2013•2014
FACULTY OF SCIENCES
Master of Statistics

Master's thesis
Determining the determinants of breastfeeding in South Africa

Promotor :
Prof. dr. Cristina SOTTO

Promotor :
Prof. KHANGELANI ZUMA

Olina Ngwenya
Thesis presented in fulfillment of the requirements for the degree of Master of Statistics

2013•2014
FACULTY OF SCIENCES
Master of Statistics

Master's thesis

Determining the determinants of breastfeeding in South Africa

Promotor :
Prof. dr. Cristina SOTTO

Promotor :
Prof. KHANGELANI ZUMA

Olina Ngwenya
Thesis presented in fulfillment of the requirements for the degree of Master of Statistics

universiteit hasselt | Maastricht University

Master of Statistics

Determining the Determinants of Breastfeeding in South Africa

Author: Olina Ngwenya

Supervisors: Prof. dr. Cristina Sotto
Dr. K. Zuma

10 September, 2014

Acknowledgements

First of all I would like to thank the Almighty God for giving me this opportunity to come and pursue a course of my dreams a Master of Science in Biostatistics. I extend my gratitude to my sponsor (VLIR-OUS) who made it possible for me to come here.

I would also like to thank my internal supervisor Professor Cristina Sotto and external supervisor Dr. Khangelani Zuma for their great support and guidance throughout the thesis write up period. Many thanks to HSRC for allowing me to use their data.

I would also like to thank Hasselt Censtat staff from whom I gained a lot of knowledge of Biostatistics. A big thank you goes to an ever smiling lady Ms Martine Machiels, she was ever ready to help students whenever need arises.

My acknowlegement goes to my fellow classmates, my group mates and above all to my good friend Anjullo Belay Belete for his assistance and contributions in all the difficult times that we endured together.

I wish to express my gratitude to my family and my family in Christ especially my God mother Sister Rihanna and Pastor Grace for your prayers. Faith Ndzimande and Josephine Shabani thank you so much for being such good friend. Theo thank you from the bottom of my heart for being there when I needed you most.

I dedicate this thesis to my lovely son Sijabuliso Hilton, Sonny thanks for the gift of time.

Abstract

Background:

World Health Organization global strategy of infant and young child feeding states that lack of breastfeeding, particularly exclusive breastfeeding (EBF), in the first six months, is a risk factor for the morbidity and mortality among children. However in South Africa, few infants are exclusively breastfed for the first six months of age. In this study we focus on the factors associated with exclusive breastfeeding.

Methods:

Data used is from a South African National HIV Prevalence, Incidence, Behaviour and Communication Survey conducted by Human Sciences Research Council in South Africa in 2008. The survey included a master sample of 1 000 enumeration areas (communities) and 15 households were selected per enumeration area. Within a household, at most 4 eligible individuals were selected. In this study we focus on children under 2 years and a guardian/parent (youth aged 15 to 24 years or adults aged 25 years or above) who participated in the survey. The pairs of a child and a guardian/parent added up to 1715. Logistic regression models are used to study the association between exclusive breastfeeding with child and mother characteristics. Analysis was done not accounting for survey design and later it was done accounting for the survey design where the univariate and multiple logistic regression models are fitted.

Results:

Race is significantly associated with EBF. The odds of being exclusively breastfed for Coloured children was found to be more than 2 times that of Indians.

Conclusion:

More awareness programmes promoting exclusive breastfeeding should be implemented. In addition, policy-makers must pay more attention on teaching the particular races which have low numbers of females excusively breastfeeding to do it and let them know why it is important to do so.

Keywords:

Exclusive breastfeeding, Logistic Regression, HIV, Survey data.

Acronyms and Abbreviations

WHO - World Health Orgarnization
HIV - Human Immunodeficiency Virus
EBF - Exclusive Breastfeeding
AIDS - Acquired Immune Deficiency Syndrome
PMTCT - Prevention of Mother to Child Transmission
EM - Expectation-Maximization
EMis - EM with Importance resampling
HSRC - Human Sciences Resources Council

Contents

List of Tables

List of Figures

1 Introduction

Exclusive breastfeeding (EBF) means that an infant receives only breast milk with no additional foods or liquids, not even water, with the exception of drops or syrups consisting of vitamins, mineral supplements, or medicines (WHO, 1991; Nutrition, 2013). If there is an introduction of any other food then the child is said to be mixed fed.

The advantages of breastfeeding are that breastmilk contains all nutrients and water a baby needs for the first six months of life. It reduces the baby's risk of getting diarrhoea, pneumonia and malnutrition allergies. It also bonds the mother and the child and saves money. Breastfeeding reduces the mother's risk of menopausal breast and ovarian cancer (Health, 2013). Also, it contributes to a delay in the return of fertility and helps protect women against anaemia by conserving iron.

Safe substitutes for exclusive breastfeeding can be assured in the developed world. Therefore, HIV-infected mothers, have always been advised not to breastfeed but to use formula feeds. However in resource-poor countries, which also harbour the greatest HIV/AIDS burden, the benefits of breastfeeding are far greater than the risks of formula feeding from a public health perspective (Ssenyonga et al., 2004). WHO recommends women to exclusively breastfeed their infants for up to 6 months (Indicators, 2008).

The reasons why some women introduce fluids or feeds other than breast milk include perceived insufficiency of milk and concerns of a crying baby. Beliefs and traditions surrounding feeding, and early technical difficulties, including cracked nipples, also play important roles in woman's confidence in, and ability to breastfeed exclusively (Ritchter et al., 1998; Guerrero et al., 1999)

It is more advantageous to exclusively breastfeed than to mix feed. EBF carries a significantly lower risk of mother to child transmission of HIV than mixed feeding (Coovadia et al., 2007; Ssenyonga et al., 2004). Also by exclusively breastfeeding, mothers avoid use of feeds potentially contaminated by pathogens due to the poor preparation and storage (Mortajemi, 1994; Iroegbu, 2000). Mixed feeding, especially with feeds prepared in unhygienic environments results in the disruption and damage to the epithelium thereby making HIV transmission more likely (Ssenyonga et al, 2004). On the other hand, EBF protects the integrity of the intestinal mucosa, which thereby presents a more effective barrier to HIV (Coovadia et al.,2007).

Determinants for breastfeeding are factors which influence mothers to breastfeed. It is important to understand these factors that influence EBF. Knowledge about the type of breastfeeding and important determinants is essential for increased initiation and extended duration. The determinants for breastfeeding can be categorized into different classes which include; demographic attributes, psycho-social attributes, health care related attributes, community attributes and public policy related factors (Yngve et al., 2001). Some of these determinants are modifiable but some are not. Awareness about the importance of these determinants for breastfeeding is vital for building effective promotion programmes.

