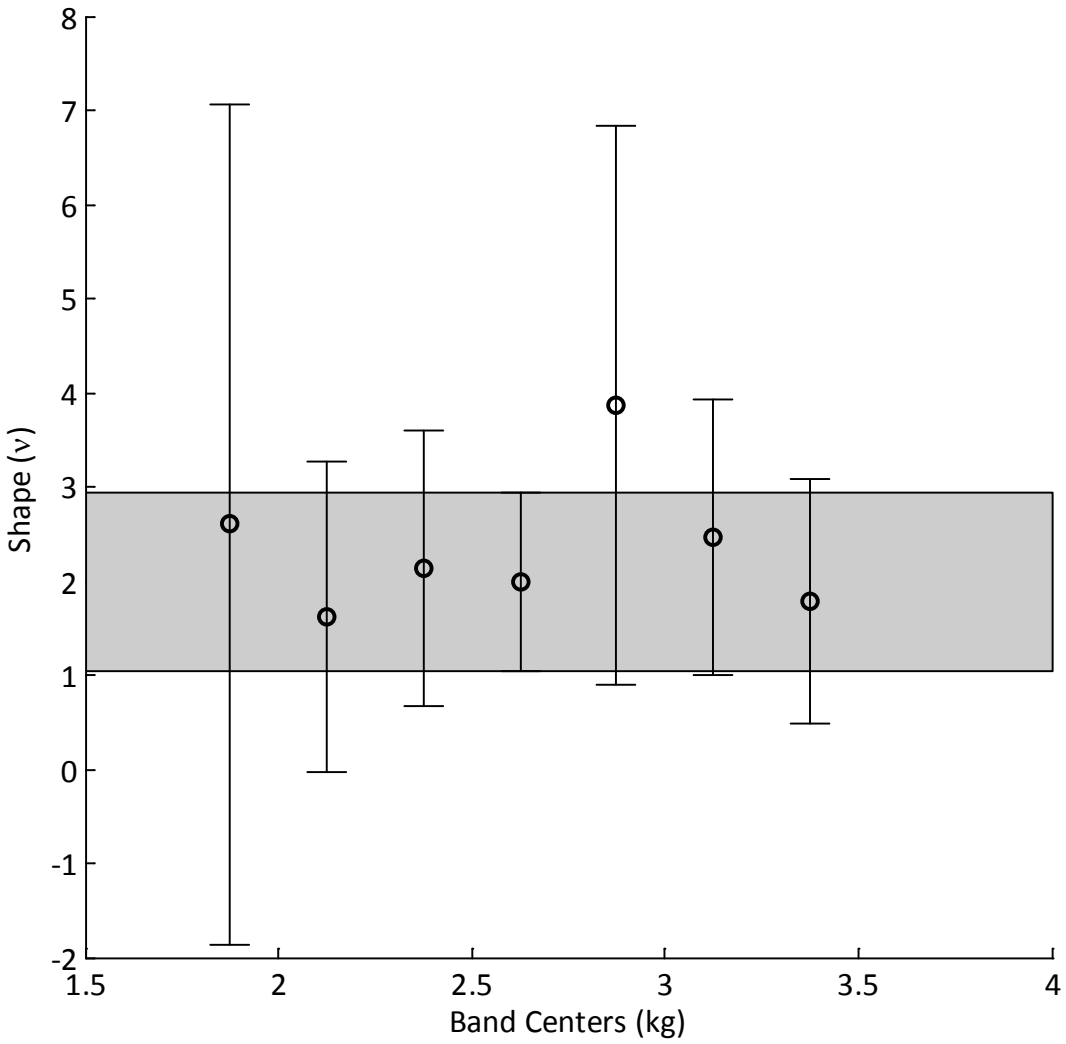

Figure 7.3: Band centers versus Shape Parameter (v)

APPENDIX B: ANALYSIS CODE

Manuscript 1: Modified Statistical Calibration and Multiple Imputation

MATLAB 7.11 R2010B

```
function imputed = generateimputed(analogBW,pd,muBeta,muDig,sigmaDig,nImputedSets)
% GENERATEIMPUTED - generates an imputed data set
% imputed = generateimputed(analagBW,pd,muBeta,muDig,sigmaDig,nImputedSets)
%
% Inputs:
% analogBW: vector of analog birth weights to be corrected
% pd: t-location scale probability distribution object describing the
%     distribution of the recentered calibration set
% muBeta: 2-by-1 vector of linear regression parameters, beta, defining
%     analog weight as a fn of digital weight. (muBeta(1) is y-intercept,
%     muBeta(2) is slope)
% muDig: scalar value of mean digital weight of calibration set
% sigmaDig: scalar value of stdev of digital weight in calibration set
% nImputedSets: scaler, number of imputed data sets to generate
%
% Outputs:
% imputed: matrix of imputed data sets.  each column is an imputed set.


% get number of birthweight samples
nSamples = length(analogBW);

% allocate memory for imputed data
imputed = zeros(nSamples,nImputedSets);

% find set of unique analog weights
[uniqueAnalogWeight, im, in] = unique(analogBW);

% define min and max allowable values for BW
maxVal = muDig+10*sigmaDig;
minVal = 0;


h = waitbar(0);
nSamplesEvaluated = 0;
% loop through all unique analog BW values
for j = 1:length(uniqueAnalogWeight)
```

```
   % construct distribution object for jth analog weight
   xcenter = (uniqueAnalogWeight(j)-muBeta(1))/muBeta(2);
   pd_iter = ProbDistUnivParam('tlocationscale', [pd.mu+xcenter,pd.sigma,pd.nu]);

   % impute values for all records with jth analog weight
   iVal = find(analogBW == uniqueAnalogWeight(j));
   y = rand(length(iVal),nImputedSets);
   imputed(iVal,:) = icdf(pd_iter,y);

   % re-impute those values that are too small
   while 1
      iTooSmall = find(imputed(iVal,:)<minVal);
      if isempty(iTooSmall)
         break
      end
      [a,b]=ind2sub(size(imputed(iVal,:)),iTooSmall);
      y = rand(length(a),length(b));
      imputed(iVal(a),b) = icdf(pd_iter,y);
   end
   % re-impute those values that are too large
   while 1
      iTooLarge = find(imputed(iVal,:)>maxVal);
      if isempty(iTooLarge)
         break
      end
      [a,b]=ind2sub(size(imputed(iVal,:)),iTooLarge);
      y = rand(length(a),length(b));
      imputed(iVal(a),b) = icdf(pd_iter,y);
   end

   % update waitbar
   nSamplesEvaluated = nSamplesEvaluated + length(iVal);
   waitbar(nSamplesEvaluated/nSamples,h);
end
close(h)
```

Manuscript 2: Plotting birth-weight frequency distributions and weight specific infant mortality curves

