Chapter 3

Methodology

3.1 Study design

This research is a hospital-based case-control study conducted during the period of June 2006 to December 2007. It was done to determine the association between dietary isothiocyanates, *GSTs* polymorphisms and oral cancer risk. Cases group comprised of patients with oral cancer while the control group comprised of those without oral cancer.

3.2 Population and sample

3.2.1 Reference population

The reference population to which the results of this study would be inferred to was all the oral cancer patients in Malaysia.

3.2.2 Source population

The source population from which the sampling frame of cases and controls recruited from was all patients in the Oral Cancer Research & Coordinating Centre (OCRCC), University of Malaya database. This OCRCC database is an oral cancer data bank which comprises information of related parameters elicited from patients who attended nine selected centers namely the Dental Faculty University of Malaya, Universiti Sains Malaysia, Universiti Kebangsaan Malaysia, and the Ministry of Health Malaysia specialists' clinic at the General Hospitals of Kuala Lumpur, Selangor, Perak, Kelantan,

Sabah and Sarawak. These OCRCC data was collected in a standardized manner pertaining to socio-demographic data, risk habits, diet, diagnosis, clinical staging, histological grading, gene expression and follow-up information for future auditing against assessments of disease outcome and behavior.

3.2.3 Sampling frame

All patients that fulfilled both the inclusion and exclusion criteria have formed the sampling frame for the study.

3.2.3.1 Inclusion criteria

For cases, we included:

i. New patients who were diagnosed pathologically with squamous cell carcinoma of the oral cavity (OSCC) at the nine selected centers
ii. Patients with complete diet data
iii. Patients with genomic DNA in the nuclei acid bank (at OCRCC-UM and CARIF – Cancer Research Initiatives Foundation) or blood samples

For controls, we included:

i. Patients who do not have oral cancer, potentially malignant lesions and as well as no other cancers
ii. Patients with complete diet data
iii. Patients with genomic DNA in the nuclei acid bank (OCRCC-UM and CARIF) or blood samples were included in this study

These patients could be attending the centers for minor ailments.

3.2.3.2 Exclusion criteria

For cases, we excluded patients who were non-Malaysian citizens and those who already had cancer and are currently undergoing treatment. This is crucial to eliminate patients with recurrence cancer. For controls, we excluded patients who had cardiovascular disease (CVD), hypertension, gastrointestinal (GI) tract disease, liver disease associated with risk factors of interest in the study (smoking, alcohol drinking, betel-quid chewing) because this would introduce bias. For both case and control groups, patients and healthy individuals without genomic DNA from the nuclei acid bank (OCRCC-UM and CARIF) or blood samples and incomplete diet data were also eliminated.

3.2.4 Sample

As it was expected that there would not be many patients available in the sampling frame, thus, we have decided to include all patients with informed and written consent as the subjects. Therefore, no sampling method was employed in this study. Informed consent was obtained for the umbrella project Oral Cancer & Precancer in Malaysia – Risk Factors, Prognostic Markers, Genetic Expression & Impact on Quality of Life, IRPA RMK 8 Project No: 06-02-03-0174 PR 0054/05-05 where this current project is a part of the umbrella project. The ethical approval for the umbrella project was also obtained with the Medical Ethics code no. DF OP0306/0018/(L) and endorsed by the Ministry of Health Malaysia. Finally, case comprised of 115 patients and control group comprised of 116 patients selected from within the database. The control patients were unmatched for age, gender and ethnicity.

Below is the summary of the study presented in a flowchart (Figure 3.1).