1.1 Objectives of the Study

South African Health Ministry policy has also mandated the establishment and promotion of successful exclusive breastfeeding as part of the Prevention of Mother to Child Transmission (PMTCT) programme towards the reduction of infant mortality. Despite all the above stated advantages of EBF, EBF rates remain low in South Africa. Once these determinants are known, it would help the policy makers to improve awareness projects to make people aware of the importance of EBF. The aim of this study is to investigate factors associated with EBF in South Africa, a country with high HIV/AIDS burden. However, there is limited data regarding the determinants of EBF in South Africa.

1.2 Thesis overview

The thesis is structured as follows: The data description as well as the statistical methodology are described in Chapter 2. Results which include exploratory data analysis and results from the statistical methods are presented in Chapter 3. Discussion of the results, conclusions drawn from the analysis and limitatons of the study are in Chapter 4.

2 Materials and Methods

2.1 Data

The binary response variable was constructed in line with the definition of EBF by the World Health Organization, where EBF assumes value 1, if a child is exclusively breastfed and 0 if a child mixed fed. Many studies on breastfeeding practices have been hindered by methodological limitations, mainly as a result of lack of a uniform definition of breastfeeding forms. Therefore a call has been made for use of consistent and strict definitions in breastfeeding research (Sseny-onga et al., 2004).

Mixed feeding includes breastfeeding, formula feeding and solid intake. There were some children for which it was not known whether they breastfed or not. This led to missing values in the response variable. The determinants considered in the analysis were related to the mothers and children. In case of the mothers, we considered age, number of children, HIV test and PMTCT programme participation. Sex, racial group, genotype and HIV status of the child were considered. The children's characteristics were recorded as

- Gender: 1=male, 2=female
- Race: 1=African, 2=White, 3=Coloured, 4=Indian
- Geotype: 1=urban formal, 2=urban informal, 3=rural formal, 4=rural informal

while the mother's characteristics were recorded as

- Number of children a mother has: 1, 2-3, ≥ 4
- HIV test (whether the mother tested for HIV or not): 1=Yes, 2=No
- Participated in PMTCT programme: 1=Yes, 2=No
- Age: less than 18 years (young mothers), $18-35$ years (normal child bearing age mothers), greater than 35 years (old mothers).

2.2 Study design and Sample

The data that will be used for this study is from a cross-sectional population-based household survey (South Africa National HIV Prevalence, Incidence, Behaviour and Communication Survey (Shisana et al., 2008)), which was conducted using a multistage stratified sampling approach by the HSRC in 2008. The first survey of this nature was carried out in 2002 and then followed by another one in 2005. A third survey was done in 2008. This survey included individuals of all ages in South Africa. All persons living in selected households were eligible to participate, including those living in hostels, but persons staying in educational institutions, old-age homes, hospitals and uniformed-service barracks, as well as homeless people, were excluded from the survey.

The main idea behind sampling is to generalize results from part of the population to the whole population at large. A sample must be selected such that it fairly represents the entire population, to avoid bias. Simple random sampling is the most basic sampling method that minimizes sample bias by allowing elements in the whole population to have an equal chance of being selected. Sometimes due to the heterogeneity of the population it may not be appropriate to adopt simple random sampling alone but rather combine it with other sampling schemes e.g. multistage sampling (Chambers et al., 2003). Therefore in this survey a multistage disproportionate, stratified sampling approach was used. A total of 1000 census enumeration areas (EAs) from the 2001 population census were selected from a database of 86 000 EAs. The selection of EAs was stratified by province and locality type. The locality types were identified as urban formal, urban informal, rural formal (including commercial farms), and rural informal. Race was also used as a third stratification variable (based on the predominant race group in the selected EA at the time of the 2001 census), in the formal urban areas. The allocation of EAs to different stratification categories was disproportionate; that means, over-sampling or over-allocation of EAs was done, for example, in areas that were dominated by Indian, coloured or white race groups to ensure that the minimum required sample size in those smaller race groups was obtained (Sishana et al, 2009).

The selected 1000 EAs formed the primary sampling units and households were used as secondary sampling units. These households were selected systematically from each EA. Within each household, eligible individuals selected for the survey represented the ultimate sampling unit. With the view to obtain an approximately self-weighted sample of households (secondary sample units), the EAs were sampled with the probability proportional to the size of the EA using the 2001 census estimate of the number of households in the EA database. It can be seen from the map (Figure 1) that the highly populated cities, like Johannesburg and Durban, have more EAs than other places which are less populated. A random sample of 15 households was selected from each of the 1000 EAs, yielding a total sample size of 15 000 households (Shisana et al, 2008). Within each household, only one person within each age group was selected, subject to there being at least one eligible person in the specified age group. Four mutually exclusive age groups were used for sampling respondents: children under 2 years, children aged between 2-14 years, youth aged between 15-24 years and adults aged 25 years and above.

Owing to the sampling design of the survey, some individuals had a higher or lower probability of selection than others. To correct this problem, sample weights were introduced to correct for bias at the EA, household, and individual levels and also adjust for non-response. Information on the sampling weights can be found in Sishana et al (2009).

Since we are dealing with the determinants of breastfeeding, the data that concerns us is limited to the age groups for children under 2 years and a youth aged between 15-24 years or an adult aged 25 years or above. Dried blood spot specimens collected by heel-prick in infants were tested for HIV antibodies and a detailed questionnaire soliciting information related to knowledge, attitudes, practice, behaviours, and demographic factors was administered to the parent / guardian (Shisana et al., 2009). Among individuals who participated in the study, 1715 children under 2 years and their guardians/parents were involved.