MATLAB 7.11 R2010B

```
function [xMin,yMin,y2500,yAtPDFmax,frequencyAtxMin] =
stackedplot(weightbins,birthfreq,mortality,bw)

% fit normal curve to bw data
[muo,sigma]=normfit(bw);
Y = normpdf(weightbins,muo,sigma);

% fit 3rd-order polynomial to mortality curve
logy = log(mortality);
I = find(logy>0);
[p,S,mu] = polyfit(weightbins(I),logy(I),3);
xFit = linspace(min(weightbins),max(weightbins),1001);
yFit = polyval(p,xFit,S,mu);

figure;
% plot mortality curve
h(1) = subplot(2,1,1);
semilogy(weightbins,mortality,'.','linewidth',2,'markersize',20,'color',[.5 0 0])
set(gca,'fontsize',12);
set(h(1),'xTickLabel','');

% underlay polyfit to mortality curve
hold on;
semilogy(xFit,exp(yFit),'linewidth',2);
hh = get(gca,'Children');
set(gca,'Children',hh(2:-1:1));


% plot birthweight hist
h(2) = subplot(2,1,2);
hb = bar(weightbins,birthfreq/1000);
set(gca,'fontsize',12);
set(hb,'FaceColor',[.5 0 0])
xlabel('Birthweight (g)');

% calculate enclosed area of histogram
xd = get(hb,'Xdata');
nbins = length(weightbins);
```

```
n = sum(birthfreq);
rangex = max(xd(:)) - min(xd(:));
binwidth = rangex/nbins;
area = n * binwidth;

% overlay pdf on histogram
hold on
plot(weightbins,area*Y/1000,'linewidth',2)

% link axes
linkaxes(h,'x')

% bring axes in contact with eachother
p1 = get(h(1),'position');
p2 = get(h(2),'position');
 top = p1(4)+p1(2);
bottom = p2(2);
height = (top-bottom)/2;

p2(4) = height;
p1(4) = height;

set(h(2),'position',p2);
set(h(1),'position',[p1(1) p2(2)+p2(4) p1(3:4)])


yTicLabel = get(h(2),'ytickLabel');
yTicNum = double(yTicLabel);
[nRows,nCols] = size(yTicNum);
for j=1:nCols
   yTicNum(nRows,j) = 32;
end
yTicLabel = char(yTicNum);
set(h(2),'ytickLabel',yTicLabel);

% adjust yaxis of mortality curve
set(gcf,'CurrentAxes',h(1));
ca = axis;
iOver1 = find(mortality>=1);
mortalityMin = min(mortality(iOver1));
pwr = 0;
while 10^pwr<mortalityMin
   pwr = pwr+.5;
end
```

```
pwr = pwr-.5;
axis([ca(1:2) 10^pwr ca(4)]);

% calculate parameters
y2500 = exp(polyval(p,2500,S,mu));
[yMin,iMin] = min(yFit);
yMin = exp(yMin);

xMin = xFit(iMin);


yAtPDFmax = exp(polyval(p,muo,S,mu));
frequencyAtxMin = normpdf(xMin,muo,sigma)*area/1000;


% add min line
set(gcf,'CurrentAxes',h(1));
ca = axis;
plot([xMin xMin],[ca(3) yMin],'--k');
plot([ca(1) xMin],[yMin yMin],'--k');

set(gcf,'CurrentAxes',h(2));
ca = axis;
plot([xMin xMin],ca(3:4),'--k');
%plot([ca(1) xMin],[frequencyAtxMin frequencyAtxMin],'--k');

% read birthweight intervals, frequencies, and mortality rates from excel
% file
data = xlsread('NFHS Birthweight and IM freq per 100g.xlsx','Sheet1','A2:D60');
wi = data(:,1);
mortality = data(:,3);
birthfreq = data(:,4);

% read raw birthweight data
bw = xlsread('NFHS Birthweight Data w intervals of 100g.xls','NFHS Birthweight Data w
interva','E2:E20947');

% throw out all records greater than 7000 grams
I = find(bw>7000);
bw(I) = [];


% generate stacked plots, and return values of interest
[xMin,yMin,y2500,yAtPDFmax,frequencyAtxMin] = stackedplot(wi,birthfreq,mortality,bw);
```

```
%% load data
data = xlsread('BW and IM freq per 100g.xlsx','Sheet1','A2:D60');
wi = data(:,1);
birthfreq = data(:,4);
mortality = data(:,3);

bw = load('nat2004us.txt');
bw = bw(:,1);
I = find(bw>7000);
bw(I) = [];

%% plot
figure(1);clf;

h(1) = subplot(2,1,1);
semilogy(wi,mortality,'.','linewidth',2,'markersize',20,'color',[.5 0 0])
set(gca,'fontsize',12);
set(h(1),'xTickLabel','');


h(2) = subplot(2,1,2);
hb = bar(wi,birthfreq/1000);
set(gca,'fontsize',12);
set(hb,'FaceColor',[.5 0 0])
xlabel('Birthweight (g)');
[mu,sigma]=normfit(bw);
Y = normpdf(wi,mu,sigma);
muo = mu;

xd = get(hb,'Xdata');
nbins = length(wi);
n = sum(birthfreq);
rangex = max(xd(:)) - min(xd(:));
binwidth = rangex/nbins;
area = n * binwidth;

hold on
plot(wi,area*Y/1000,'linewidth',2)

linkaxes(h,'x')

ca = axis;
axis([ca(1:3) 250]);
```

```
yTicLabel = get(h(2),'ytickLabel');
yTicNum = double(yTicLabel);
yTicNum(6,:) = [32 32 32];
yTicLabel = char(yTicNum);
set(h(2),'ytickLabel',yTicLabel);

p1 = get(h(1),'position');
p2 = get(h(2),'position');
 top = p1(4)+p1(2);
bottom = p2(2);
height = (top-bottom)/2;

p2(4) = height;
p1(4) = height;

set(h(2),'position',p2);
set(h(1),'position',[p1(1) p2(2)+p2(4) p1(3:4)])

logy = log(mortality);
I = find(logy>0);
[p,S,mu] = polyfit(wi(I),logy(I),3);
xFit = linspace(min(wi),max(wi),1001);
yFit = polyval(p,xFit,S,mu);

y2500 = exp(polyval(p,2500,S,mu));

set(gcf,'CurrentAxes',h(1));
hold on;
semilogy(xFit,exp(yFit),'linewidth',2);
hh = get(gca,'Children');
set(gca,'Children',hh(2:-1:1));

figure(2);clf;
plot(wi,logy,xFit,yFit);

[yMin,iMin] = min(yFit);
xMin = xFit(iMin)
hold on

plot(xMin,yMin,'or')
```

Manuscript 3: General Descriptive Analysis

SAS 9.2

```
PROC IMPORT OUT= WORK.REACH
        DATAFILE= "D:\My Documents\GSPH-Dissertation\REACH LBW paper\Spring 2010
Analysis\REACH DATA2004-2009.xls"
        DBMS=EXCEL2000 REPLACE;
    SHEET="'data'";
    GETNAMES=YES;
RUN;

proc print data=work.REACH (obs=10);
run;

/**CREATING/ALTERING VARIABLES**/
DATA REACH;
set work.REACH;
if NumberANCVisits ge 3 then ANC=1;
if NumberANCVisits lt 3 then ANC=0;
run;
DATA REACH;
set work.REACH;
if BW = ' ' then LBW = '.' ;
if BW lt 2.5 then LBW=1;
if BW ge 2.5 then LBW=0;
run;
DATA REACH;
set work.REACH;
if BW = 2.5 then BW25 = 1;
run;
DATA REACH;
set work.REACH;
if DOBLMP lt 259 then preterm=1;
if DOBLMP ge 259 then preterm=0;
if DOBLMP = ' ' then preterm= ' ';
run;
DATA REACH;
set work.REACH;
MothersAgeAtBirth = MothersAge - (2010 - Yearofbirth);
run;
DATA REACH;
set work.REACH;
```

```
if YearOfBirth lt 2007 then y2007 = 0 ;
if YearOfBirth ge 2007 then y2007 = 1 ;
run;

/**GENERAL FREQUENCIES**/
proc means data=REACH; var BW; run;
proc freq data=REACH; table BW; run;
proc freq data=REACH; table BW25; run;
proc freq data=REACH; table isthereabirthweightrecord; run;
proc freq data=REACH; table ANC; run;
proc freq data=REACH; table infantdeath childdeath; run;
proc means data=REACH; var AgeAtDeathDays; run;
proc freq data=REACH; table LBW; run;

proc freq data = REACH;
where IsThereABirthWeightRecord = 'YES';
table LBW preterm PretermBirthWeeks sexofchild yearofbirth InfantDeath ChildDeath
placeofbirth deliverytype familytype religion caste dietary Motherseducation Fatherseducation
Cosanguinous_marriage111 Marriage_relation112 literate116 religion121 cast122
Mother_occupation128 Previous_child_death206 Asthma diabetes ANC;
run;

proc freq data = REACH;
where IsThereABirthWeightRecord = 'YES';
table    thyroid malaria jaundice panmasala alcohol smoke TB_19 TB_Symptoms_21
Drinking_H2O_23 cooking_H2O_24 H2O_source_location_25 Treat_H2O_28       toilet30
religion33 caste34      scheduled_caste35 cookf37a cookt38 cookc39 HOUSE_TYPE42
own_house44 own_land45 ration_Card49 salt52 ANC;
run;

proc means data=REACH;
where IsThereABirthWeightRecord='YES';
var BW MothersAge DOBLMP AgeAtDeathDays NumberANCvisits FamilyIncomeInformation
Mother_education114 Father_education124 Birthweight435 rooms43 acres_owned462
        agri471;
run;
```