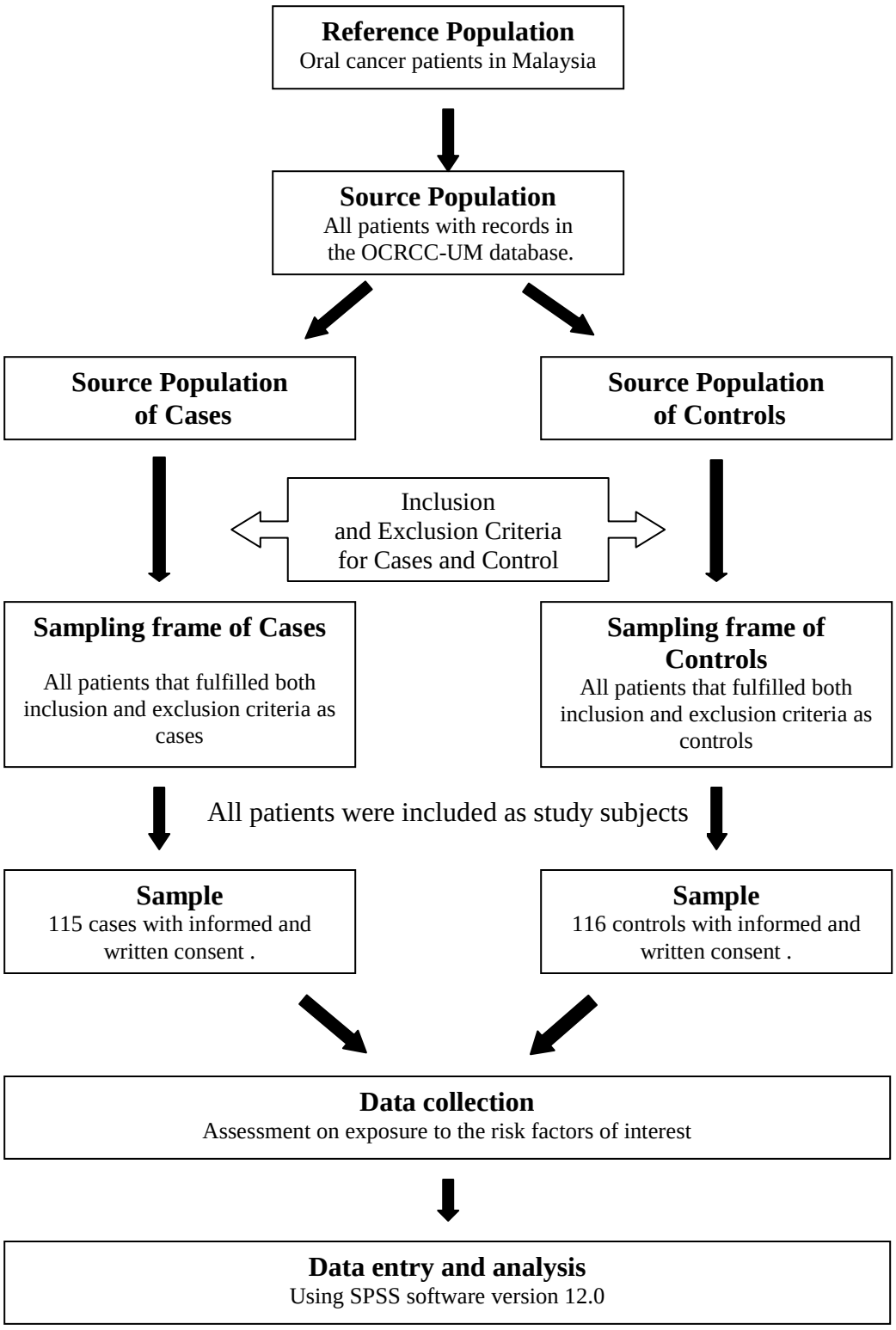


Figure 3.1: Flowchart of the methodology of the study

3.3 Sample size estimation

Using the Epi Info Version 3.3.2, the sample size was determined for all objectives that have the information needed available. Estimation based on the objective to determine association between dietary ITC and oral cancer risk has yielded the largest affordable sample size (appendix A). The estimation took into account the requirement for significance level of 0.05 and having at least 80% power of study. The ratio of control to cases was 1: 1 and the odds ratio (OR) of exposure among cases for detection was 0.45 (with 95% confidence interval (CI)). The parameter used was 70% prevalence of high dietary intake of ITC in population (London *et al*., 2000, Zhao *et al*., 2001, Seow *et al*., 2002 and Wang *et al*., 2004). Thus, the estimated sample size for one control to one cancer patient generated was 115 participants in each group, making a total of 230 subjects. Some of the constraints that were also taken into considerations are limited time to conduct the study and higher cost incurred in obtaining DNA extraction and PCR kit, reagents, restriction enzyme and the agarose gel.

3.4 Variables

The outcome or the dependent variable was having oral cancer. Risk factors of interest or the independent variables were grouped into socio-demographic factors, dietary ITC intakes and genetic polymorphism. Below is the summary of the independent and dependent variables used in this study (Figure 3.2).