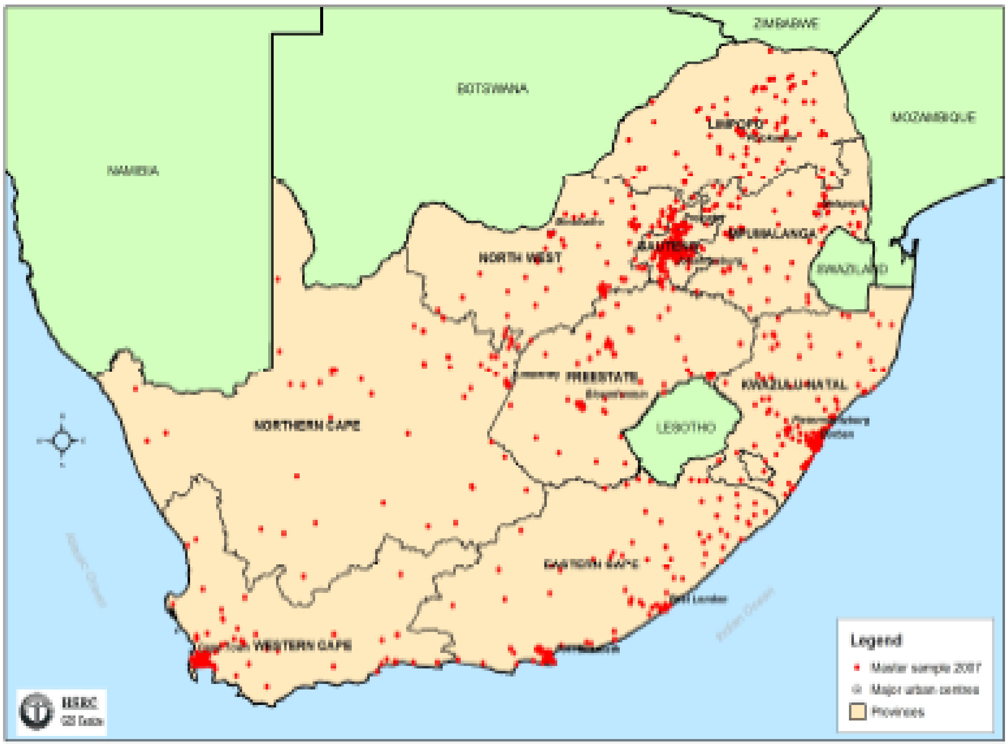

Figure 1: Master Sample for 2007

2.3 Methods

2.3.1 Logistic regression

Since we are dealing with a binary response variable, logistic regression was used to find out if there was association between EBF and the predictor variables: age, number of children, HIV test and PMTCT programme participation, sex of the child, geotype and the racial group of the child. Let $\pi(x)$ be the probability of exclusive breastfeeding for the given covariate. The model for $\pi(x) = P(Y = 1)$ at $\mathbf{X}$ where $\mathbf{X}$ takes the value x of the predictor variable is

$$logit[\pi(x)] = log(\frac{\pi(x)}{1-\pi(x)}) = \alpha + \beta x$$

2.3.2 Multiple logistic regression

Like ordinary regression, logistic regression extends to models with multiple explanatory variables. Let $\pi(x)$ be the probability of exclusive breastfeeding for the given vector of covariates. The model for $\pi(\mathbf{x}) = P(Y = 1)$ at values $\mathbf{X} = (x_1, x_2, ..., x_p)$ of p predictors is

$$logit[\pi(\mathbf{x})] = log(\frac{\pi(x)}{1-\pi(x)})\alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p$$

The alternative formula, directly specifies $\pi(x)$, as

$$\pi(\mathbf{x}) = \frac{exp(\alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p)}{1 + \alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p}$$

The parameters β_i refer to the effect of x_i on the log odds that $Y = 1$ controlling for the other x_j. An explanatory variable can be qualitative, using dummy variables for categories (Agresti, 2002). This is the case for the explanatory variables in the current study.

Therefore factors associated with exclusive breastfeeding since childbirth were found through univariate analysis (logistic regression) and multiple logistic regression.

As discussed in section 2.2, sampling of children and guardians was done by using multistage sampling design, in which stratification was done at provincial level, as well as weighting was taken into account. Multiple logistic regression models were fitted with and without taking survey design into account.

2.3.3 Missing data and multiple imputation

Most surveys suffer from the problem of missing data. Common reasons for missing data include refusal to answer sensitive questions e.g. in this study it could be your child's HIV status, insufficient knowledge (in case a mother is not there to answer questions about her child) and dropout due to loss of contact with respondents in case of longitudinal surveys. Missing data complicates the statistical analysis of clinical trials, surveys, etc, because the analysis methods must make assumptions about the missingness mechanism to deal with incomplete cases (Bohdana et al, 2011). The assumptions may be that the missingness is Missing Completely at Random (MCAR), Missing at Random (MAR), or Missing Not at Random (MNAR).

MCAR means that missingness is independent of both the observed and unobserved data, i.e. the probability of an observation being missing is independent of the response. Under MCAR various statistical methods, i.e. the frequentist, likelihood and Bayesian procedures, can be performed ignoring the process that generated the missing values (Molenberghs et al, 2005). Although MCAR is convenient given the simplicity of ignoring the missingness, it is not applicable in many situations. MAR means that the missingness depends only on observed responses. This implies that the probability of a missing observation is conditionally independent of the unobserved data given the values of the observed data and covariates (Molenberghs et al, 2005). Analysis based on the direct likelihood are valid under this mechanism. MNAR is used to describe missingness that depends on both observed and non-observed responses.

Multiple imputation is a technique that replaces each missing value with two or more acceptable values representing a distribution of possibilities (Rubin, 1987). There are several imputation methods which one can use, for instance observed values - information on the same respondent can be used (e.g. last observation carried forward), also the information can be borrowed from other subjects (e.g. mean imputation). The ultimate goal in using these methods is to reduce bias in survey estimates arising from missing data as well as to allow analysis to be conducted as if the dataset were complete, thus making analysis easier to conduct and results easier to interpret (Kalton, 1982).

The principle underlying multiple imputation is to fill up each missing value, say M times (usually 3-5 times) (Rubin, 1987). The M imputations for each missing value therefore give rise to M complete data sets, each of which is analyzed using standard complete data procedures as if the imputed data were real data obtained from the nonrespondents. The results from the M analyses are then averaged across the M samples to produce a single inference. The three commonly used methods of imputing are regression method, propensity score method and markov chain monte carlo method (MCMC). Multiple imputation is valid under the missing at random (MAR) assumption. It is useful to use next to direct likelihood.

The advantages of multiple imputation are that replication of imputations which is done in attempting to represent the distribution of the missing data increases the efficiency of estimation, i.e. good estimates of standard error are obtained because uncertainty resulting from filling in missing values is taken into account. Also as a result of replication, multiple imputations introduce random error into the imputation process thereby making it possible to obtain nearly unbiased estimates of all parameters.

Imputation has some desirable properties but it also has some drawbacks. Imputation methods do not necessarily lead to estimates that are less biased than those obtained from an incomplete dataset. In fact biases could be even much greater (Kalton, 1982). It is also stated that the performance of imputation techniques is ureliable since it is difficult to distinguish situations where they work and those where they prove misleading (Verbeke et al, 2000).

