Synopsis
1. Name of the Student: Rajini Thakur Nagrani
2. Name of the Constituent Institution: Tata Memorial Centre
3. Enrolment No. : HLTH09200904001
4. Title of the Thesis: Risk Factors of Breast Cancer in Rural & Urban India
5. Board of Studies: Health Sciences
INTRODUCTION

In 2012, 1.7 million women globally were diagnosed with breast cancer and there were 6.3 million women alive who had been diagnosed with breast cancer in the previous five years [1]. In countries with high and medium Human Development Index (HDI) an increased risk in female breast cancer has been observed [2]. Further, the pattern of age adjusted incidence rates as observed in Cancer Incidence in Five Continents report clearly indicate that breast cancer rates are high in developed countries and are much lower in less developed countries including India [3].

Within India, there are likely substantial differences in the incidence rates of breast cancer in rural and urban areas: rates observed in urban registries are in the range of 29 – 35 per 100,000 whereas those observed in rural registries vary from 10-12 per 100,000. The lowest breast cancer incidence rates are found among women from the rural area of Barshi in Western India, and Dindigul Amblikkai, another rural area in the more developed South of India. Among urban Indian women, breast cancer incidence rates are almost three times higher than in rural women [4]. A twofold increased risk was observed in urban areas and a threefold increased risk was observed in metro areas compared to rural area [5]. The cause of this strong urban rural difference is not known although it is likely to be due to one or more westernised lifestyle related factors such as parity, breastfeeding, age at first birth and obesity, prevalence of which differs strongly between rural and urban women.

Nulliparity and late age at first birth are the most consistently observed risk factors for breast cancer [6]. The risk among women who have their first child after the age of 30 is about twice that of women who have a first child before the age of 20. Similarly, women who start menstruating early in life, or have a late menopause, also have an increased risk of developing breast cancer [7], possibly because of the increased number of ovulatory cycles and exposure
to estrogens and other breast tissue proliferative hormones. It is also possible that extensive breastfeeding reduces the risk of breast cancer by suppressing the number of ovulatory cycles, although the evidence based on studies conducted in western populations is unclear [8]. Overall however, these established risk factors account for only a small part of the large difference in incidence between developed and developing countries and other important risk factors for breast cancer remain to be identified.

Higher body mass index has been found to increase the risk of breast cancer after menopause, although this has not been observed in all cohort studies which have examined the association. Similarly, weight gain during adulthood has been identified as a risk factor for breast cancer in most studies in which it was investigated in post menopausal women. Some studies have also observed that the weight gain at the age of 20 years increases the overall breast cancer risk [9]. Physical activity has also been hypothesized to protect against the development of breast cancer [10]. Green et.al has shown individual height as an independent risk factor in breast cancer [11]. There have been large Genome Wide Association Studies (GWAS) on breast cancer in most developed countries [12-15] showing low to modest associations between common polymorphisms and breast cancer risk. In India, however, there have been no GWAS studies and few properly designed retrospective studies with smaller sample size on genetic susceptibility to study this risk [16-18].

**GAPS IN LITERATURE**

It has been observed for long time that the rates of breast cancer differ in rural and urban areas. However, there are very few studies in literature to address the reasons for the differences in the breast cancer rates of rural and urban area. Obesity has been observed to be risk factor for postmenopausal breast cancer. However the contribution of different measures of obesity and their role in pre- and postmenopausal women is still not clear. In Indian context, there are no
large studies to address the issue of reproductive factors, obesity, age at last pregnancy, oral contraceptive use and genetic susceptibility in development of breast cancer.

The present thesis proposal is designed to understand more clearly the reasons for rural-urban differences, and role of genetic susceptibility in development of breast cancer.

HYPOTHESIS

Anthropometric and Lifestyle related variables are the cause of large differences in occurrence of breast cancer in rural and urban areas.

AIM

Primary: To study role of anthropometric and other lifestyle related variables in causation of breast cancer in rural and urban areas.