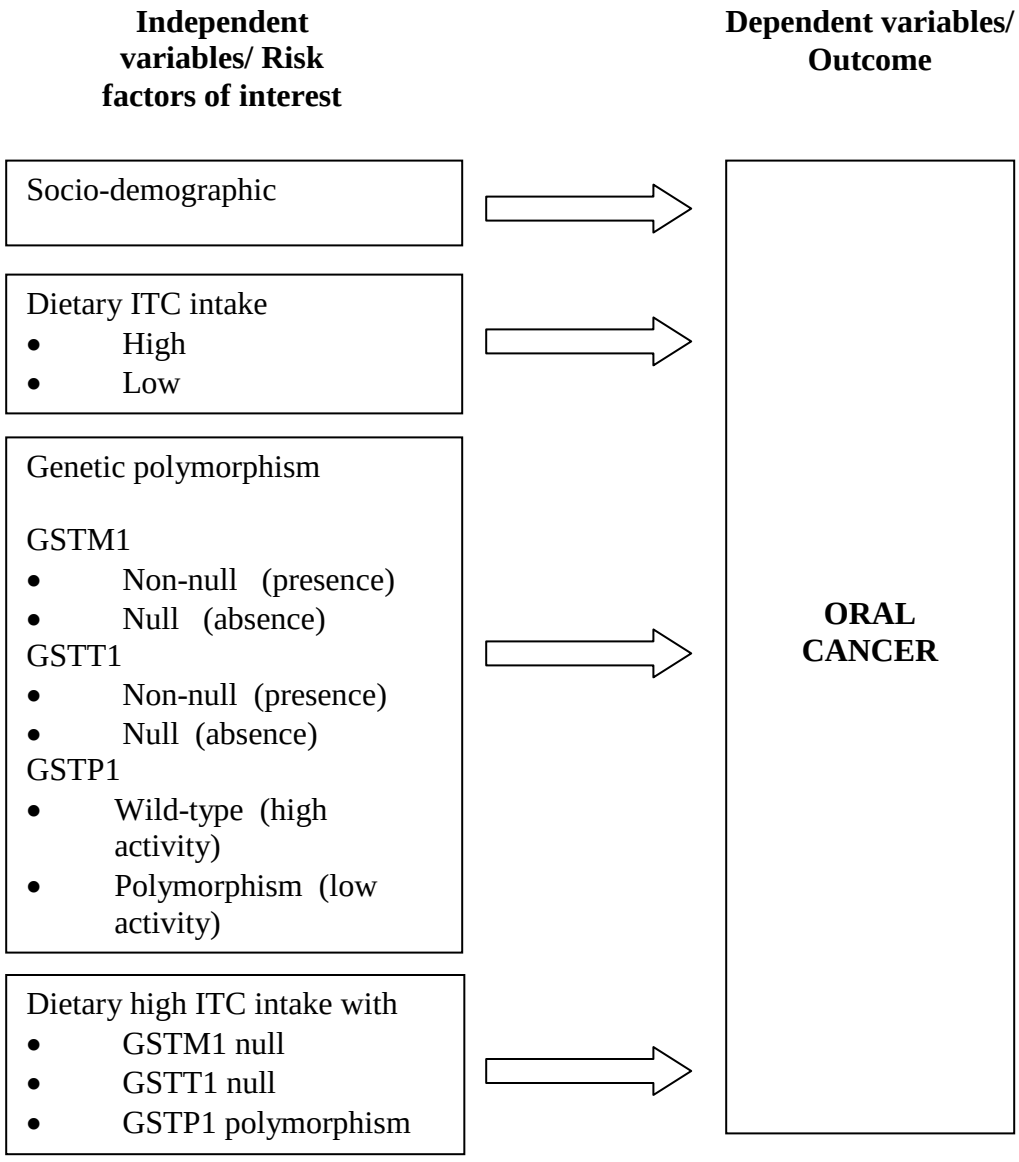


Figure 3.2: Independent and dependent variables of this study

3.5 Measurement tools

Since part of this study was using secondary data from OCRCC database, the validity of the information is crucial. Taking that into account, efforts were made to ensure

that measurement tools used to collect the information needed were reliable. Different measurement tools were used to measure different exposures or risk factors of interest.

3.5.1 Dietary ITC intake

Two measurement tools were used in gathering information on dietary ITC intake. The first was a semi-quantitative food frequency questionnaire (FFQ). The researcher then used the secondary data or information obtained from the FFQ to quantify the mean energy and daily nutrient intake using NutrieMart Version 2.

i. Semi-quantitative food frequency questionnaire (FFQ)

FFQ has been validated specifically for the Malaysian population according to various ethnic groups to assess habitual diet. This FFQ contains 9 major food groups such as dairy products, fruits, vegetables, meat/fish/seafood, bread/cereals, beverages, snacks/miscellaneous, fermented and processed food. It was an interview guided questionnaire where the research clerks at the respective hospitals and clinics were trained prior to conducting the interview. Interviewer would then ask the oral cancer patients to estimate their usual dietary intake 1 year prior to diagnosis, whereas controls were asked to estimate their usual dietary intake for the previous year. For each individual food, subjects were given option to select from nine frequency categories (ranging from ‘never’ to ‘more than six times a day’). For each of the food items in the FFQ, a score based on the daily equivalent was tabulated to determine the estimated intake of food by the patient.

ii. NutrieMart Version 2

This NutrieMart software is a reliable and valid nutritional software used as a tool to quantify the total energy and daily nutrient intake of individual food consumption. The software is designed to quantify specific macro and micronutrients of each food items such as protein, vitamins, minerals and etc. This software used in this study was calibrated periodically to ensure the reliability and accuracy of measurements.

3.5.2 GSTM1, GSTT1 and GSTP1 determination

In the genotyping assays to determine the *GSTM1*, *GSTT1* and *GSTP1* genes, the measurement tools involved are QIAamp® Blood Mini Kit (Qiagen), multiplex polymerase chain reaction (PCR), polymerase chain reaction – restriction fragment length polymorphism (PCR-RFLP) and the agarose gel.

i. Qiagen Blood Mini Kit

Qiagen Blood Mini Kit was used to extract genomic DNA from the buffy coat of the blood samples. This kit is reliable to produce a substantial amount of genomic DNA with good quality for reliable PCR downstream application. In fact, many had also used this Qiagen Blood Mini Kit to extract the genomic DNA in their studies (Hahn *et al*., 2002; Nazar-Stewart *et al*., 2003; Schneider *et al*., 2004; Sugimura *et al*., 2006).

ii. PCR and PCR-RFLP

PCR is widely used in biological research to amplify part of a gene. In this study, however, in order to amplify two genes concurrently from the genomic DNA, multiplex PCR was commonly practiced. This multiplex PCR is a fast and reliable technique to determine the presence or absence of both the *GSTM1* and *GSTT1* gene simultaneously. This time-saving genotyping procedure which includes albumin as the internal positive control was employed in many studies (Nair *et al*., 1999; Rossini *et al*., 2002; Nazar-Stewart *et al*., 2003; Sweeney *et al*., 2003, Gattas *et al*., 2004; Naveen *et al*., 2004; Cho *et al*., 2005; Sugimura *et al*., 2006).

PCR-RFLP is another method to detect the *GSTP1* gene. In order to determine the ile/ile, ile/val and val/val genotypes of *GSTP1*, an additional procedure which required the PCR products to be digested with restriction enzymes were applied. This method was proven to be reliable as shown in studies by Park *et al*., (1999), Tan *et al*., (2000), Wang *et al*., (2003) and Chan-Yeung *et al*., (2004). Finally, from PCR and PCR-RFLP procedures, the output would be the three PCR products which contained the specific required genes - *GSTM1, GSTT1* and *GSTP1*.

iii. Agarose gel

GSTM1, GSTT1 and *GSTP1* genes in the PCR products need to be separated into non-null or null genotype for *GSTM1* and *GSTT1* and wild type and polymorphism genotype for *GSTP1*. For this purpose, agarose gel was used. It is the most common and effective way to isolate the PCR and PCR-RFLP product. Agarose gel electrophoresis

allowed the bands to be visualized under the UV transmission. These bands indicate the corresponding base pair for the respective genotype of interest. In fact, this reliable method was done in almost all PCR and PCR-RFLP studies (Olshan *et al*., 2000; Wang *et al*., 2003; Srivastava *et al*., 2005).

3.6 Data collection

Data collection was done in over a period of six months. It involved case ascertainment and other background variables, gathering the information on diet and also laboratory work to determine the genotypes.

3.6.1 Case ascertainment and socio-demographic profile

Oral cancer cases were identified through the OCRCC database in which oral cancer patients were diagnosed clinically and also confirmed by histopathology examination. Socio-demographic information such as patients' habits for tobacco smoking, alcohol drinking and betel-quid chewing, occupation, medical history, family history of cancer, age, gender and ethnicity were also obtained from the OCRCC database. These socio-demographic data were prior collected via interview by trained personnel using a structured questionnaire.