Rather than imputing missing values or using a complete case analysis, direct likelihood analysis can be employed (Molenberghs et al, 2007). With this approach all collected data are used without discarding or imputing any missing value.

In this study, statistical package Amelia II in R was used to generate M=5 filled-in datasets. Amelia II (Honaker et al, 2011) utilizes a bootstrapping-based EM (Expectation-Maximization) algorithm (e.g. EMis (King et al, 2001)) that is both fast and robust. EMis (EM with importance resampling) is a computational algorithm which gives draws from posterior distribution and it does not have convergence or independence difficulties.

Amelia II includes features for imputing cross-sectional surveys, time series data, and time-series/cross-sectional data. In this study, Amelia II performed the imputation step. Separate analyses and combination of results was undertaken in SAS. We assumed that the data are MAR. Multiple imputation was used as a supporting analysis to check how the results differ under survey design corrected analysis without imputation and the survey design corrected analysis with multiply-imputed data.

The variables in the model are assumed to be jointly multivariate normal. Assuming such is obviously an approximation, as few data sets have variables that are all continuous and unbounded. Fortunately, many researchers have found that it works as well as more complicated alternatives specially designed for categorical or mixed data (King et al, 2001). Therefore it is not a bad idea to assume the multivariate normality assumption.

3 Results

In this section different analyses are presented. Exploratory data analysis (EDA) is presented first, followed by results from statistical modelling that were obtained by applying statistical methods mentioned in Chapter 2.

3.1 Exploratory Data Analysis

Exploratory data analysis was done, to get some insights about the data. The data used in this study comprised of 1715 children under 2 years and their parents/guardians who participated in the survey. Out of these 1715 children, 110 had missing EBF observations. 1033 (60.2%) children exclusively breastfed, while 572 (33.4%) children did not exclusively breastfeed. Table 8 in the Appendix gives the frequencies of the attributes and the frequencies of their missing values. 40 children were HIV positive, while 726 children were negative and 949 children had a missing HIV status in the data set. But it was not known if they were negative in the beginning/after birth, therefore HIV status of the child will not be included in the analysis.

68 health cards for the children had PMTCT coding implying that the mothers were HIV positive, and 1003 children health cards had no PMTCT coding. There were 644 missing values on the PMTCT coding. 1188 mothers answered yes to having tested for HIV during pregnancy and 66 did not test. 461 had missing values for the HIVTEST variable. Mother's age was categorized into < 18 years representing young mothers, 18-35 years representing normal child bearing age mothers and > 35 years for old mothers. About 5% (65) of the interviewed mothers were young mothers, about 80% (1115) were normal child bearing age mothers and 15% (210) were old mothers. There were 325 missing values on the age of the mother variable. Twenty one mothers had no child before, 602 mothers had 1 child besides the new born, 705 mothers had 2-3 children and 261 mothers had more than 3 children. There were 126 missing values in the number of children variable.

The child and the mother attributes for those who exclusively breastfed and those who mixed fed are compared in Table 1. As for the child's attributes, geotype and child's race were significantly different for the two groups but sex of the child was not significant. The mother's characteristics (HIV Test, PMTCT, age of the mother, number of children) were not significantly different for the two groups (exclusively breastfeeding group and the mixed feeding group).

Table 1: *Child and Mother attributes in the study*

	Not EBF(572)	EBF(1033)	p value
Child's Attributes			
Sex			
Male	281 (49.13%)	538 (52.08%)	0.2566
Female	291(50.87%)	495(47.92%)	
Race			
African	396(69.35%)	748(72.90%)	0.0005
White	40(7.01%)	30(2.92%)	
Coloured	90(15.76%)	186(18.13%)	
Indian	45(7.88%)	62(6.04%)	
Geotype			
Urban Formal	411(73.79%)	692(68.38%)	0.0044
Urban Informal	30(5.39%)	95(9.39%)	
Rural Formal	100(17.95%)	173(17.09%)	
Rural Informal	16(2.87%)	52(5.14%)	
Mother's Attributes			
HIV Test			
Yes	385(94.36%)	798(94.89%)	0.6977
No	23 (5.64%)	43(5.11%)	
PMTCT			
Yes	31(7.85%)	37(5.50%)	0.1288
No	364(92.15%)	636(94.50%)	
Age of the mother			
less than 18	18(4.04%)	41(4.57%)	0.6536
18-35	366(82.06%)	718(79.96%)	
greater than 35	62(13.90%)	139(15.48%)	
Number of children			
1	230(48.63%)	372(44.77%)	0.4028
2-3	152(32.14%)	289(34.78%)	
4+	91(19.24%)	170(20.46%)	

Figure 2 show the descriptive statistics of the attributes of the children grouped according to the type of breastfeeding. These statistics vary among different attributes indicated in the descriptive statistics. It can be seen that for all children attributes, there is generally a high proportion of children who exclusively breastfeed as compared to those who did not exclusively breastfeed.

3.1.1 Exploring missingness

Table 7 in the Appendix shows missingness patterns together with the frequencies of patients exhibiting them. It is calculated that 41.28% of the patients have complete data whilst 12.48% are drop-outs. About 46.24% patients had missingness of the intermittent type. This shows that missingness is largely due to the intermittent type.