Secondary: To study role of genetic susceptibility in breast cancer.

PRIMARY AIM

To study role of anthropometric and other lifestyle related variables in causation of breast cancer in rural and urban areas.

Study Population: A hospital based case-control study was conducted at Tata Memorial Hospital (TMH), Mumbai during the period of January 2009 to September 2013.

Criteria for enrolment of cases: The cases were female breast cancer patients coming to TMH. Only primary breast cancer cases aged 20-69 were enrolled in the study with date of diagnosis not more than 6 months before the date of interview. All the breast cancer cases enrolled in the study were histologically confirmed.

Criteria for enrolment of controls: All female visitors with no history of cancer coming along with any site cancer patient aged 20-69 were included in the study. Visitor controls coming to various Disease Management Group (DMGs) have been enrolled. Not more than 20% controls have been enrolled from any of the DMGs, to avoid selection bias.
The study has been approved by TMH Institutional Review Board. Written informed consent was obtained from all participants before enrolling them in the study.

**Data Collection:** In-person interview of each case and control was conducted by trained interviewers using a pre-tested structured questionnaire covering demographic and socioeconomic variables, reproductive history, time spent in household activities on a normal day, residential history, occupational history, personal and family medical history, tobacco and alcohol habits, and diet. Controls were frequency matched to cases on age and region of residence (South, North, East, West and Central India). Anthropometric measurements were taken at the end of the interview.

**Blood Collection:** A 10ml blood sample was collected from each study participant and centrifuged into plasma and buffy coat. The blood components were then stored at -80°C.

**Definition of Rural and Urban areas:** All study participants were asked to list all places of residence where they had lived for at least 1 year, starting with the place of birth. The rural and urban residence status was self reported by study participant. Study participants were stratified into rural and urban using four different definitions as follows:

1. **Ever lived in a rural area:** If a study participant has lived in a rural area for 1 year or more in life is termed as a “rural participant”, whereas any participant who has never lived or lived less than 1 year in a rural area is termed as “urban participant”.

2. **First 20 years of life lived in a rural area:** If a study participant has lived first 20 years of her life in a rural area, i.e., from age 0 to age 20, then participant is classified as “rural participant,” whereas any participant who has lived less than 20 years in a rural area is classified as “urban participant”.

3. **Currently living in a rural area:** Any study participant who has a current residence (at the time of enrolment) of 1 year or more in a rural area is termed as “rural participant”, versus a current residence in an urban area is an “urban participant”.

6
4. **Total years lived in a rural area:**
   
   a. **1-10 years:** A minimum of 1 year and a maximum of 10 years lived in a rural area versus never lived in a rural area are categorized as rural and urban participants respectively.
   
   b. **>10 years:** If total years lived in a rural area is more than 10 years, study participant is categorized as rural or else urban.

**Anthropometric Measurements:** Height (without shoes in cm) and weight in light clothing (in kg) of each study participant were measured using standard equipment. Weight was measured with light clothing. Waist size (in cm) was measured using a tape around the narrowest part of the trunk between the lower rib and level of the highest point of the margin of the hipbone, and hip size (in cm) was measured with light clothing at the widest part. All measurements were done twice in succession and averaged for a final value. Waist-to-hip ratio was computed by taking the ratio of waist size (in cm) and hip size (in cm).

**Definition of Menopausal status:** Women with no history of menstrual period during the last 12 months were classified as postmenopausal. The rest were treated as premenopausal.

**Quality Assessment for Questionnaire Based Data**

**Preparation of Instruction Manual for filling up the Questionnaire in Case-Control Studies:** In order to assure the homogeneity of data collection by the Social Investigators, an instruction manual and video recording was prepared. The instruction manual contains detailed guidelines and figures wherever required for better understanding of questions by the social investigator as well as the respondent [19].