Manuscript 3: Outcome Analysis

Birth-weight as a continuous variable

```
PROC GLM DATA=REACHDATA;
        CLASS EXCRETA_RECODE;
        MODEL IMPUTEDBW = EXCRETA_RECODE / SOLUTION SS3;
        TITLE 'BW CONTINUOUS VS EXCRETA';
RUN;
QUIT;
PROC GLM DATA=REACHDATA;
        CLASS REFUSE_RECODE;
        MODEL IMPUTEDBW = REFUSE_RECODE / SOLUTION SS3;
        TITLE 'BW CONTINUOUS VS REFUSE';
RUN;
QUIT;
PROC GLM DATA=REACHDATA;
        CLASS SEWAGE_RECODE;
        MODEL IMPUTEDBW = SEWAGE_RECODE / SOLUTION SS3;
        TITLE 'BW CONTINUOUS VS SEWAGE';
RUN;
QUIT;
PROC GLM DATA=REACHDATA;
        CLASS WATER_RECODE;
        MODEL IMPUTEDBW = WATER_RECODE / SOLUTION SS3;
        TITLE 'BW CONTINUOUS VS WATER';
RUN;
QUIT;
PROC GLM DATA=REACHDATA;
        CLASS COOKING_RECODE;
        MODEL IMPUTEDBW = COOKING_RECODE / SOLUTION SS3;
        TITLE 'BW CONTINUOUS VS COOKING';
RUN;
QUIT;
PROC GLM DATA=REACHDATA;
        CLASS DRINKINGH20_RECODE;
        MODEL IMPUTEDBW = DRINKINGH20_RECODE / SOLUTION SS3;
        TITLE 'BW CONTINUOUS VS DRINKING WATER';
RUN;
QUIT;
PROC REG DATA=REACHDATA;
        MODEL IMPUTEDBW = DRINK_IMP;
        TITLE 'BW CONTINOUS VS DRINKING WATER DICHOTOMIZED';
RUN;
```

```
QUIT;
PROC GLM DATA=REACHDATA;
        CLASS TOILET_RECODE;
        MODEL IMPUTEDBW = TOILET_RECODE / SOLUTION SS3;
        TITLE 'BW CONTINUOUS VS TOILET';
RUN;
QUIT;
PROC REG DATA=REACHDATA;
        MODEL IMPUTEDBW = TOILET_IMP;
        TITLE 'BW CONTINOUS VS TOILET DICHOTOMIZED';
RUN;
QUIT;
PROC GLM DATA=REACHDATA;
        CLASS COOKINGH20_RECODE;
        MODEL IMPUTEDBW = COOKINGH20_RECODE / SOLUTION SS3;
        TITLE 'BW CONTINUOUS VS COOKING WATER';
RUN;
QUIT;
PROC REG DATA=REACHDATA;
        MODEL IMPUTEDBW = COOK_IMP;
        TITLE 'BW CONTINOUS VS COOKING WATER DICHOTOMIZED';
RUN;
QUIT;
PROC GLM DATA=REACHDATA;
        CLASS COOKTYPE_RECODE;
        MODEL IMPUTEDBW = COOKTYPE_RECODE / SOLUTION SS3;
        TITLE 'BW CONTINUOUS VS COOK TYPE';
RUN;
QUIT;
PROC GLM DATA=REACHDATA;
        CLASS COOKFUEL_RECODE;
        MODEL IMPUTEDBW = COOKFUEL_RECODE / SOLUTION SS3;
        TITLE 'BW CONTINUOUS VS COOKING FUEL';
RUN;
QUIT;
PROC REG DATA=REACHDATA;
        MODEL IMPUTEDBW = COOK_CLEAN;
        TITLE 'BW CONTINOUS VS COOKING FUEL DICHOTOMIZED';
RUN;
QUIT;
```

```
*UNIVARIATE ASSOCIATIONS WITH VARIABLES OTHER THAN SANITATION VARIABLES;
PROC REG DATA=REACHDATA;
        MODEL IMPUTEDBW = OWN_HOUSE44;
        TITLE 'BW CONTINOUS VS HOME OWNERSHIP';
RUN;
QUIT;
PROC REG DATA=REACHDATA;
        MODEL IMPUTEDBW = OWN_LAND45;
        TITLE 'BW CONTINOUS VS LAND OWNDERSHIP';
RUN;
QUIT;
PROC REG DATA=REACHDATA;
        MODEL IMPUTEDBW = SEXOFCHILD;
        TITLE 'BW CONTINOUS VS SEX OF CHILD';
RUN;
QUIT;
PROC REG DATA=REACHDATA;
        MODEL IMPUTEDBW = YEAROFBIRTH;
        TITLE 'BW CONTINOUS VS YEAR OF BIRTH';
RUN;
QUIT;
PROC REG DATA=REACHDATA;
        MODEL IMPUTEDBW = NUMBERANCVISITS;
        TITLE 'BW CONTINOUS VS NUMBER OF ANC VISITS';
RUN;
QUIT;
PROC GLM DATA=REACHDATA;
        CLASS PLACEOFBIRTH;
        MODEL IMPUTEDBW = PLACEOFBIRTH / SOLUTION SS3;
        TITLE 'BW CONTINUOUS VS PLACEOFBIRTH';
RUN;
QUIT;
PROC GLM DATA=REACHDATA;
        CLASS DELIVERYTYPE;
        MODEL IMPUTEDBW = DELIVERYTYPE / SOLUTION SS3;
        TITLE 'BW CONTINUOUS VS DELIVERYTYPE';
RUN;
QUIT;
PROC REG DATA=REACHDATA;
        MODEL IMPUTEDBW = FAMILYTYPE;
        TITLE 'BW CONTINOUS VS FAMILYTYPE';
RUN;
QUIT;
PROC GLM DATA=REACHDATA;
```

```
        CLASS RELIGION;
        MODEL IMPUTEDBW = RELIGION / SOLUTION SS3;
        TITLE 'BW CONTINUOUS VS RELIGION';
RUN;
QUIT;
PROC GLM DATA=REACHDATA;
        CLASS CASTE;
        MODEL IMPUTEDBW = CASTE / SOLUTION SS3;
        TITLE 'BW CONTINUOUS VS CASTE';
RUN;
QUIT;
PROC GLM DATA=REACHDATA;
        CLASS DIETARY;
        MODEL IMPUTEDBW = DIETARY / SOLUTION SS3;
        TITLE 'BW CONTINUOUS VS DIETARY';
RUN;
QUIT;
PROC GLM DATA=REACHDATA;
        CLASS MOTHERSEDUCATION;
        MODEL IMPUTEDBW = MOTHERSEDUCATION / SOLUTION SS3;
        TITLE 'BW CONTINUOUS VS MOTHERSEDUCATION';
RUN;
QUIT;
PROC GLM DATA=REACHDATA;
        CLASS FATHERSEDUCATION;
        MODEL IMPUTEDBW = FATHERSEDUCATION / SOLUTION SS3;
        TITLE 'BW CONTINUOUS VS FATHERSEDUCATION';
RUN;
QUIT;
PROC GLM DATA=REACHDATA;
        CLASS HTYPE;
        MODEL IMPUTEDBW = HTYPE / SOLUTION SS3;
        TITLE 'BW CONTINUOUS VS HOUSE TYPE';
RUN;
QUIT;
PROC REG DATA=REACHDATA;
        MODEL IMPUTEDBW = RATION_POOR;
        TITLE 'BW CONTINOUS VS RATION_POOR';
RUN;
QUIT;
PROC REG DATA=REACHDATA;
        MODEL IMPUTEDBW = familyincomeinformation;
        TITLE 'BW CONTINOUS VS familyincomeinformation';
RUN;
```

```
QUIT;
PROC REG DATA=REACHDATA;
        MODEL IMPUTEDBW = mothersage;
        TITLE 'BW CONTINOUS VS mothersage';
RUN;
QUIT;
```