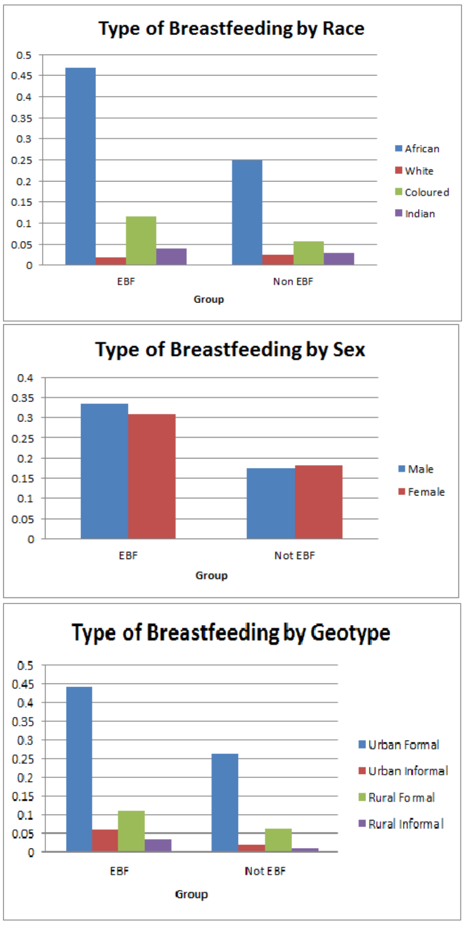

Figure 2: *Type of breastfeeding by children's characteristics*

3.2 Statistical Analysis

The data used in the analysis had missing values and SAS discards the observations with missing EBF values. Therefore instead of using data for 1715 children and their guardians, data for 1605 children and guardians was used in the analysis if the covariate had no missing values. For a case where the covariates had missing values, more lesser data was used. For multiply imputed data the whole data set with data from 1715 children and their guardians together with imputed missing data was used. Logistic regression was used to find out if there was association between EBF and the predictor variables: age, number of children, attendance, HIV test and PMTCT programme participation, sex of the child, geotype and the racial group of the child. Then multiple logistic was done not taking survey design and taking survey design into account. The interactions between attributes were found not to be significant.

Table 2: *Logistic Regression, Not Taking Survey Design into Account: Parameter Estimates for the Predictors of EBF*

Parameter	Estimate	Std error	P-value
Intercept	0.5004	0.6455	0.4382
Sex	-0.1086	0.1638	0.5073
Urban Formal vs Rural Informal	0.0416	0.3598	0.9080
Urban Informal vs Rural Informal	-0.2514	0.4076	0.5374
Rural Formal vs Rural Informa	0.1351	0.3760	0.7194
Age:less than 18 vs greater than 35	0.5379	0.4714	0.2539
Age: 18-35 vs greater than 35	0.2680	0.2898	0.3550
African vs Indian	0.5528	0.3212	0.0852
White vs Indian	-0.9069	0.6049	0.1338
Coloured vs Indian	0.6041	0.3373	0.0343
HIVTEST	0.2972	0.3959	0.4528
PMTCT	-0.6622	0.2991	0.0268
Number of children 1 vs 4+	-1.1020	0.3162	0.0005
Number of children 2-3 vs 4+	-0.7119	0.3063	0.0201

3.2.1 Logistic Regression not taking survey design into account

To obtain etimates of the odds ratios of the covariates, univariate logistic regression was fit to all children's and mother's characteristics, followed by a multiple logistic regression model. Hosmer and Lemeshow recommendation to include any covariate in the multivariate analyses which had p-value less than 0.25 in the univariate analysis (Agresti, 2002). A characteristic significantly associated with exclusive breastfeeding, was seen from the p-values less than 0.05 and the 95% confidence interval of the odds ratios excluding one.

Table 3 presents association of variables with EBF. In the univariate analysis, race of the child (p-value=0.0008) and geotype (p-value=0.0051) were significantly associated with exclusive breastfeeding. Children from areas classified as rural informal had higher estimated odds of being exclusively breastfed as compared to children from urban formal and rural formal. Whereas children from rural informal areas had similar estimated odds of being breastfed to children from urban informal (OR: 0.974; CI: 0.486-1.952). The odds of being exclusively breastfed for a White child are at least 47% less than of an Indian child (OR:0.529; CI:0.364-0.770), whereas African children (OR:1.333; CI:1.104-1.609) and Coloured children (OR:1.458; CI: 1.142-1.862) had higher odds of being exclusively breastfed as compared to the Indian children. The odds of being exclusively breastfed for boy children are at least 13% higher than girl children, however this effect is not statistically significant. All the mother's characteristics had an insignificant effect on EBF, in the univariate analysis, as seen from Table 3.

A multiple logistic regression model including any covariate which had p-value less than 0.25 in the unaivarite analysis was fit and the significant results are presented on the right hand side of Table 3. Table 2 shows the parameter estimates found from the multiple regression model not taking survey design into

Table 3: *Logistic Regression, Not Taking Survey Design into Account: Odds Ratio Estimates for the Predictors of EBF*

	Univariate		Multiple	
Mothers' characteristics	OR(95% CI)	P-value	OR(95% CI)	P-value
HIVTEST (Yes vs No)	1.109(0.659-1.866)	0.6978		
PMTCT	0.683(0.417-1.120)	0.1307	0.516(0.287-0.927)	0.0268
Number of children				
1 vs 4+	0.866(0.639-1.172)	0.1924	0.333(0.179-0.617)	0.0005
2-3 vs 4+	1.024(0.760-1.380)	0.4786	0.491(0.269-0.894)	0.0201
Age of the mother				
less than 18 vs greater than 35	1.016 (0.541-1.907)	0.7792		
18-35 vs greater than 35	0.875(0.632-1.211)	0.4137		
Children's characteristics				
Sex(Male vs Female)	1.126(0.917-1.381)	0.2567		
Race				
African vs Indian	1.333(1.104-1.609)	0.0028	1.738(0.926-3.261)	0.0852
White vs Indian	0.529(0.364-0.770)	0.0009	0.40(0.123-1.32)	0.1338
Coloured vs Indian	1.458(1.142-1.862)	0.0025	1.830 (1.145-3.543)	0.0343
Geotype				
Urban Formal vs Rural Informal	0.518(0.292-0.919)	0.0017		
Urban Informal vs Rural Informal	0.974(0.486-1.952)	0.0858		
Rural Fomal vs Rural Informal	0.532(0.289-0.982)	0.0204		

account. On multiple regression analysis, PMTCT (p-value=0.0268), number of children a mother had (p-value=0.0014) and race of the child (p-value=0.0216) had a significant association with EBF. PMTCT and the number of children a mother had were not significant in the univariate analysis, but after adjusting for various characteristics they became significant. Geotype was significant under the univariate analysis but turned out to be insignificant in the multiple logistic regression. Race was significant in both univariate and multiple logistic regression. The odds of being exclusively breastfed for Coloured children is almost 2 times that of Indians, holding other characteristics constant. The odds of exclusive breastfeeding for mothers who attended PMTCT are at least 48% less than those who did not attend PMTCT, while holding other characteristics constant. Mothers with 4 or more children had higher odds to exclusively breastfeed than mothers with 2-3 or 1 child, holding other characteristics constant.