3.6.2 Information on dietary ITC intake

Information on dietary ITC intake was extracted in two stages. The first stage was utilizing secondary data which was collected before through a semi-quantitative food frequency questionnaire (FFQ). The researcher then used the information obtained from the FFQ to quantify the total energy and daily nutrient intake including ITC using NutrieMart Version 2. For the purpose of this study, the vegetables component of the FFQ was the main focus, thus only this component was assessed to extract the information on dietary ITC intake. Of the 21 vegetables listed, there are three members of the *Brassicaceae* family. These three cruciferous vegetables were commonly consumed by Malaysian. There are cabbage (*Brassica oleracea var. capitata*), cauliflower (*Brassica oleracea var. botrytis*) and kai lan (*Brassica oleracea var. alboglabra*, also known as Chinese kale) / choy sum (*Brassica oleracea var. parachinensis*, also known as Chinese flowering cabbage). Watercress (*Nasturtium officinale*) and broccoli (*Brassica oleracea var. italica*) are infrequently consumed in this population thus were not included in the questionnaire. For each of the food items, a score based on the daily equivalent was tabulated to determine the estimated intake of food by the patient. The scores for the estimated intake of each food were as follows (Table 3.1):

Table 3.1: Scores for estimated intake of food

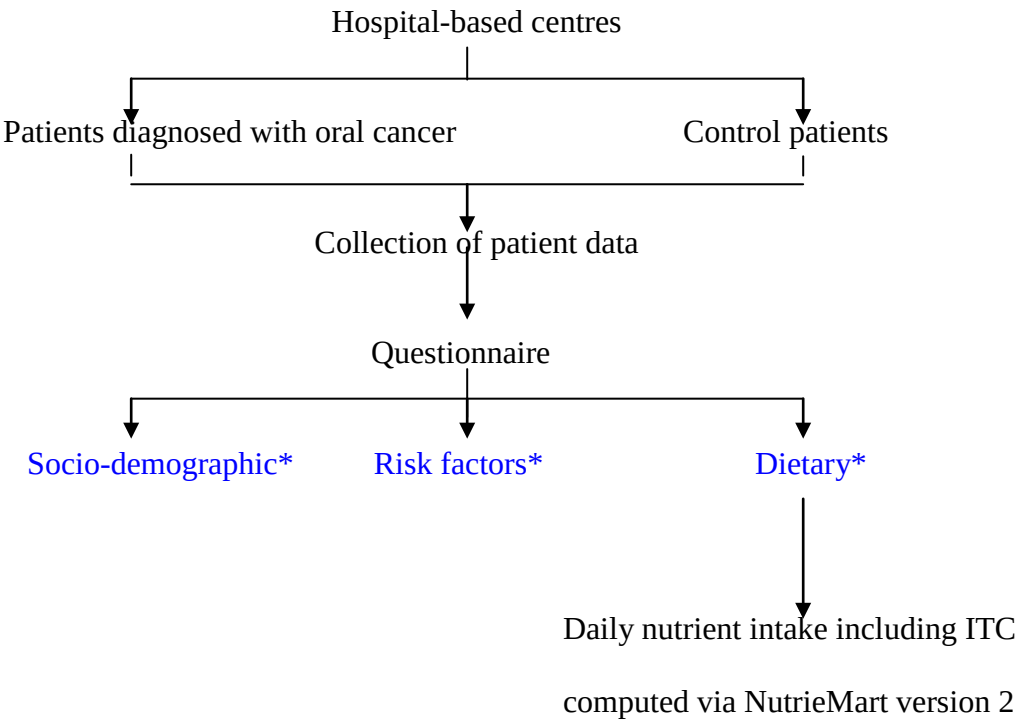
Frequencies categories	Daily equivalent score
Seldom/never	0/7 = 0
1-3 times per month	3/30 = 0.10
Once per week	1/7 = 0.14
2-4 times per week	4/7 = 0.57
5-6 times per week	6/7 = 0.86
Once per day	1
2-3 times per day	3
4-5 times per day	5
More than 6 times per day	6

For each patient, the scores for the estimated intake of each food obtained were entered into nutritional software called NutrieMart version 2. This was to compute the total energy intake and daily nutrient intake as the composition of nutrients for each food items which was based on the Nutrient Composition of Malaysian Foods guidebook compiled by Tee *et al*., (1997) (Appendix B). The guidebook mainly consists of the basic macro and micro-nutrient and minerals for almost all types of Malaysian food. The final output from NutrieMart software would be the total energy intake and daily nutrient intake for each food item that also include dietary ITC intake.

However, because each patient total energy or calories intake may differ from the recommended daily allowance, it was not justified to quote the dietary ITC intake directly from the nutrient data computation. Therefore, to avoid biasness, the total dietary ITC intake for each patient was divided by his total calories intake times a thousand. This

standardized dietary ITC intake would be based on a thousand of the total energy intake (per 1000kcal) of the patient and finally was used as estimated dietary ITC intake for each patient.

Despite all this, NutrieMart has some limitations whereby a few exotic food items in the FFQ were not included in the software. Therefore, these food items such as palm shoot, bat and bambangan were excluded from computation and this may underestimate the total energy and daily nutrient intake. However, effort has been taken where the researcher has checked and found that the amount consumed on the excluded food item was not much and may not carry weightage or any significant difference in the total calorie intake. The collection of patients' information was summarized as below in Figure 3.3.



* Secondary data

Figure 3.3: Collection of patient information

3.6.3 Determination GSTM1, GSTT1 and GSTP1 genotypes

In general, determination of *GSTM1*, *GSTT1* and *GSTP1* genotypes involved isolation of genomic DNA, genotyping of *GSTM1*, *GSTT1* and *GSTP1* genes and finally, determination of each gene genotypes either presence/high activity (non-null or wild type) or absence/low activity (null or polymorphism).