Birth-weight as a categorical variable

*1. Sanitation

```
PROC LOGISTIC DATA=REACHDATA;
        CLASS EXCRETA_DISP (REF='HOUSEHOLD') / PARAM=REF;
        MODEL BW_CAT (EVENT='1')= EXCRETA_DISP / LINK=CLOGIT;
RUN;
PROC LOGISTIC DATA=REACHDATA;
        CLASS REFUSE_DISP (REF='COMPOSTING') / PARAM=REF;
        MODEL BW_CAT (EVENT='1')=REFUSE_DISP / LINK=CLOGIT;
RUN;
PROC LOGISTIC DATA=REACHDATA;
        CLASS SWATER_DISP (REF='CLOSED DRIANAGE') / PARAM=REF;
        MODEL BW_CAT (EVENT='1')=SWATER_DISP / LINK=CLOGIT;
RUN;
PROC LOGISTIC DATA=REACHDATA;
        CLASS WATER (REF='HOUSEHOLD SUPPLY') / PARAM=REF;
        MODEL BW_CAT (EVENT='1')=WATER / LINK=CLOGIT;
RUN;
PROC LOGISTIC DATA=REACHDATA;
        CLASS COOKING (REF='ELECTRIC') / PARAM=REF;
        MODEL BW_CAT (EVENT='1')=COOKING / LINK=CLOGIT;
RUN;
PROC LOGISTIC DATA=REACHDATA;
        CLASS DRINK_IMP (REF='1') / PARAM=REF;
        MODEL BW_CAT (EVENT='1')=DRINK_IMP / LINK=CLOGIT;
RUN;
PROC LOGISTIC DATA=REACHDATA;
        CLASS DRINKINGH20_RECODE (REF='Bottled water') / PARAM=REF;
        MODEL BW_CAT (EVENT='1')=DRINKINGH20_RECODE / LINK=CLOGIT;
RUN;
proc freq data=reachdata;
        table cookf37a;
run;
```

```
PROC LOGISTIC DATA=REACHDATA;
        CLASS COOKF37A  (REF='Coal Lignite') / PARAM=REF;
        MODEL BW_CAT (EVENT='1')=COOKF37A  / LINK=CLOGIT;
RUN;
PROC LOGISTIC DATA=REACHDATA;
        CLASS COOK_CLEAN (REF='1') / PARAM=REF;
        MODEL BW_CAT (EVENT='1')=COOK_CLEAN / LINK=CLOGIT;
RUN;

*UNIVARIATE ASSOCIATIONS WITH VARIABLES OTHER THAN SANITATION VARIABLES;
PROC FREQ DATA=REACHDATA;
TABLE BW_CAT*SEXOFCHILD / CHISQ;
TITLE 'IM VS SEX OF CHILD';
RUN;
PROC LOGISTIC DATA=REACHDATA;
CLASS SEXOFCHILD (REF='MALE') / PARAM=REF;
MODEL BW_CAT (EVENT='1') = SEXOFCHILD / LINK=CLOGIT;
TITLE 'BW_CAT VS SEX OF CHILD';
RUN;
*
PROC LOGISTIC DATA=REACHDATA;
MODEL BW_CAT (EVENT='1') = YEAROFBIRTH / LINK=CLOGIT;
TITLE 'BW_CAT VS YEAR OF BIRTH';
RUN;
*
proc logistic data=REACHDATA;
class PLACEOFBIRTH(ref='PVT') / param=ref;
model BW_CAT(event='1')= PLACEOFBIRTH / LINK=CLOGIT;
title 'BW_CAT vs PLACE OF BIRTH';
RUN;
*
proc logistic data=REACHDATA;
class DELIVERYTYPE (ref='NORMAL') / param=ref;
model BW_CAT(event='1')= DELIVERYTYPE / LINK=CLOGIT;
TITLE 'BW_CAT VS DELIVERY TYPE';
RUN;
PROC FREQ DATA=REACHDATA;
TABLE BW_CAT*DELIVERYTYPE / CHISQ;
TITLE 'BW_CAT VS DELIVERY TYPE';
RUN;
*
PROC FREQ DATA=REACHDATA;
TABLE BW_CAT*FAMILYTYPE / CHISQ;
TITLE 'BW_CAT VS FAMILY TYPE';
```

```
RUN;
PROC LOGISTIC DATA=REACHDATA;
CLASS FAMILYTYPE (REF='NUCLEAR FAMILY') / PARAM=REF;
MODEL BW_CAT(EVEN='1')=FAMILYTYPE / LINK=CLOGIT;
TITLE 'BW_CAT VS FAMILY TYPE';
RUN;
*
proc logistic data=REACHDATA;
class Caste (ref='FORWARD CASTE') / param=ref;
model BW_CAT(event='1')= CASTE / LINK=CLOGIT;
title 'BW_CAT vs CASTE';
RUN;
*
proc logistic data=REACHDATA;
class MOTHERSEDUCATION (ref='HIGH SCHOOL') / param=ref;
model BW_CAT(event='1')= MOTHERSEDUCATION / LINK=CLOGIT;
title 'BW_CAT vs MOTHERS EDUCATION';
RUN;
PROC FREQ DATA=REACHDATA;
TABLE BW_CAT*MOTHERSEDUCATION / CHISQ;
TITLE 'BW_CAT VS MOTHERS EDUCATION TABLE';
RUN;
*
proc logistic data=REACHDATA;
class FATHERSEDUCATION (ref='HIGH SCHOOL') / param=ref;
model BW_CAT(event='1')= FATHERSEDUCATION / LINK=CLOGIT;
title 'BW_CAT vs FATHERS EDUCATION';
RUN;
*

*
PROC FREQ DATA=REACHDATA;
TABLE BW_CAT*REL_OCCU / CHISQ;
TITLE 'BW VS OCCUPATION';
RUN;
PROC FREQ DATA=REACHDATA;
TABLE BW_CAT*VILLAGE / CHISQ;
TITLE 'BW VS VILLAGE';
RUN;
*
PROC FREQ DATA=REACHDATA;
TABLE BW_CAT*RELIGION33 / CHISQ;
TITLE 'BW VS RELIGION33';
RUN;
```

```
PROC LOGISTIC DATA=REACHDATA;
CLASS RELIGION33 (REF='1') / PARAM=REF;
MODEL BW_CAT=RELIGION33 / LINK=CLOGIT;
RUN;
*
PROC FREQ DATA=REACHDATA;
TABLE BW_CAT*scheduled_caste35 / CHISQ;
TITLE 'BW VS scheduled_caste35';
RUN;
*

*Multivariate logistic regression for sginificant correlates
Human waste disposal, community toilet
Drinking water, community tap
Unclean Cooking Fuel
BIRTH-WEIGHT
Birth-weight, continuous
Birth-weight, dichotomous (<2.5kg)
Birth-weight, categorical, very small (<1.7kg);
*Adjusted models for BW_CAT as the outcome and sanitation variables predicting;
PROC LOGISTIC DATA=REACHDATA;
        CLASS EXCRETA_DISP (REF='HOUSEHOLD') / PARAM=REF;
        MODEL BW_CAT3 (EVENT='1') = EXCRETA_DISP MOTHERSAGE SEXOFCHILD POB_HOME
POB_GOVT
                YEAROFBIRTH MOTHED_UNED MOTHED_PRIM MOTHED_MS FATHED_UNED
FATHED_PRIM CASTE_ST CASTE_BC CASTE_SC/ LINK=CLOGIT;
RUN;
PROC LOGISTIC DATA=REACHDATA;
        CLASS SWATER_DISP (REF='CLOSED DRIANAGE') / PARAM=REF;
        MODEL BW_CAT1 (EVENT='1') = SWATER_DISP MOTHERSAGE SEXOFCHILD POB_HOME
POB_GOVT
                YEAROFBIRTH MOTHED_UNED MOTHED_PRIM MOTHED_MS FATHED_UNED
FATHED_PRIM CASTE_ST CASTE_BC CASTE_SC/ LINK=CLOGIT;
RUN;
PROC LOGISTIC DATA=REACHDATA;
        CLASS DRINKINGH20_RECODE (REF='Bottled water') / PARAM=REF;
        MODEL BW_CAT1 (EVENT='1')=DRINKINGH20_RECODE MOTHERSAGE SEXOFCHILD
POB_HOME POB_GOVT
                YEAROFBIRTH MOTHED_UNED MOTHED_PRIM MOTHED_MS FATHED_UNED
FATHED_PRIM CASTE_ST CASTE_BC CASTE_SC/ LINK=CLOGIT;
RUN;
PROC LOGISTIC DATA=REACHDATA;
        CLASS DRINK_IMP (REF='2.00') / PARAM=REF;
```

```
        MODEL BW_CAT1 (EVENT='1')=DRINK_IMP MOTHERSAGE SEXOFCHILD POB_HOME
POB_GOVT
                YEAROFBIRTH MOTHED_UNED MOTHED_PRIM MOTHED_MS FATHED_UNED
FATHED_PRIM CASTE_ST CASTE_BC CASTE_SC/ LINK=CLOGIT;
RUN;
PROC LOGISTIC DATA=REACHDATA;
        CLASS COOK_CLEAN (REF='1.00') / PARAM=REF;
        MODEL BW_CAT1 (EVENT='1')=COOK_CLEAN MOTHERSAGE SEXOFCHILD POB_HOME
POB_GOVT
                YEAROFBIRTH MOTHED_UNED MOTHED_PRIM MOTHED_MS FATHED_UNED
FATHED_PRIM CASTE_ST CASTE_BC CASTE_SC/ LINK=CLOGIT;
RUN;
PROC LOGISTIC DATA=REACHDATA;
        CLASS REFUSE_DISP (REF='COMPOSTING') / PARAM=REF;
        MODEL BW_CAT1 (EVENT='1')=REFUSE_DISP MOTHERSAGE SEXOFCHILD POB_HOME
POB_GOVT
                YEAROFBIRTH MOTHED_UNED MOTHED_PRIM MOTHED_MS FATHED_UNED
FATHED_PRIM CASTE_ST CASTE_BC CASTE_SC/ LINK=CLOGIT;
RUN;
```