3.2.2 Logistic Regression taking survey design into account

Table 4: *Multiple Logistic Regression, Taking Survey Design into Account: Parameter Estimates for the Predictors of EBF*

Parameter	Estimate	Std error	P-value
Intercept	0.5972	0.9184	0.5155
Sex	-0.0369	0.2323	0.8738
Urban Formal vs Rural Informal	0.1404	0.4200	0.7382
Urban Informal vs Rural Informal	-0.2825	0.4618	0.5407
Rural Fomal vs Rural Informal	0.2155	0.4293	0.6157
Age: less than 18 vs greater than 35	0.4181	0.5960	0.4829
Age: 18-35 vs greater than 35	-0.3987	0.3911	0.3080
African vs Indian	0.7142	0.5111	0.1623
White vs Indian	-1.6267	0.9089	0.0735
Coloured vs Indian	0.7719	0.3711	0.0297
HIVTEST	0.3685	0.5277	0.4850
PMTCT	-0.4810	0.4421	0.2766
Number of children 1 vs 4+	-0.6365	0.4213	0.1262
Number of children 2-3 vs 4+	-0.4865	0.4078	0.2329

When the survey design is taken into account we found out that the odds ratio parameter estimates change (either become smaller or larger) and the standard errors become larger, this can be seen from Table 4 compared to Table 2. In the univariate analysis, the age of the mother (p-value=0.0420) and race (p-value=0.0095) are significantly associated with exclusive breastfeeding. But when the survey design was not taken into account race of the child and geotype were significantly associated with breastfeeding. So there are now changes in what is significant or not.

The odds of exclusive breastfeeding for normal aged mothers (18-35 years) (OR:0.598; CI:0.380-0.940) are at least 40% less than the old aged mothers. The odds of being exclusively breastfed for Coloured children is more than 2 times that of Indians.

In the multiple logistic regression analysis, where we are now adjusting for various characteristics, only race of the child was significantly associated with EBF as seen in Table 5. The odds of being exclusively breastfed for Coloured children (OR: 2.164 CI:1.146-4.482) is more than 2 times that of Indians holding other characteristics constant.

Table 5: *Logistic Regression, Taking Survey Design into Account: Odds Ratio Estimates for the Predictors of EBF*

	Univariate		Multiple	
Mothers' characteristics	OR(95% CI)	P-value	OR(95% CI)	P-value
HIVTEST (Yes vs No)	0.975(0.476-2.001)	0.9460		
PMTCT	0.839(0.406-1.734)	0.6362		
Number of children				
1 vs 4+	0.880(0.571-1.355)	0.5614		
2-3 vs 4+	1.374(0.875-2.157)	0.1672		
Age of the mother				
less than 18 vs greater than 35	1.059 (0.481-2.332)	0.8897		
18-35 vs greater 35	0.598(0.380-0.940)	0.0375		
Children's characteristics				
Sex(Male vs Female)	1.215(0.897-1.645)	0.2086		
Race				
African vs Indian	1.717(0.906-3.252)	0.0975	2.043(0.750-5.561)	0.1623
White vs Indian	0.602(0.232-1.561)	0.2966	0.197(0.033-1.167)	0.0735
Coloured vs Indian	2.194(1.075-4.479)	0.0308	2.164 (1.146-4.482)	0.0397
Geotype				
Urban Formal vs Rural Informal	0.896(0.543-1.479)	0.6687		
Urban Informal vs Rural Informal	0.917(0.528-1.591)	0.7575		
Rural Fomal vs Rural Informal	1.175(0.715-1.931)	0.5251		

3.2.3 Logistic Regression taking survey design into account applied onto multiple imputed data

As indicated in EDA, almost all the variables had missing data, multiple imputation was considered. The incomplete data was imputed 5 times (M=5). Survey design corrected analysis was applied onto these data sets.
After multiple imputation of the missing values, logistic regression was done taking survey design into account, we found the above results. Table 6 shows odds ratio estimates based on design corrected analysis done after imputation of missing values. Race was significantly associated with EBF at 5% level of significance, both on the univariate analysis (p=0.0052) and multiple logistic regression (p=0.0033) analysis. The odds of being exclusively breastfed for Coloured children (OR: 2.216 ; CI: 1.188-4.513) is more than 2 times that of Indians after controlling for other variables.

These results are similar to the ones produced by the survey design corrected analysis without imputation of missing data, this however gives more confidence in our results obtained from available case analysis (survey design corrected analysis without imputation of missing data).

Table 6: *Logistic Regression, Taking Survey Design into Account, Multiply Imputed Data: Odds Ratio Estimates for the Predictors of EBF*

	Univariate		Multiple	
Mothers' characteristics	OR(95% CI)	P-value	OR(95% CI)	P-value
HIVTEST (Yes vs No)	1.192 (0.816-1.741)	0.3636		
PMTCT	0.944 (0.622-1.433)	0.7870		
Number of children				
1 vs 4+	0.855(0.603-1.212)	0.4536		
2-3 vs 4+	1.231(0.857-1.768)	0.2204		
Age of the mother				
less than 18 vs greater than 35	1.056(0.678-1.646)	0.5412		
18-35 vs greater than 35	1.028(0.737-1.435)	0.8361		
Children's characteristics				
Sex(Male vs Female)	1.218 (0.902-1.643)	0.1977		
Race				
African vs Indian	1.736(0.938-3.212)	0.1377	1.642(0.823-3.277)	0.1594
White vs Indian	0.623(0.241-1.611)	0.2372	0.626(0.216-1.811)	0.1290
Coloured vs Indian	2.214(1.130-4.340)	0.0445	2.216 (1.188-4.513)	0.0436
Geotype				
Urban Formal vs Rural Informal	0.909(0.556-1.488)	0.7092		
Urban Informal vs Rural Informal	0.932(0.541-1.606)	0.8003		
Rural Fomal vs Rural Informal	1.181(0.725-1.923)	0.5102		

.

4 Discussion and Conclusion

This study was aimed at studying the factors that are associated with EBF in South Africa. This was done using data from the South African National HIV Prevalence, Incidence, Behaviour and Communication Survey conducted by Human Sciences Research Council in South Africa in 2008 (Shisana et al., 2009). The analysis was based on children under 2 years and youth aged 15 to 24 years or adults aged 25 years or above. The outcome variable was a binary variable exclusive breastfeeding of the child in the first 6 months.

In the first part of the analysis a univariate logistic analysis and multiple logistic regression were done, not taking the sampling design into account. In addition, both univariate logistic regression and multiple logistic regression were done, taking survey design into account with stratification done at provincial level, as well as weighting. Then lastly multiple imputation of missing values was done followed by design-corrected univariate and multiple logistic regression.

From the two approaches it was observed that it is important to take the survey design into account since if ignored there is a higher probability of obtaining biased results.