**Preparation of Instruction Manual for Data Entry:** In order to assure the homogeneity while entering the data, clear and precise instructions with predefined logical checks have been listed in the form of Manual [20].
Monitoring of Daily Work: All forms were regularly checked for errors after conducting the interviews and after the data has been entered in the database. Weekly meetings were conducted to understand and resolve the problems of data collection. Training program was conducted every quarter so as to ensure the quality of interviews. The questionnaire was checked daily for completeness of information.

Quality Checks on Data Entry: Logical Checks were prepared to identify errors in the data entry. The data was entered twice and corrected for errors between 2 entries, if any, occurred while entering the data.

Reproducibility of Questionnaire: Abbreviated questionnaire was designed. This questionnaire contained constant (non changing in recent time) variables such as number of pregnancies, height, vegetarian /non-vegetarian status. The reproducibility questionnaire was completed for 249 study participants (approx 8% of total enrolled in study). Details of main measured exposures are shown in Table 1.

Table 1: Reproducibility of Measured Exposure

<table>
<thead>
<tr>
<th>Variable</th>
<th>Study Mean (Reproducibility Mean)</th>
<th>Coefficient of Correlation (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>46.90 (47.17)</td>
<td>92.25</td>
</tr>
<tr>
<td>Number of Pregnancies</td>
<td>4.06 (3.99)</td>
<td>91.07</td>
</tr>
<tr>
<td>Height</td>
<td>156.92 (157.18)</td>
<td>96.51</td>
</tr>
</tbody>
</table>

Correction of differences between Data entry 1 & 2: There were 207 variables which were corrected for any differences observed between the 2 data entries.

Statistical Analysis: The odds ratios (OR)s of developing breast cancer and their 95% confidence intervals (CI)s for anthropometric measurements and reproductive factors were estimated separately by residential status (Rural/Urban) and menopausal status. Unconditional logistic regression models were adjusted for potential confounders (age, region of residence, rural-urban status, education, age at first full-term pregnancy, waist-to-hip ratio (WHR), height, menopausal status, number of abortion and miscarriage). The ORs for age at first full-
term pregnancy were estimated after adjusting for number of full-term pregnancies in addition to above mentioned variables. Weight in kg as continuous variable was entered in the model for estimating OR for WHR and height in addition to above variables. All analysis were performed using the statistical package Stata version 12 [21]. Tests for linear trend across levels of exposure categories were performed on the continuous categorical variables entered as ordered, quantitative variables into the models.

**Result & Discussion:** Questionnaire data was collected on 1637 breast cancer cases and 1515 female controls. All the results were adjusted for the confounding variables unless mentioned otherwise.

A protective association was observed using all the definitions of “rural” [Ever lived in rural area – OR=0.81; 95% CI - 0.71-0.94); More than 10 years lived in rural area – OR=0.81; 95% CI - 0.70-0.93); first 20 years of life lived in rural area – OR=0.65; 95% CI - 0.56-0.76)] except in women who were currently residing in rural area at the time of enrolment after adjusting for confounding variables such as age and region of residence. On further adjustment for additional risk factors viz. age, region of residence, education, age at first full-term pregnancy, height, WHR and menopausal status; only women who lived first 20 years of life in rural area showed protection against breast cancer (OR=0.77; 95% CI - 0.65-0.92).

However, most of the etiological studies have used current area of residence as definition for rural [22-23] and limited studies which have taken early years of life spent [24] or place of birth [25-26] in rural areas as definition for “rural”. The current residence as demonstrated in the present study is not a good marker for studying the effect of rural environment on the risk of breast cancer, as exposures in early life may be more important in the development of breast cancer compared to current exposures [27]. For instance, strenuous physical activity at younger age can delay both menarche and onset of regular menstrual cycle [28]. Further, the individuals migrating from rural area to urban area in adulthood might not change their
lifestyle and adhere to rural life; therefore they may continue to get protection from breast cancer even if they are currently residing in urban areas which have been clearly demonstrated in the present study. Therefore, in all the further analysis women who lived first 20 years of life in rural area were designated as ‘rural participants’ else stratified as urban.