Infant mortality outcome

```
*Showing relationship between continuous BW variable and IM;
proc logistic data=REACHdata descending;
        model infantdeath (event='no')=IMPUTED1 / link=clogit;
run;
*categorized BW variables vs IM;
PROC LOGISTIC DATA=REACHDATA;
        MODEL INFANTDEATH (EVENT='no')= LBW/ LINK=CLOGIT;
RUN;
PROC LOGISTIC DATA=REACHDATA;
        CLASS BW_CAT1 (REF='3') / PARAM=REF;
        MODEL INFANTDEATH (EVEN='YES') = BW_CAT1 / LINK=CLOGIT;
RUN;

proc freq data=reachdata;
table LBW1*infantdeath;
table LBW2*infantdeath;
table LBW3*infantdeath;
table BW_CAT1*infantdeath;
table BW_CAT2*infantdeath;
table BW_CAT3*infantdeath;
run;
proc means data=reachdata;
```

```
var imputed1 imputed2 imputed3;
run;

*proc logistic data = hsb2 ;
PROC LOGISTIC DATA=REACHDATA;
        CLASS EXCRETA_DISP (REF='HOUSEHOLD') / PARAM=REF;
        MODEL INFANTDEATH (EVENT='YES')= EXCRETA_DISP / LINK=CLOGIT;
RUN;
*
PROC LOGISTIC DATA=REACHDATA;
        CLASS REFUSE_DISP (REF='COMPOSTING') / PARAM=REF;
        MODEL INFANTDEATH (EVENT='YES')=REFUSE_DISP / LINK=CLOGIT;
RUN;
*
PROC LOGISTIC DATA=REACHDATA;
        CLASS SWATER_DISP (REF='CLOSED DRIANAGE') / PARAM=REF;
        MODEL INFANTDEATH (EVENT='YES')=SWATER_DISP / LINK=CLOGIT;
RUN;
*
PROC LOGISTIC DATA=REACHDATA;
        CLASS WATER (REF='HOUSEHOLD SUPPLY') / PARAM=REF;
        MODEL INFANTDEATH (EVENT='YES')=WATER  / LINK=CLOGIT;
RUN;
*
PROC LOGISTIC DATA=REACHDATA;
        CLASS DRINK_IMP (REF='1') / PARAM=REF;
        MODEL INFANTDEATH (EVENT='YES')=DRINK_IMP  / LINK=CLOGIT;
RUN;
PROC LOGISTIC DATA=REACHDATA;
        CLASS DRINKINGH20_RECODE (REF='Bottled water') / PARAM=REF;
        MODEL INFANTDEATH (EVENT='YES')=DRINKINGH20_RECODE / LINK=CLOGIT;
RUN;
PROC FREQ DATA=REACHDATA;
TABLE DRINKINGH20_RECODE;
TABLE INFANTDEATH*DRINKINGH20_RECODE;
TABLE BW_CAT1*DRINKINGH20_RECODE;
RUN;
*
RUN;


*UNIVARIATE ASSOCIATIONS WITH VARIABLES OTHER THAN SANITATION VARIABLES;
proc logistic data=REACHDATA;
class PLACEOFBIRTH(ref='PVT') / param=ref;
```

```
model infantdeath(event='YES')= PLACEOFBIRTH / LINK=CLOGIT;
title 'IM vs PLACE OF BIRTH';
RUN;
*
proc logistic data=work.REACH;
class DELIVERYTYPE (ref='NORMAL') / param=ref;
model infantdeath(event='YES')= DELIVERYTYPE / LINK=CLOGIT;
TITLE 'IM VS DELIVERY TYPE';
RUN;
PROC FREQ DATA=REACHDATA;
TABLE INFANTDEATH*DELIVERYTYPE / CHISQ;
TITLE 'IM VS DELIVERY TYPE';
RUN;

PROC FREQ DATA=REACHDATA;
TABLE infantdeath*FAMILYTYPE / CHISQ;
TITLE 'IM VS FAMILY TYPE';
RUN;
proc logistic data=reachdata;
class familytype (ref='NUCLEAR FAMILY') / param=ref;
model infantdeath(event='yes')=familytype / link=clogit;
title 'IM vs family type';
run;
*
proc logistic data=REACHDATA;
class Caste (ref='FORWARD CASTE') / param=ref;
model infantdeath(event='YES')= CASTE / LINK=CLOGIT;
title 'IM vs CASTE';
RUN;
*
proc logistic data=REACHDATA;
class MOTHERSEDUCATION (ref='HIGH SCHOOL') / param=ref;
model infantdeath(event='YES')= MOTHERSEDUCATION / LINK=CLOGIT;
title 'IM vs MOTHERS EDUCATION';
RUN;
proc logistic data=REACHDATA;
class FATHERSEDUCATION (ref='HIGH SCHOOL') / param=ref;
model infantdeath(event='YES')= FATHERSEDUCATION / LINK=CLOGIT;
title 'IM vs FATHERS EDUCATION';
RUN;

proc logistic data=REACHdata descending;
        model infantdeath (event='YES')= mothersage / link=clogit;
        title 'IM vs mothers age';
```

```
run;
*
PROC FREQ DATA=REACHDATA;
TABLE infantdeath*REL_OCCU / CHISQ;
TITLE 'IM VS OCCUPATION';
RUN;
*Significant by chisquare, but can't make any sense of it;
PROC FREQ DATA=REACHDATA;
TABLE infantdeath*Isthereabirthweightrecord / CHISQ;
TITLE 'IM VS Isthereabirthweightrecord';
RUN;
*

*Multivariate logistic regression for sginificant correlates
Human waste disposal, community toilet
Drinking water, community tap
Unclean Cooking Fuel
BIRTH-WEIGHT
Birth-weight, continuous
Birth-weight, dichotomous (<2.5kg)
Birth-weight, categorical, very small (<1.7kg)
;

*Adjusted models for IM as the outcome and sanitation variables predicting;
PROC LOGISTIC DATA=REACHDATA;
        CLASS EXCRETA_DISP (REF='HOUSEHOLD') / PARAM=REF;
        MODEL INFANTDEATH (EVENT='YES') = EXCRETA_DISP MOTHERSAGE
POB_HOME*DELIVERYTYPE FAMILYTYPE MOTHED_UNED FATHED_UNED
                FATHED_PRIM CASTE_ST BW_CAT1/ LINK=CLOGIT;
RUN;
PROC LOGISTIC DATA=REACHDATA;
        CLASS SWATER_DISP (REF='CLOSED DRIANAGE') / PARAM=REF;
        MODEL INFANTDEATH (EVENT='YES') = SWATER_DISP MOTHERSAGE POB_HOME
DELIVERYTYPE FAMILYTYPE MOTHED_UNED FATHED_UNED
                FATHED_PRIM CASTE_ST BW_CAT/ LINK=CLOGIT;
RUN;
PROC LOGISTIC DATA=REACHDATA;
        CLASS DRINKINGH20_RECODE (REF='Bottled water') / PARAM=REF;
        MODEL INFANTDEATH (EVENT='YES')=DRINKINGH20_RECODE MOTHERSAGE
POB_HOME FAMILYTYPE MOTHED_UNED FATHED_UNED
                FATHED_PRIM CASTE_ST BW_CAT3/ LINK=CLOGIT;
RUN;
PROC LOGISTIC DATA=REACHDATA;
        CLASS DRINK_IMP (REF='1') / PARAM=REF;
```

```
        MODEL INFANTDEATH (EVENT='YES')=DRINK_IMP MOTHERSAGE POB_HOME
DELIVERYTYPE FAMILYTYPE MOTHED_UNED FATHED_UNED
                FATHED_PRIM CASTE_ST BW_CAT/ LINK=CLOGIT;
RUN;
PROC LOGISTIC DATA=REACHDATA;
        CLASS COOK_CLEAN (REF='1') / PARAM=REF;
        MODEL INFANTDEATH (EVENT='YES')=COOK_CLEAN MOTHERSAGE POB_HOME
DELIVERYTYPE FAMILYTYPE MOTHED_UNED FATHED_UNED
                FATHED_PRIM CASTE_ST BW_CAT/ LINK=CLOGIT;
RUN;
PROC LOGISTIC DATA=REACHDATA;
        CLASS REFUSE_DISP (REF='COMPOSTING') / PARAM=REF;
        MODEL INFANTDEATH (EVENT='YES')=REFUSE_DISP MOTHERSAGE POB_HOME
DELIVERYTYPE FAMILYTYPE MOTHED_UNED FATHED_UNED
                FATHED_PRIM CASTE_ST BW_CAT/ LINK=CLOGIT;
RUN;



*BW VS IM, ALL THE DIFF BW VARIABLES;
proc logistic data=REACHdata descending;
        model infantdeath (event='no')=IMPUTED1 MOTHERSAGE POB_HOME DELIVERYTYPE
FAMILYTYPE MOTHED_UNED FATHED_UNED
                FATHED_PRIM CASTE_ST / link=clogit;
run;
*categorized BW variables vs IM;
PROC LOGISTIC DATA=REACHDATA;
        MODEL INFANTDEATH (EVENT='no')= LBW MOTHERSAGE POB_HOME DELIVERYTYPE
FAMILYTYPE MOTHED_UNED FATHED_UNED
                FATHED_PRIM CASTE_ST / LINK=CLOGIT;
RUN;
PROC LOGISTIC DATA=REACHDATA;
        CLASS BW_CAT (REF='3') / PARAM=REF;
        MODEL INFANTDEATH (EVENT='YES') = BW_CAT MOTHERSAGE POB_HOME
DELIVERYTYPE FAMILYTYPE MOTHED_UNED FATHED_UNED
                FATHED_PRIM CASTE_ST / LINK=CLOGIT;
RUN;

proc freq data=REACHDATA;
        table Drinking_H2O_23;
        table BW_cat;
        table EXCRETA_DISP;
        table caste*Drinking_H2O_23;
run;
```