On the univariate analysis which did not take sampling design into account, race of the child and geotype were significantly associated with EBF. When the sampling design was taken into account, the results of the univariate logistic regression changed and then age of the mother and race of the child were significantly associated with EBF.

On the multiple logistic regression analysis, when sampling design was not taken into account, variables PMTCT, race of the child and the number of children a mother had turned out to have a significant association with EBF. When sampling design was taken into account, only race emerged to have a significant association with EBF, while taking other characteristics to be constant. It appeared as if correcting for the survey design we pay the price in terms of efficiency. This reflected by the larger standard errors in the design-corrected analysis as compared to the analysis where survey design was not taken into account. Loss of precision here is not a problem as we are doing the correct analysis that will produce unbiased estimates.

Therefore from the survey design corrected analysis, we concluded that the odds of being exclusively breastfed for Coloured children (OR: 2.164; CI:1.146-4.482) are more than 2 times that of Indians, while holding other characteristics constant.

Due to the fact that there was some missingness in the variables involved in the analysis, we decided to do multiple imputation and then a survey design corrected analysis followed and we came to the same conclusion that the odds of being exclusively breastfed for Coloured children (OR: 2.216 ; CI: 1.188-4.513) are more than 2 times that of Indians, while holding other characteristics constant.

Therefore we concluded that race is a determinant of exclusive breastfeeding in this current study. The other child characteristics like mother HIVTEST and others were not significantly associated with EBF. Therefore infant feeding services in South Africa, should target particular races (Whites, Africans and Indians) which tend to exclusively breastfeed less as a way of promoting EBF.

In the study by Ssenyonga et al. (2004), they found out that gender, delivery at a government health unit, type of delivery and child age group greater than 3 months were significantly associated with EBF. For the study by Bland et al. (2002), the only factor associated with EBF from 0 to 6 weeks was delivery in the district hospital rather than in clinics, facilities outside the district or at home.

It was noted that in rural areas of South Africa more than 75% of poor households have no access to piped water or sanitation (Bland et al, 2002). Even if formula feed were provided freely, in such areas with the highest prevalences of HIV in the world, women do not have the facilities to prepare feeds safely. Currently, the most common mode of feeding is mixed breastfeeding, which is associated with increased mortality and morbidity and the highest rates of MTCT (mother to child transmission) (Bland et al, 2002). Therefore HIV services in South Africa, and all other developing countries formulating policies concerning HIV and infant feeding must promote EBF.

Generalizing these results may not be possible, but the study offers us some insights about EBF.

There are some practical difficulties which makes it hard for mothers to follow the policy of exclusive breastfeeding. Some mothers do not get the right information and support prenatally, during pregnancy and postnatally. Therefore the Health system should always try to deliver the important information to mothers at child bearing age.

Also the maternity leave of three months helps but in the last three months the mothers have to go back to work and in some cases some mothers don't benefit from maternity leave. Therefore in such cases it becomes hard for a baby to be exclusively breastfeed for 6 months and to solve this issue it may be better for maternity leave to go up to 6 months.

The limitations of the study are that; it was found out that demographic and health surveys usually lead to an overestimation of EBF (Ssenyonga et al, 2004). Therefore a better way of getting good information can be through following up mothers when they visit the hospital for child immunization and family planning like in the studes by Ssenyonga et al (2004) and Bland et al (2002). It is a better way because there are inclusion criteria, for example if the person bringing a kid to the hospital is not the mother of the child, the person is not enrolled in the study. This is a great advantage in the sense that you would not end up getting wrong answers or a lot of missing data because the person does not know some issues about the child and the mother. In a study like that of by Ssenyonga et al (2004) and Bland et al (2002), it becomes a bit easy to get information about certain mother aspects for example, mode of delivery (whether the child was delivered by Caesarean section or normal section) and type of clinic delivered in, etc.

In the current study, at times the mother of the child was not there to answer questions and any guardian could answer the questions thereby leading to less information at times or wrong answers. Therefore to gather better information to determine determinants for exclusive breastfeeding, it advisable to talk directly to the mothers only.

The HIV status of the child was checked at the time of the survey but we do not know if the mothers knew the child's HIV status before the study, so we could not include it in the analysis, since for a variable to be a determinant it has to be known before hand.

5 References

Agresti, Alan., (2002). *Categorical Data Analysis.* John Wiley & Sons, Inc., Hoboken, New Jersey.

Bland, R M., Rollins N K., Coutsoudis A., and Coovardia H M. (2002). Breastfeeding practices in an area of high HIV prevalence in rural South Africa. *Acta Paediatrica*; Vol. 91, Issue 6 pg 704–711.

Bohdana, R., and Michael, O. (2011). Implementation of Pattern-Mixture Models Using standard SAS/STAT Procedures. *PharmaSU2011-Paper SP04.*

Chambers, R L., and Skinner, C J. (2003). *Analysis of Survey Data.* New York: Wiley and Sons, Inc.

Coovadia, H M.,Rollins,N C., Bland ,R M., Little, K., Coutsoudis, A., Bennish, M L.,and Newell, M. (2007). Mother-to-child transmission of HIV-1 infection during exclusive breastfeeding in the first 6 months of life: an intervention cohort study, *The lancet* Vol 369.

Exclusive Breastfeeding.; Accessed on 16 September 2013; http://www.who.int/ nutrition/topics /exclusive_breastfeeding/en / (Nutrition).

Guerrero, ML., Morrow RC., Calva, JJ., Ortega-Gallegos, H., Weller, SC., Ruiz-Palacios, GM., and Morrow, AL. (1999). Rapid ethnographic assessment of breastfeeding practices in periurban Mexico City. *Bull World Health Organ*, 77(4) pg 323-30.

Health.; Accessed on 16 September 2013; http://www.kznhealth.gov.za/ exclusivebreastfeeding.htm (Health).

Honaker, J., Gary K., and Matthew, B. (2011). Amelia II: A Program for Missing Data. *Journal of Statistical Software*, 45, no. 7 pg 1-47.

Iroegbu, CU., Ene Obong, HN., Uwaegbute, AC., and Amazigo, UV., (2000). Bacteriological quality of weaning food and drinking water given to children of market women in Nigeria: Implications for control of diarrhoea. *Journal of Health,Population and Nutition*, 18(3): pg 157-62.

Kalton, G. (1982). Accessed on 10 July 2014. Imputing for missing survey responses. Available on: http://www.amstat.org/sections/ srms/Proceedings/ papers/1982_004.pdf.