Prevalence of ER/PR negative (60.9%) cases was higher in rural area compared to urban area where the prevalence was observed to be 54.3%. A statistically significant difference (P = 0.018) in the prevalence of Triple Negative Breast Cancer (TNBC) tumours was observed in the rural area (44.2%) compared to urban area (34.3%).

Women who had 4 or more live births showed a protective association with OR = 0.66 (95% CI - 0.49-0.87) as compared to women with 1 live birth after adjusting for confounding variables and without stratification for rural-urban status. The significant protection was observed only in rural women (OR = 0.42; 95% CI-0.24-0.75). Age at first full-term pregnancy proved to be an important risk factor in the development of breast cancer. Women who had their first full-term pregnancy after age 25 had a significantly elevated risk of breast cancer compared with women who had first full-term pregnancy below 20 years of age (OR = 1.83; 95% CI-1.41-2.36). This protection was observed in both rural areas (OR = 2.24; 95% CI-1.13-4.43) and urban areas (OR = 1.78; 95% CI-1.32-2.41). A statistically significant linear trend was observed among the categories of age at first full-term pregnancy. The lifestyle patterns among women living in urban areas has changed considerably, with women attaining higher level of education, postponing marriage, postponing their first child to an older age, and having fewer pregnancies over time [29-30]. An indication has been observed in this study that use of OC may increase the risk of breast cancer particularly for women residing in urban area (OR = 1.28; 95% CI-0.93-1.76). Two or more than 2 induced abortions has been observed to be a risk factor of breast cancer overall (OR = 1.65; 95% CI-1.25-2.17), urban (OR = 1.58; 95% CI-1.15-2.16) and rural women (OR = 2.08; 95% CI-1.16-3.72). Even
a single miscarriage provides a protection against breast cancer in rural (OR = 0.62; 95% CI-0.41-0.95) and urban women (OR = 0.79; 95% CI-0.58-1.06) possibly due to its protection acquired by pregnancy. However the results observed for abortion and miscarriage has to be interpreted considering the possibility of recall bias, a limitation of case-control studies. A time difference of 10 years or more between age at menarche and age at first full-term pregnancy was observed to be significantly associated in urban women (OR = 1.36; 95% CI-1.11-1.68), but not in rural women (OR = 1.43; 95% CI-0.91-2.24).

Height has been consistently associated with an increase in breast cancer risk [11, 31]. In the present study, for every 5 cm increase in height the OR of 1.10 (95% CI-1.02-1.19) was observed in the urban area, but not in rural area (OR = 1.05; 95% CI-0.93-1.19). The increased risk of breast cancer for WHR of ≥0.95 when compared to WHR of <0.85 was observed to be OR = 3.78 (95% CI-2.92-4.89) without stratifying on rural-urban status; in urban women (OR = 4.07; 95% CI-3.00-5.53) and rural women (OR = 3.00; 95% CI-1.84-4.90). A significant positive association with WHR has been reported in both pre- and postmenopausal women, a result similar to that observed in two meta-analysis report [32-33].

SECONDARY AIM

To study role of genetic susceptibility in breast cancer.

DNA Preparation: Buffy coat samples were available for 1214 cases and 1293 controls. Genomic DNA was extracted from buffy coat using Qiagen QiAamp Blood DNA MidiKit and Macherey Nagel Nucleomag Blood kit. Concentration of each DNA sample was determined by the optical density (OD) at 260 nm and the purification was evaluated by OD 260/280 ratio. All DNA samples were also quantitated using Quant-iT PicoGreen dsDNA reagent, and purity was assessed by measuring the UV absorbance for accuracy. The quality of genomic DNA was assessed on 5% samples using 0.8% agarose gel. 1204 cases and 1212 controls had sufficient yield to proceed with genotyping. DNA concentrations were adjusted
to 50ng/µl and verified using Quant-iT PicoGreen dsDNA reagent. The aliquots of DNA were stored at -20°C.