3.6.3.1 Isolation of genomic DNA

While most of the genomic DNA from blood samples was obtained from the nuclei acid bank (at OCRCC-UM and CARIF – Cancer Research Initiatives Foundation, Subang), the researcher had also extracted some of the genomic DNA for this study when the genomic DNA was not available in the nuclei acid bank and only blood samples in buffy coat was available. These blood samples were obtained with prior consent from the patients. The genomic DNA was extracted using QIAamp® Blood Mini Kit (Qiagen, Inc, Chartworth, CA) following the manufacturing instructions with slight modifications. During the final elution step, 100µl of Buffer AE was added to the QIAamp Spin Column to increase the final DNA concentration in the eluate. The QIAamp Spin Column loaded with Buffer AE was incubated at room temperature (15-25^{o}C) for 5 minutes before centrifugation to increase the DNA yield. For each of the genomic DNA obtained, both the $A_{260/280}$ ratio and the quantity of DNA (ng/µl) were recorded using NanoDrop Spectrophotometer. To prevent contamination and cross-contamination between samples during polymerase chain reaction (PCR), careful attention was given during DNA extraction. The extraction of DNA samples was performed in a location distant from the workstation where PCR amplifications were performed.

3.6.3.2 Genotyping GSTM1, GSTT1 and GSTP1 assays

For the genotyping of *GSTM1* and *GSTT1*, some of the data were extracted from a concurrent study done at OCRCC (Zain *et al*., 2006). However, for some of the newly recruited control patients' samples, the genotyping of *GSTM1* and *GSTT1* of these samples were carried out similarly following the protocol from Nair *et al*., (1999). The genotyping of *GSTM1* and *GSTT1* were determined using a modified multiplex PCR approach for simultaneous replication of both genes. The co-amplification of an albumin gene fragment served as internal positive control for a successful amplification reaction. Primers were used to identify the targeted genes *GSTM1* and *GSTT1* from the extracted genomic DNA. The sequences for the primer pairs were as follows:

Table 3.2: Primer sequences for *GSTM1*, *GSTT1* and albumin

Genotype	Primer sequences
GSTM1	
Forward	5' - GAA CTC CCT GAA AAG CTA AAG C - 3'
Reverse	5' - GTT GGG CTC AAA TAT ACG GTG G - 3'
Albumin	
Forward	5' - GCC CTC TGC TAA CAA GTC CTA C - 3'
Reverse	5' - GCC CTA AAA AGA AAA TCG CCA ATC - 3'
GSTT1	
Forward	5' - TTC CTT ACT GGT CCT CAC ATC TC - 3'
Reverse	5' - TCA CCG GAT CAT GGC CAG CC – 3'

PCR was performed in a 50µl reaction volume containing 50mM KCl, 1.5mM $MgCl_2$, 20mM Tris-HCl (pH 9.0), 200µM deoxynucleotide triphosphate (dNTPs), *GSTM1* primers at 3µg/ml each, *GSTT1* primers at 2µg/ml each, albumin primers 1µg/ml each, GoTag® DNA Polymerase (2.5U; Promega, USA) and 50-100ng of genomic DNA. These PCR reaction mix were then placed in a GeneAmp PCR system 2727 (Perkin-Elmer, Applied Biosystems). After an initial denaturation at 95^{o}C for 5 minutes, amplification was carried out for 40 cycles at 94^{o}C (denaturation) for 1 minute, 64^{o}C (annealing) for 1 minute and 72^{o}C (elongation) for 1 minute, followed by final elongation at 72^{o}C for 7 minutes. Finally, the PCR products containing the *GSTM1* and *GSTT1* genes were separated by agarose gel electrophoresis (2%) and visualized with ethidium bromide to determine the genotype either non-null or null.

As for genotyping the *GSTP1*, all genomic DNA samples were screened for the presence of the *GSTP1* gene codon 105 polymorphism using PCR-restriction fragment length polymorphism (RFLP) analysis, following closely the procedures described by Park *et al*., (1999). For the *in vitro* amplification of codon 105 of the *GSTP1* gene fragments from each sample, the primers homologous to exon 4 and intron 5 of the *GSTP1* gene have been synthesized and were purchased from Sigma-Proligo Inc, Singapore. Primers were used to identify the targeted genes *GSTP1* from the extracted genomic DNA. The primers used were as follows:

Table 3.3: Primer sequence for GSTP1

GSTP exon 4	Primer sequence
Forward	5'-AAT ACC ATC CTG CGT CAC CT-3'
Reverse	5'-TGA GGG CAC AAG AAG CCC CTT-3'

The reaction mix for each of the reaction was prepared to amplify the samples and the negative control. The reaction was set up as follows:

Table 3.4: 1st reaction master mix

5X Green GoTag® Flexi Buffer	10µl
$MgCl_2$ (25mM)	3µl
dNTP Mix (10mM)	1µl
GSTP1 Forward	1µl
GSTP1 Reverse	1µl
GoTaq® DNA Polymerase (5U/µl)	0.5µl
Total	16.5µl

The volume of 16.5µl of the PCR mix was added to the appropriately labeled 0.2ml microcentrifuge tubes containing 50-100ng of genomic DNA and distilled water such as the following table. No DNA was added to the control reactions.