King, G., Honaker, J., Joseph, A., and Scheve, K. (2001). *Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation. American Political Science Review*, 95(1), 49(69).

Mortajemi, Y., Kaferstein, F., Moy, G., and Quevedo, F. (1994). *World Health Forum*, 15(1) pg 169-74.

Molenberghs, G. and Verberke, G. (2005). *Models for Discrete Longitudinal data.* New York: Springer.

Molenberghs, G. and Kenward, M. G. (2007). *Missing data in clinical studies.* New York: Wiley.

Ritchter, L. and Griesel, D. (1998). Breastfeeding and infant care in the context of HIV/AIDS. *Psychology in society (PINS)*, 24, pg 40-56.

Rubin, D. B. (1987). *Multiple imputation for Nonresponse in Surveys.* New York: Wiley.

Shisana, O., Rehle, T., Simbayi, L., Zuma, K., and Jooste, S.(2009). South African National Prevalence, Incidence, Behaviour and Communication Survey, 2008. A Turning Tide Among Teenagers?. Cape Town, South Africa: HSRC Press.

Ssenyonga, R., Muwonge, R., and Nankya, I. (2004). Towards a Better Understanding of Exclusive Breastfeeding in the Era of HIV/AIDS: A Study of Prevalence and Factors Associated with Exclusive Breastfeeding from Birth, in Rakai, Uganda, *Journal of Tropical Pediatrics*, Vol. 50, No.6, pg 348-353.

Verbeke, G. and Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data.* New York: Springer.
World Health Organization., Indicators for Assessing Breastfeeding Practices. WHO/CDD/SER/91.4 (WHO), Geneva, (1991).

World Health Organization., Indicators for assessing infant and young child feeding practices(2008). Accessed on 2 March 2014. http://www.who.int/maternal_child_adolescent /documents/9789241596664/en/ (Indicators).

Yngve, A., and Sjostrom, M. (2001). Breastfeeding determinants and a suggested framework for action in Europe., *Public Health and Nutrition*, 4(2B), pg 729-739.

6 Appendix

Table 7: *Missingness patterns and their frequencies. 'X' indicates observed and '.' indicates missing)*

EBF	Sex	Geotypec	agec	racec	HIVTEST	PMTCT	Childrenc	Freq	Percent
X	X	X	X	X	X	X	X	708	41.28
X	X	X	X	X	X	X	.	138	8.05
X	X	X	X	X	X	.	X	322	18.78
X	X	X	X	X	X	.	.	67	3.91
X	X	X	X	X	.	X	X	53	3.09
X	X	X	X	X	.	X	.	12	0.70
X	X	X	X	X	.	.	X	27	1.57
X	X	X	X	X	.	.	.	9	0.52
X	X	X	X	.	X	X	X	3	0.17
X	X	X	X	.	X	X	.	1	0.06
X	X	X	X	.	X	.	X	2	0.12
X	X	X	X	.	X	.	.	1	0.06
X	X	X	X	.	.	X	X	1	0.06
X	X	X	.	X	X	X	X	2	0.12
X	X	X	.	X	X	.	X	4	0.23
X	X	X	.	X	X	.	.	1	0.06
X	X	X	.	X	.	X	X	112	6.53
X	X	X	.	X	.	X	.	38	2.22
X	X	X	.	X	.	.	X	70	4.08
X	X	X	.	X	.	.	.	34	1.98
.	X	X	X	X	X	X	X	1	0.06
.	X	X	X	X	X	X	.	1	0.06
.	X	X	X	X	X	.	X	3	0.17
.	X	X	X	X	.	.	X	18	1.05
.	X	X	X	X	.	.	.	22	1.28
.	X	X	X	.	.	.	X	1	0.06
.	X	X	.	X	.	X	.	1	0.06
.	X	X	.	X	.	.	X	3	0.17
.	X	X	.	X	.	.	.	60	3.50

Table 8: *Frequencies of the attributes and their missingness*

Attribute	Frequency (Percentage)
Child's Attributes	
Sex	
Male	864 (50.38%)
Female	851(49.62%)
Missing	0 (0%)
Race	
African	1203(70.15%)
White	81(4.72%)
Coloured	301(17.55%)
Indian	121 (7.06%)
Missing	9(0.52%)
Geotype	
Urban Formal	873(50.90%)
Urban Informal	232(13.53%)
Rural Formal	478(27.87%)
Rural Informal	132(7.70%)
Missing	0 (0%)
Mother's Attributes	
HIV Test	
Yes	1188(69.27%)
No	66(3.84%)
Missing	461 (26.88%)
PMTCT	
Yes	68(3.97%)
No	1003(58.48%)
Missing	644(37.55%)
Age of the mother	
less than 18	65(3.79%)
18-35	1115(65.01%)
greater than 35	210(12.24%)
Missing	325(18.95%)
Number of children	
1	612(35.69%)
2-3	453(26.41%)
4+	265(15.45%)
Missing	385(22.45%)

Auteursrechtelijke overeenkomst

Ik/wij verlenen het wereldwijde auteursrecht voor de ingediende eindverhandeling:
Determining the determinants of breastfeeding in South Africa

Richting: **Master of Statistics-Biostatistics**
Jaar: **2014**

in alle mogelijke mediaformaten, - bestaande en in de toekomst te ontwikkelen - , aan de Universiteit Hasselt.

Niet tegenstaand deze toekenning van het auteursrecht aan de Universiteit Hasselt behoud ik als auteur het recht om de eindverhandeling, - in zijn geheel of gedeeltelijk -, vrij te reproduceren, (her)publiceren of distribueren zonder de toelating te moeten verkrijgen van de Universiteit Hasselt.

Ik bevestig dat de eindverhandeling mijn origineel werk is, en dat ik het recht heb om de rechten te verlenen die in deze overeenkomst worden beschreven. Ik verklaar tevens dat de eindverhandeling, naar mijn weten, het auteursrecht van anderen niet overtreedt.

Ik verklaar tevens dat ik voor het materiaal in de eindverhandeling dat beschermd wordt door het auteursrecht, de nodige toelatingen heb verkregen zodat ik deze ook aan de Universiteit Hasselt kan overdragen en dat dit duidelijk in de tekst en inhoud van de eindverhandeling werd genotificeerd.

Universiteit Hasselt zal mij als auteur(s) van de eindverhandeling identificeren en zal geen wijzigingen aanbrengen aan de eindverhandeling, uitgezonderd deze toegelaten door deze overeenkomst.

Voor akkoord,

Ngwenya, Olina

Datum: **22/09/2014**