BIBLIOGRAPHY

1. Attaran A. An immeasurable crisis? A criticism of the millennium development goals and why they cannot be measured. PLoS Med 2005;2:e318.
2. Sachs JD. Health in the developing world: achieving the Millennium Development Goals. Bull World Health Organ 2004;82:947-9; discussion 50-2.
3. Bale JR, Stoll BJ, Lucas AO. Improving birth outcomes meeting the challenges in the developing world. In. Washington, DC: National Academies Press; 2003:372 p.
4. Barker DJ. Developmental origins of adult health and disease. J Epidemiol Community Health 2004;58:114-5.
5. Osmani S, Sen A. The hidden penalties of gender inequality: fetal origins of ill-health. Econ Hum Biol 2003;1:105-21.
6. Yajnik CS, Fall CH, Coyaji KJ, et al. Neonatal anthropometry: the thin-fat Indian baby. The Pune Maternal Nutrition Study. Int J Obes Relat Metab Disord 2003;27:173-80.
7. Eriksen W, Sundet JM, Tambs K. Birth weight standardized to gestational age and intelligence in young adulthood: a register-based birth cohort study of male siblings. Am J Epidemiol 2010;172:530-6.
8. McCormick MC. The Contribution of Low Birthweight to Infant Mortality and Childhood Morbidity. New England Journal of Medicine 1985;312:82-90.
9. Silverman WA, Sinclair JC. Infants of low birth weight. N Engl J Med 1966;274:448-50.
10. Public Health Aspects of Low Birth Weight: Third Report of the Expert Committee on Maternal and Child Health. Geneva, Switzerland: World Health Organization,; 1961.
11. Wilcox AJ. On the importance--and the unimportance-- of birthweight. Int J Epidemiol 2001;30:1233-41.
12. Steffensen FH, Sorensen HT, Gillman MW, et al. Low birth weight and preterm delivery as risk factors for asthma and atopic dermatitis in young adult males. Epidemiology 2000;11:185-8.
13. Richards M, Hardy R, Kuh D, Wadsworth ME. Birth weight and cognitive function in the British 1946 birth cohort: longitudinal population based study. BMJ 2001;322:199-203.
14. Richards M, Hardy R, Kuh D, Wadsworth ME. Birthweight, postnatal growth and cognitive function in a national UK birth cohort. Int J Epidemiol 2002;31:342-8.
15. Barker DJ, Osmond C, Simmonds SJ, Wield GA. The relation of small head circumference and thinness at birth to death from cardiovascular disease in adult life. BMJ 1993;306:422-6.
16. Basso O. Birth weight is forever. Epidemiology 2008;19:204-5.
17. Godfrey KM, Barker DJ. Fetal nutrition and adult disease. Am J Clin Nutr 2000;71:1344S-52S.
18. MacDorman MF, Atkinson JO. Infant mortality statistics from the 1997 period linked birth/infant death data set. Natl Vital Stat Rep 1999;47:1-23.
19. Rothman KJ, Greenland S, Lash TL. Modern epidemiology. 3rd ed. Philadelphia: Wolters Kluwer Health/Lippincott Williams & Wilkins; 2008.
20. Blanc AK, Wardlaw T. Monitoring low birth weight: an evaluation of international estimates and an updated estimation procedure. Bull World Health Organ 2005;83:178-85.
21. Boerma JT, Weinstein KI, Rutstein SO, Sommerfelt AE. Data on birth weight in developing countries: can surveys help? Bull World Health Organ 1996;74:209-16.
22. Robles A, Goldman N. Can accurate data on birthweight be obtained from health interview surveys? Int J Epidemiol 1999;28:925-31.
23. Rekha C, Whelan RM, Reddy P, Reddy PS. Evaluation of adjustment factors to reduce heaping in birth-weight data from rural India. In preparation 2011.