Table 3.5: 50µl reaction set up

Tube	Sample	Negative control
Volume of genomic DNA	1-10µl DNA	-
Volume of dH_2O	23.5-32.5µl	24.5µl
PCR 1st reaction master mix	16.5µl	16.5µl
Total	50µl	50µl

The 50µl PCR mix will consist of 1-10µl of the genomic DNA solution, 50mM KCl, 1.5mM $MgCl_2$, 10mM Tris-HCl (pH 8.5), 200µM of each of the four dNTPs, 2.5U of GoTag® DNA Polymerase (5U/µl, Promega, USA) and 0.2µM of each of the primers. The reaction was mixed by gently tapping the tube and quick spin to collect the reaction at the bottom of the tube. The tubes were then put into the GeneAmp® PCR System 2727 (PE Applied Biosystems) programmed for the following temperatures:

Table 3.6: Conditions for the reaction mixtures following incubations in the thermocycler

Program	Temperature	Duration	Cycle
Initial denaturation	95°C	2 minutes	1 X
Denaturation	94°C	30 seconds	
Annealing	58°C	30 seconds	40 X
Elongation	72°C	30 seconds	
Final elongation	72°C	10 minutes	1 X
Hold/soak	4°C	∞	

After confirmation of the amplified fragment of the expected size (568bp) on the agarose gel, the PCR product was digested with 2.5 units of restriction enzyme (RE) BsmAI (New England Biolabs, USA). The restriction enzyme reaction was prepared as follows:

Table 3.7: 2nd restriction enzyme reaction master mix

10X NEBuffer 3	5µl
PCR product	20µl
BsmAI (2.5U)	0.5µl
Sterile dH_2O	24.5µl
Total	50µl

The reaction was incubated at 55°C for 16 hours. Finally, the PCR products contained the *GSTP1* genes were later separated by agarose gel electrophoresis (2%) and stained with ethidium bromide to determine the genotypes either wild type or polymorphism.

3.6.3.3 Determination of *GSTM1, GSTT1* and *GSTP1* genotypes

Agarose gel was used to determine the genotypes of *GSTM1, GSTT1* and *GSTP1* genes in the PCR products either non-null or null genotype for *GSTM1* and *GSTT1* and wild-type or polymorphism genotype for *GSTP1*. For this purpose, the 2% agarose gel was prepared by microwaving 2g of agarose powder in 200ml of 1X TAE buffer (appendix C) until the agarose was fully dissolved. After the solution was cooled to about 55°C,

ethidium bromide was added to a final concentration of 0.5µg/ml and mixed thoroughly by gently swirling and the gel was poured into the gel mold with a well comb. After the gel was completely set, it was mounted in an electrophoresis tank and freshly made 1X TAE buffer was added until the gel was covered.

i. Agarose gel electrophoresis of GSTM1 and GSTT1

In order to finally determine the genotypes of *GSTM1* and *GSTT1*, both PCR products were resolved by agarose gel electrophoresis. Agarose gel electrophoresis would allow the bands to be visualized under the UV transmission. The presence of the bands at the specific base pair represented the expression of the respective genes. The presence of the 350bp DNA fragment corresponding to the albumin gene product provides an internal positive control for each reaction and should be seen in all PCR reactions as an indicative of a successful PCR. Without the presence of the internal positive control, the expression of *GSTM1* and *GSTT1* gene were considered invalid. The gene product of *GSTM1* null and *GSTT1* null were evidenced by the absence of a 215bp and 480bp fragment respectively as simplified in Table 3.8.

Table 3.8: Determination of GSTM1 and GSTT1 genotypes

DNA fragments	Absence	Presence
GSTM1 215bp	Null	Non-null
GSTT1 480bp	Null	Non-null
Albumin 350bp	Results for both GSTM1 and GSTT1 not valid due to absence of internal positive control	Indicative of a successful PCR

Figure 3.4 shows the representative banding patterns of *GSTM1* and *GSTT1* genotypes in this study. The descriptions for the electrophoresis gel banding patterns of *GSTM1* and *GSTT1* genotypes could be explained as belows :

- Lane 1: 100bp DNA ladder (Promega, USA)
- Lane –ve: Negative control
- Lane 2-16: Samples

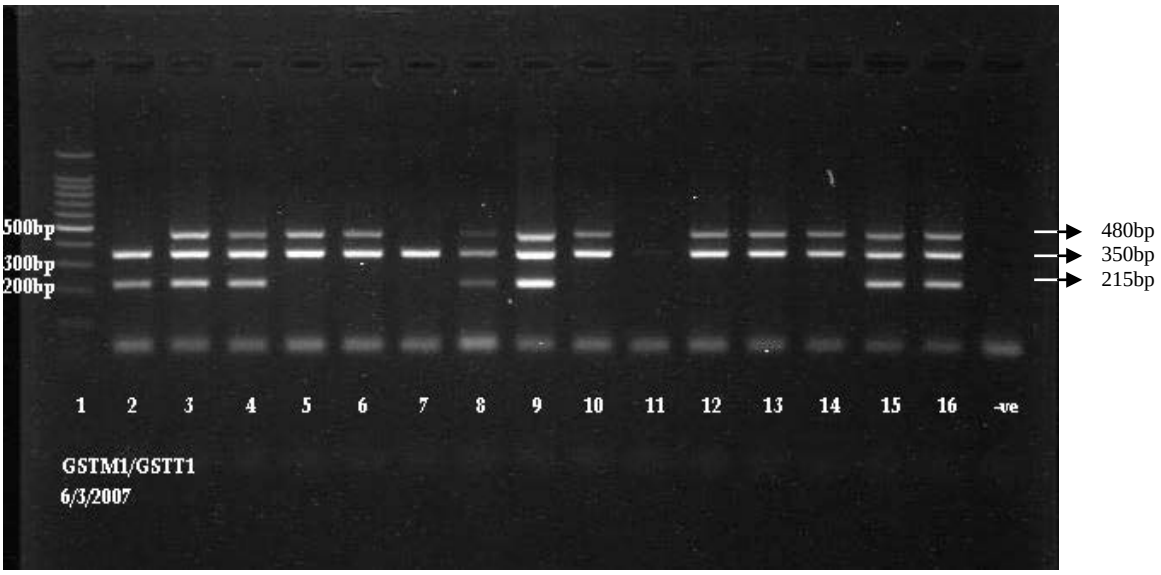


Figure 3.4: Banding patterns of *GSTM1* and *GSTT1* genotypes. Numbers on the left and right indicate base pairs of the DNA ladder and the *GSTM1* and *GSTT1* genotypes.