**Design of Custom SNP Panel:** A customized panel of 384 single nucleotide polymorphisms (SNPs) was designed using a mixture of 3 strategies which are as follows

1. **Candidate SNP Studies:** All candidate SNP studies which have been significantly associated with breast cancer and suggested by collaborator on basis of animal experiments were included under this criterion using HuGE Navigator [34]. Total SNPs selected from this category were 96.

2. **GWAS:** The GWAS snps were identified using HuGE Navigator [35] and NIH GWAS Catalog [36]. The SNPs which were positively associated (p value < 10^{-5}) with following diseases or traits (the number of snps selected in the respective category mentioned in parenthesis) were included:

   - Body Mass Index (37)
   - Breast Cancer (51)
   - Insulin Like Growth Factors (1)
   - Menstruation and Menopause (41)
   - Obesity (29)
   - Waist-to-Hip Ratio (2)

   A total of 161 snps were identified using this strategy.

3. **Bioinformatics Tool:** 127 tag snps were selected using this strategy. Obesity search term was used in Gene evidence [37] tab of HuGE Navigator. Thirty three genes had a score of 0.05 or more which were uploaded on the Candidate gene SNP selection (Genepipe) pipeline of “SNPinfo” a web-based SNP selection tool [38]. The algorithm used for selecting SNPs is as follows: Five kb upstream and 1 kb downstream of the gene coordinate were included in the selection. Only SNPs showing a minor allele frequency (MAF) of 0.05 or greater were included. Tagging proportion cut-off to filter gene was kept 0.8 and LD threshold cut off was kept 0.8. Minimum number of snps tagged by a tag snp was kept as 3. In order to ensure that each gene has some coverage a minimum of 1 tag snp to a maximum of 100 tag snps per gene were included. Further SNPs were filtered using the functional SNP prediction in “Genepipe”
that cause an amino acid change, those that may alter the functional or structural properties of
the translated protein, disrupt transcription factor binding sites, disrupt splice sites or other
functional sites.

**Quality Assessment of Genotyping:** Genotyping was performed on the Illumina Hi-Scan
using GoldenGate Genotyping (GGGT) Custom SNP Panel assay (Illumina Inc., San Diego,
CA). GGGT assay was performed on 1204 cases and 1212 controls (Total: 2416) for 384
custom selected SNPs. Intraplate and interplate replicates (7% approx.) were included on all
plates and in all batches. Blinded duplicates were also included on all plates as another QC
measure. The reproducibility rate of all the replicate samples (n=160) for all the assays was
>98%. Also negative controls were run in some of the assays to check for any inter sample
contamination. After excluding 17 samples with call rate <90%, a total of 2399 samples were
included in final analysis. Further, 16 SNPs with diffused clusters, 6 SNPs with call frequency
<95%, 4 SNPs with MAF<1% and 6 SNPs with substantial deviation from Hardy-Weinberg
Equilibrium (p<0.001) were excluded to have a list of 352 SNPs for final analysis. All SNPs
had a Gen train score value of 0.4 and above leading to no exclusions of SNPs due to poor
cluster quality.

**Statistical Analysis:** A chi-square test was used to verify whether the observed genotype
frequencies were in Hardy-Weinberg equilibrium. Principal Component Analysis was
conducted to evaluate the potential effects of population structure between the samples. There
was no significant difference in eigenvector loadings for the first five factors showing that the
regional differences in structure were a minor source of population variability. Therefore, the
analysis were not conditioned on region. Unconditional logistic regression was used to
estimate OR and corresponding 95% CI between genotypes and case status. The genotypes
were coded as 0=wild type, 1=heterozygous and 2=homozygous variant. The models fitted
were additive (continuous effect of increasing number of variant alleles - 0 versus 1 versus 2),
dominant (0 versus 1 and 2), recessive (0 and 1 versus 2) and genotypic (0 versus 1, 0 versus 2). Positive associations were defined as an OR larger than 1, whereas an inverse association was specified by an OR below 1. To limit the probability of false-positives due to multiple testing, a false discovery rate method of Benjamini and Hochberg [39] was used to calculate $q$-value. A false discovery rate cut-off of 0.05 was applied to select the top SNPs, which limited the probability of false-positives due to multiple tests that were carried out. All the analysis were performed using the statistical software Stata version 12.0 and PLINK v1.07 [40-41].