24. Fox MP, Lash TL, Greenland S. A method to automate probabilistic sensitivity analyses of misclassified binary variables. Int J Epidemiol 2005;34:1370-6.
25. Lash TL, Fink AK. Semi-automated sensitivity analysis to assess systematic errors in observational data. Epidemiology 2003;14:451-8.
26. Lash TL, Fox MP, Fink AK. Applying quantitative bias analysis to epidemiologic data. In: Statistics for biology and health. Dordrecht ; New York: Springer; 2009:xii, 192 p.
27. Carroll RJ, Stefanski LA. Approximate Quasi-Likelihood Estimation in Models with Surrogate Predictors. Journal of the American Statistical Association 1990;85:652-63.
28. Carroll RJ, Stefanski LA. Measurement Error, Instrumental Variables and Corrections for Attenuation with Applications to Metaanalyses. Statistics in Medicine 1994;13:1265-82.
29. Carroll RJ, Ruppert D, Stefanski LA. Measurement Error in Non-Linear Models. London: Chapman and Hall; 1995.
30. Thurston SW, Spiegelman D, Ruppert D. Equivalence of regression calibration methods in main study/external validation study designs. J Stat Plan Infer 2003;113:527-39.
31. Thurston SW, Williams PL, Hauser R, Hu H, Hernandez-Avila M, Spiegelman D. A comparison of regression calibration approaches for designs with internal validation data. J Stat Plan Infer 2005;131:175-90.
32. Spiegelman D, Carroll RJ, Kipnis V. Efficient regression calibration for logistic regression in main study/internal validation study designs with an imperfect reference instrument. Stat Med 2001;20:139-60.
33. Rosner B, Spiegelman D, Willett WC. Correction of logistic regression relative risk estimates and confidence intervals for random within-person measurement error. Am J Epidemiol 1992;136:1400-13.
34. Rosner B, Willett WC, Spiegelman D. Correction of logistic regression relative risk estimates and confidence intervals for systematic within-person measurement error. Stat Med 1989;8:1051-69; discussion 71-3.
35. Rosner B, Spiegelman D, Willett WC. Correction of logistic regression relative risk estimates and confidence intervals for measurement error: the case of multiple covariates measured with error. Am J Epidemiol 1990;132:734-45.
36. Pethybridge RJ, Ashford JR, Fryer JG. Some features of the distribution of birthweight of human infants. Br J Prev Soc Med 1974;28:10-8.
37. Wilcox AJ, Russell IT. Birthweight and perinatal mortality: I. On the frequency distribution of birthweight. Int J Epidemiol 1983;12:314-8.
38. Umbach DM, Wilcox AJ. A technique for measuring epidemiologically useful features of birthweight distributions. Stat Med 1996;15:1333-48.
39. Wilcox AJ, Russell IT. Birthweight and perinatal mortality: III. Towards a new method of analysis. Int J Epidemiol 1986;15:188-96.
40. Wilcox AJ, Skjaerven R. Birth weight and perinatal mortality: the effect of gestational age. Am J Public Health 1992;82:378-82.
41. Wilcox AJ, Russell IT. Birthweight and perinatal mortality: II. On weight-specific mortality. Int J Epidemiol 1983;12:319-25.
42. The Analysis of Birth Weight and Infant Mortality: An Alternative Hypothesis. 2002. (Accessed 2010, at http://eb.niehs.nih.gov/bwt.)
43. Wilcox A, Skjaerven R, Buekens P, Kiely J. Birth weight and perinatal mortality. A comparison of the United States and Norway. JAMA 1995;273:709-11.
44. Humphrey C, Elford J. Social class differences in infant mortality: the problem of competing hypotheses. J Biosoc Sci 1988;20:497-504.
45. Mittendorf R, Herschel M, Williams MA, Hibbard JU, Moawad AH, Lee KS. Reducing the frequency of low birth weight in the United States. Obstet Gynecol 1994;83:1056-9.

46. Yerushalmy J. The relationship of parents' cigarette smoking to outcome of pregnancy--implications as to the problem of inferring causation from observed associations. Am J Epidemiol 1971;93:443-56.
47. Hernandez-Diaz S, Wilcox AJ, Schisterman EF, Hernan MA. From causal diagrams to birth weight-specific curves of infant mortality. Eur J Epidemiol 2008;23:163-6.
48. Basso O, Wilcox AJ, Weinberg CR. Birth weight and mortality: causality or confounding? Am J Epidemiol 2006;164:303-11.
49. Basso O, Wilcox AJ. Intersecting birth weight-specific mortality curves: solving the riddle. Am J Epidemiol 2009;169:787-97.
50. Wilcox AJ. Fertility and Pregnancy: An Epidemiologic Perspective. New York, NY: Oxford University Press; 2010.
51. UN Inter-agency Group for Child Mortality Estimation. Level & Trends in Child Mortality. Report 2010.; 2010.
52. UNICEF. State of the World's Children Report 2008. . New York; 2008.
53. Reproductive and Child Health Project. 2010. (Accessed 2011, at http://mohfw.nic.in/NRHM.htm.)
54. Uner S, Cakir B, Yildirak K. Do we adequately respect the potential of routine primary health care services in reducing neonatal mortality in developing countries? The example of the Denizli cohort. Cah Sociol Demogr Med 2010;50:477-99.
55. Jackson AA, Bhutta ZA, Lumbiganon P. Nutrition as a preventative strategy against adverse pregnancy outcomes. Introduction. J Nutr 2003;133:1589S-91S.
56. Mattson S. Millennium development goals and global women's and infants' health. J Obstet Gynecol Neonatal Nurs 2010;39:573-9.
57. Claeson M, Bos ER, Mawji T, Pathmanathan I. Reducing child mortality in India in the new millennium. Bull World Health Organ 2000;78:1192-9.
58. Bassani DG, Kumar R, Awasthi S, et al. Causes of neonatal and child mortality in India: a nationally representative mortality survey. Lancet 2010;376:1853-60.
59. Pandey A, Choe MK, Luther NY, Sahu D, Chand J. Infant and Child Mortality in India. Mumbai, India: International Institute for Population Sciences and East-West Center Program on Population; 1998.
60. Lahariya C, Sudfeld CR, Lahariya D, Tomar SS. Causes of child deaths in India, 1985-2008: a systematic review of literature. Indian J Pediatr 2010;77:1303-11.
61. Ramachandran P. Nutrition and child survival in India. Indian J Pediatr 2010;77:301-5.
62. Shrimpton R. Preventing low birthweight and reduction of child mortality. Trans R Soc Trop Med Hyg 2003;97:39-42.
63. Alexander GR, Cornely DA. Prenatal care utilization: its measurement and relationship to pregnancy outcome. Am J Prev Med 1987;3:243-53.
64. Alexander GR, Kotelchuck M. Assessing the role and effectiveness of prenatal care: history, challenges, and directions for future research. Public Health Rep 2001;116:306-16.
65. Balcazar H, Hartner J, Cole G. The effects of prenatal care utilization and maternal risk factors on pregnancy outcome between Mexican Americans and non-Hispanic whites. J Natl Med Assoc 1993;85:195-202.
66. Fiscella K. Does prenatal care improve birth outcomes? A critical review. Obstet Gynecol 1995;85:468-79.
67. Heaman MI, Newburn-Cook CV, Green CG, Elliott LJ, Helewa ME. Inadequate prenatal care and its association with adverse pregnancy outcomes: a comparison of indices. BMC Pregnancy Childbirth 2008;8:15.

68. Fall CH, Yajnik CS, Rao S, Davies AA, Brown N, Farrant HJ. Micronutrients and fetal growth. J Nutr 2003;133:1747S-56S.
69. Dreyfuss ML, Msamanga GI, Spiegelman D, et al. Determinants of low birth weight among HIV-infected pregnant women in Tanzania. American Journal of Clinical Nutrition 2001;74:814-26.
70. Allen LH. Anemia and iron deficiency: effects on pregnancy outcome. Am J Clin Nutr 2000;71:1280S-4S.
71. Allen LH. Biological mechanisms that might underlie iron's effects on fetal growth and preterm birth. J Nutr 2001;131:581S-9S.
72. Brabin BJ, Ginny M, Sapau J, Galme K, Paino J. Consequences of maternal anaemia on outcome of pregnancy in a malaria endemic area in Papua New Guinea. Ann Trop Med Parasitol 1990;84:11-24.
73. van den Broek N. Anaemia in pregnancy in developing countries. Br J Obstet Gynaecol 1998;105:385-90.
74. Casey BM, Dashe JS, Wells CE, et al. Subclinical hypothyroidism and pregnancy outcomes. Obstet Gynecol 2005;105:239-45.
75. Casey BM, Dashe JS, Spong CY, McIntire DD, Leveno KJ, Cunningham GF. Perinatal significance of isolated maternal hypothyroxinemia identified in the first half of pregnancy. Obstet Gynecol 2007;109:1129-35.
76. Davis LE, Leveno KJ, Cunningham FG. Hypothyroidism complicating pregnancy. Obstet Gynecol 1988;72:108-12.
77. Leung AS, Millar LK, Koonings PP, Montoro M, Mestman JH. Perinatal outcome in hypothyroid pregnancies. Obstet Gynecol 1993;81:349-53.
78. Steketee RW. Pregnancy, Nutrition and Parasitic Diseases. Journal of Nutrition 2003;133:1161S-1167S.
79. Stratton JA, Miller RD, Schmidt P. Effect of maternal parasitic disease on the neonate. American Journal of Reproductive Immunology and Microbiology 1985;8:141-2.
80. Weigel MM, Calle A, Armijos RX, Vega IP, Bayas BV, Montenegro CE. The effect of chronic intestinal parasitic infection on maternal and perinatal outcome. International Journal of Gynecology & Obstetrics 1996;52:9-17.
81. Villar J, Klebanoss M, Kestler E. The Effect on Fetal Growth of Protozoan and Helminthic Infection During Pregnancy. Obstetrics and Gynecology 1989;74:915-20.
82. Hillier SL, Nugent RP, Eschenbach DA, et al. Association between bacterial vaginosis and preterm delivery of a low-birth-weight infant. The Vaginal Infections and Prematurity Study Group. N Engl J Med 1995;333:1737-42.
83. Svare JA, Schmidt H, Hansen BB, Lose G. Bacterial vaginosis in a cohort of Danish pregnant women: prevalence and relationship with preterm delivery, low birthweight and perinatal infections. BJOG 2006;113:1419-25.
84. Pararas MV, Skevaki CL, Kafetzis DA. Preterm birth due to maternal infection: Causative pathogens and modes of prevention. Eur J Clin Microbiol Infect Dis 2006;25:562-9.
85. Hirve SS, Ganatra BR. Determinants of low birth weight: a community based prospective cohort study. Indian Pediatr 1994;31:1221-5.
86. Pathak P, Kapil U, Yajnik CS, Kapoor SK, Dwivedi SN, Singh R. Iron, folate, and vitamin B12 stores among pregnant women in a rural area of Haryana State, India. Food Nutr Bull 2007;28:435-8.
87. Shah D, Sachdev HP. Maternal micronutrients and fetal outcome. Indian J Pediatr 2004;71:985-90.
88. Rao S, Yajnik CS, Kanade A, et al. Intake of micronutrient-rich foods in rural Indian mothers is associated with the size of their babies at birth: Pune Maternal Nutrition Study. J Nutr 2001;131:1217-1224.
89. Ramachandran P. Poverty nutrition linkages. Indian J Med Res 2007;126:249-61.