The presence of only 215bp product in sample 2 indicated that this individual has *GSTM1* non-null genotype (*GSTM1* gene was presence) and *GSTT1* null genotype (*GSTT1* gene was absent as the 480bp band was not expressed). Meanwhile the presence of both 215bp and 480bp bands in sample 3, 4, 8, 9, 15 and 16 showed that these individuals had both *GSTM1* and *GSTT1* non-null genotypes (both *GSTM1* and *GSTT1* genes were present). The presence of only 480bp product in sample 5, 6, 10, 12, 13 and 14 indicated that these individuals had *GSTT1* non-null genotype (*GSTT1* gene was present) and *GSTM1* null genotype (*GSTM1* gene was absent as the 215bp band was not expressed). Lastly, the absence of both 215bp and 480bp bands in sample 7 showed that this individual has both *GSTM1* and *GSTT1* null genotype (both *GSTM1* and *GSTT1* genes were absent). Sample 11 was repeated to confirm the expression of *GSTM1* and *GSTT1* gene.

ii. Agarose gel electrophoresis of *GSTP1*

Each PCR product containing *GSTP1* genes that have been digested with BsmA1, was electrophoresed on 2% agarose gel and stained with ethidium bromide. The nuclei acid is then separated by applying an electric current of 110 volts to the gel for 20 minutes. Finally, the gel was examined over ultraviolet light and photographed. Table 3.9 shows the banding patterns observed by restriction fragment length polymorphism (RFLP) analysis of *GSTP1* codon 105 genotypes. PCR products were incubated with BsmAI restriction enzyme. Digestion with BsmA1 proved to be very robust and reliable and the direct sequencing of PCR products from individuals with each genotype confirmed the accuracy of the method.

Table 3.9: Determination of GSTP1 genotypes

Wild type (normal)	Polymorphism	
Ile/Ile genotype (homozygote)	Ile/Val genotype (heterozygote)	Val/Val genotype (homozygote mutant)
	83 bp	83 bp
125 bp	125 bp	125 bp
138 bp	138 bp	138 bp
	222 bp	222 bp
305 bp	305 bp	

The DNA fragments after digestion with BsmAI showed the presence of 125bp and 138bp in all the *GSTP1* genotypes. A 305bp gene product is present only in those samples

containing the ile/ile genotype (wild-type). Meanwhile, in the *GSTP1* polymorphism, the ile/val genotype corresponded to the presence of the 83bp, 222bp and 305bp DNA fragments while the val/val genotype was evidenced by the presence of 83bp and 222bp DNA fragments.

Figure 3.5 and 3.6 show the representative banding patterns of *GSTM1* and *GSTT1* genotypes in this study before and after digestion of BsmA1. Before the digestion of BsmA1, the undigested PCR-amplified *GSTP1* exon 5 product produced a 568bp (Figure 3.5; Sample 2, 3, 4, 5, 6, 7 and 8). After the BsmA1 digestion, three banding patterns were observed by RFLP analysis of the *GSTP1* codon 105 polymorphism. In the *GSTP1* (ile/ile) genotype, the 568bp fragment was digested into 125, 138 and 305bp bands (Figure 3.6; Sample 5, 7 and 8). These individuals with 125, 138 and 305bp bands had wild-type genotype. Meanwhile, the presence of 83, 125, 138, 222, 305bp bands (Figure 3.6; Sample 2, 3 and 4) and 83, 125, 138, 222bp bands (Figure 3.6; Sample 6) which correspond to the *GSTP1* (ile/val) and *GSTP1* (va/val) genotype respectively, showing that these individuals had polymorphism genotype. For the purpose of analysis in this study, the *GSTP1* (ile/val) and *GSTP1* (val/val) were categorized into the polymorphism genotype.

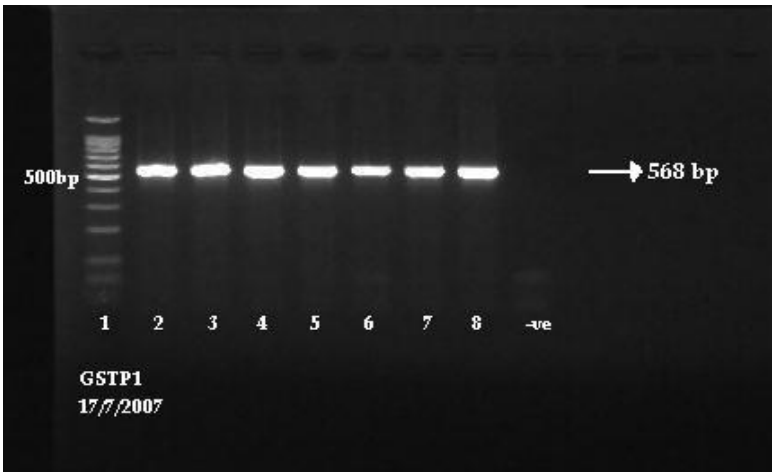


Figure 3.5: Banding patterns of the undigested PCR products of *GSTP1*. Lane 1: 100bp DNA ladder (Promega, USA); Lane 2, 3, 4, 5, 6, 7 and 8: undigested PCR-amplified *GSTP1* exon 5 product; Lane –ve: Negative control. Numbers on the left and right indicate base pairs of the DNA ladder and undigested products respectively.