**Results and Discussion:** Out of 384 SNPs genotyped a total of 32 SNPs were excluded from final analysis due to various reasons mentioned above. From 352 SNPs which were analysed 4 SNPs in FGFR2 gene using genotypic model (homozygous dominant v/s homozygous recessive) i.e. rs1219648, rs2420946, rs2981575 and rs2981582 showed positive association having OR 1.32 (1.02-1.70), 1.42 (1.10-1.82), 1.33 (1.04-1.70), 1.31 (1.02-1.68), 1.47 (1.11-1.94) respectively with breast cancer. FGFR2, fibroblast growth factor receptor 2, encodes a receptor tyrosine kinase that is amplified and over expressed in breast cancers. Polymorphisms in FGFR2 associated with breast cancer conferred a 20% increased risk of breast cancer among heterozygotes and a 60% increased risk among homozygotes with variant allele when compared to wild type homozygotes [42-43]. Recently a meta analysis has also observed a similar association of rs1219648, rs2420946 and rs2981582 in Caucasians and East Asians in ER+/PR+ tumours of breast cancer [44].

rs374748 on FBN2 (Fibrillin) which had been associated with obesity in previous studies [45] have been found to be positively associated with breast cancer in this study which may be due to well known association of obesity and breast cancer. Some of the other SNPs which had shown association with body mass index (BMI) [46] and obesity [47-48], weight gain or overweight and showed positive association with breast cancer in this study are rs2922763.
Hepatocyte Nuclear Factor 4-Gamma (HNF4G), rs2116830 (KCNMA1 - Potassium Channel, Calcium-Activated, Large Conductance, Subfamily M, Alpha Member 1) and rs10953454 (LHFPL3 - Lipoma HMGIC Fusion Partner-Like 3). A positive association was observed with SNPs rs11121832, rs16886165, rs11594610 and rs2274459 in genes MTHFR (Mitochondrial Carrier Homolog 2), MAP3K1 (Mitogen-Activated Kinase Kinase Kinase 1), TCF7L2 (Transcription Factor 7-Like 2) and MLN (Motilin) respectively.

**SUMMARY AND CONCLUSIONS**

The strongest risk factors associated with breast cancer after adjusting for confounding variables are as follows:

1. For every 2 year increase in the age at first full-term pregnancy there is a 10% increase in risk of breast cancer.
2. For every 5 cm increase in height there is an increase of breast cancer with OR = 1.09 (95% CI-1.02-1.17).
3. WHR showed strong significant positive association with breast cancer in both rural-urban areas and in pre- and postmenopausal women. The risk was more than 3-fold in highest category (≥0.95) as compared to lowest category (<0.85).
4. Four SNPs selected from FGFR2 gene were positively associated with breast cancer. Some of the other SNPs identified in this study are rs11121832, rs16886165, rs11594610, rs2116830 and rs2274459 in genes MTHFR, MAP3K1, TCF7L2, KCNMA1 and MLN respectively. These are SNPs related to inflammation, obesity and signal transduction pathway.

The current study demonstrates that protection observed for breast cancer by living in a rural area is possibly because of less prevalence of risk factors viz. late age at first full-term pregnancy and central obesity which are observed to be strongly involved in the disease etiology. It’s therefore possible to adopt public health strategies to prevent/reverse increasing
trends of breast cancer by monitoring the lifestyle. The strategies to reduce central obesity (and not only BMI) should be evolved as this will be helpful not only in the prevention of breast cancer but also other non communicable diseases. Efforts should be made to prevent late age at first pregnancy by proper counselling and informing about the risk associated with it.
BIBLIOGRAPHY


[21] StataCorp. Stata Statistical Software. College Station, TX: StataCorp LP; 2011.


[40] Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 2007;81:559–75.