90. Muthayya S, Kurpad AV, Duggan CP, et al. Low maternal vitamin B12 status is associated with intrauterine growth retardation in urban South Indians. Eur J Clin Nutr 2006;60:791-801.
91. Rao S, Kanade AN, Yajnik CS, Fall CHD. Seasonality in maternal intake and activity influence offspring's birth size among rural Indian mothers--Pune Maternal Nutrition Study. Int J Epidemiol 2009;38:1094-103.
92. Rao S, Kanade A, Margetts BM, et al. Maternal activity in relation to birth size in rural India. The Pune Maternal Nutrition Study. Eur J Clin Nutr 2003;57:531-42.
93. Alderman H, Behrman JR. Reducing the Incidence of Low Birth Weight in Low-Income Countries Has Substantial Economic Benefits. The World Bank Research Observer 2006;21:25-48.
94. Indian National Family Health Survey. In. New Delhi: Indian Department of Family Welfare; 1999.
95. Osborne C. Statistical Calibration: A Review. International Statistical Review 1991;59:309-36.
96. Schafer JL. Multiple imputation: a primer. Stat Methods Med Res 1999;8:3-15.
97. Parker JD, Klebanoff MA. Invited commentary: Crossing curves--it's time to focus on gestational age-specific mortality. Am J Epidemiol 2009;169:798-801.
98. Klebanoff MA, Schoendorf KC. Invited commentary: what's so bad about curves crossing anyway? Am J Epidemiol 2004;160:211-2; discussion 5-6.
99. Bjerkedal T, Czeizel A, Hosmer DW, Jr. Birthweight of single livebirths and weight specific early neonatal mortality in Hungary and Norway. Paediatr Perinat Epidemiol 1989;3:29-40.
100. Hollingsworth MJ. Observations on the Birth Weights and Survival of African Babies: Single Births. Ann Hum Genet 1965;28:291-300.
101. Hughes K. Comparison of birthweight and infant mortality between Singapore and England and Wales, 1980. J Epidemiol Community Health 1985;39:135-40.
102. Terrenato L, Gravina MF, Ulizzi L. Natural selection associated with birth weight. I. Selection intensity and selective deaths from birth to one month of life. Ann Hum Genet 1981;45:55-63.
103. Olsen SF, Olsen J. A birth weight adjusted comparison of perinatal mortality in the Faroe Islands and Denmark. Scand J Soc Med 1994;22:219-24.
104. Birth Cohort Linked Birth - Infant Death Data Files. In: (U.S.) NCfHS, ed. Hyattsville, MD: US Vital Statistics Data Online; 2004.
105. 2004 Birth Cohort Linked Birth/Infant Death Data Set. In: Prevention CfDCa, Statistics NCfH, Statistics DoV, eds. Public Use Data File and Documentation: Department of Health and Human Services; 2009.
106. Whelan RM, Feiner D, Rekha C, et al. Adjusting birth-weight data from developing countries to account for heaping: A novel method. in preparation 2010.
107. Wilcox A, Russell I. Why small black infants have a lower mortality rate than small white infants: the case for population-specific standards for birth weight. J Pediatr 1990;116:7-10.
108. Buekens P, Wilcox A. Why do small twins have a lower mortality rate than small singletons? Am J Obstet Gynecol 1993;168:937-41.
109. Collins JW, Jr., David RJ. Differential survival rates among low-birth-weight black and white infants in a tertiary care hospital. Epidemiology 1990;1:16-20.
110. Lahariya C, Paul VK. Burden, differentials, and causes of child deaths in India. Indian J Pediatr 2010;77:1312-21.
111. Muthayya S. Maternal nutrition & low birth weight - what is really important? Indian J Med Res 2009;130:600-8.
112. Thompson JM, Wall C, Becroft DM, Robinson E, Wild CJ, Mitchell EA. Maternal dietary patterns in pregnancy and the association with small-for-gestational-age infants. Br J Nutr 2010;103:1665-73.
113. Hull HR, Dinger MK, Knehans AW, Thompson DM, Fields DA. Impact of maternal body mass index on neonate birthweight and body composition. Am J Obstet Gynecol 2008;198:416 e1-6.

114. Sen J, Roy A, Mondal N. Association of maternal nutritional status, body composition and socio-economic variables with low birth weight in India. J Trop Pediatr 2010;56:254-9.
115. Pathak R, Mustafa MD, Ahmed T, et al. Intra uterine growth retardation: Association with organochlorine pesticide residue levels and oxidative stress markers. Reprod Toxicol 2011.
116. Sahu MT, Das V, Mittal S, Agarwal A, Sahu M. Overt and subclinical thyroid dysfunction among Indian pregnant women and its effect on maternal and fetal outcome. Arch Gynecol Obstet 2010;281:215-20.
117. Ganesh Kumar S, Harsha Kumar HN, Jayaram S, Kotian MS. Determinants of low birth weight: a case control study in a district hospital in Karnataka. Indian J Pediatr 2010;77:87-9.
118. Kalaivani K. Prevalence & consequences of anaemia in pregnancy. Indian J Med Res 2009;130:627-33.
119. Whelan RM, Feiner D. Characteristics of Birth-Weight Distributions and the Birth-weight – Infant Mortality Relationship : A Comparison of India and the United States. In preparation 2011.
120. Rubin DB. Multiple Imputation After 18+ Years. Journal of the American Statistical Association 1996;91:473-89.
121. Kramer MS, Seguin L, Lydon J, Goulet L. Socio-economic disparities in pregnancy outcome: why do the poor fare so poorly? Paediatr Perinat Epidemiol 2000;14:194-210.
122. Chan MF, Ng WI, Van IK. Socioeconomic instability and the availability of health resources: their effects on infant mortality rates in Macau from 1957-2006. J Clin Nurs 2010;19:884-91.
123. Joseph KS, Liston RM, Dodds L, Dahlgren L, Allen AC. Socioeconomic status and perinatal outcomes in a setting with universal access to essential health care services. CMAJ 2007;177:583-90.
124. Maher J, Macfarlane A. Inequalities in infant mortality: trends by social class, registration status, mother's age and birthweight, England and Wales, 1976-2000. Health Stat Q 2004:14-22.
125. Arntzen A, Samuelsen SO, Bakketeig LS, Stoltenberg C. Socioeconomic status and risk of infant death. A population-based study of trends in Norway, 1967-1998. Int J Epidemiol 2004;33:279-88.
126. Haig D. Meditations on birth weight: is it better to reduce the variance or increase the mean? Epidemiology 2003;14:490-2.
127. United Nations Children's Fund and World Health Organization. *Low Birthweight: Country, regional and global estimates.* . New York; 2004.