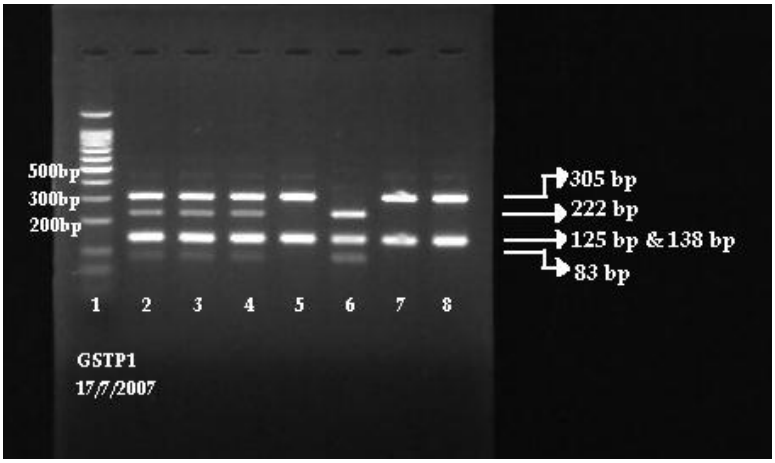


Figure 3.6: Banding patterns of the BsmA1-digested PCR amplifications product of *GSTP1*. Lane 1: 100bp DNA ladder (Promega, USA); Lane 2, 3 and 4: *GSTP1* (ile/val) polymorphism genotype; Lane 5, 7 and 8: *GSTP1* (ile/ile) wild-type genotype; Lane 6: *GSTP1* (val/val) polymorphism genotype. Numbers on the left and right indicate base pairs of DNA ladder and digested products respectively.

The genotyping assays for *GSTM1, GSTT1* and *GSTP1* was summarized as below in Figure 3.7 and appendix D.

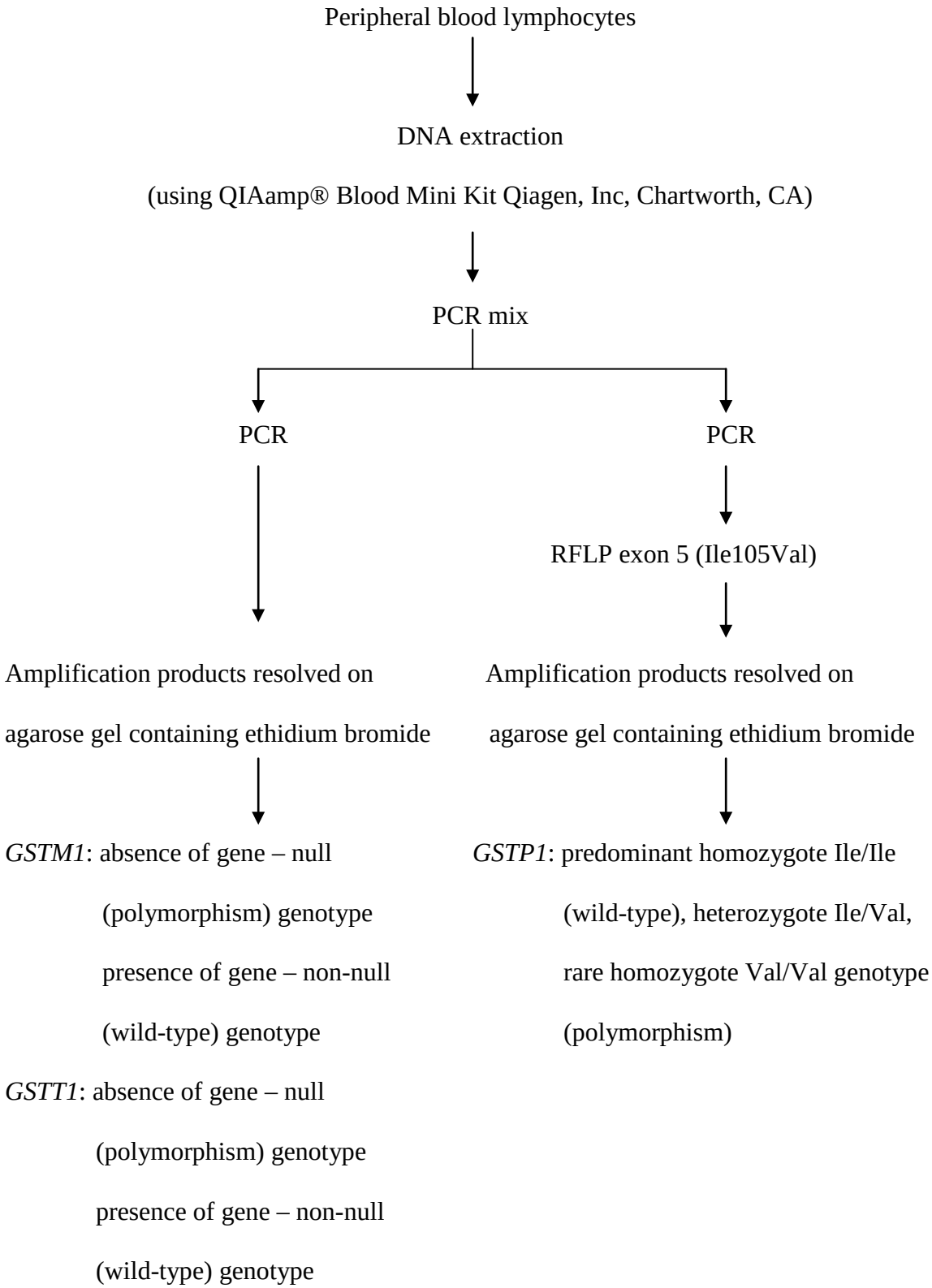


Figure 3.7: Genotyping of GSTM1, GSTT1 and GSTP1 assays

Finally, Table 3.10 summarizes the sources of data or information for the whole data collection process.

Table 3.10: Summary of sources of data or information

Variables	Sources				
	OCRCC database	FFQ	NutrieMart	Nuclei acid bank (OCRCC & CARIF)	PCR & PCR-RFLP
Dependent variable					
Patients with oral cancer – cases	X				
Patients with non oral cancer – controls	X				
Independent variables					
Sociodemographic profile	X				
Dietary ITC intake		X	X		
Genotyping assays					
- *GSTM1*				X	
- *GSTT1*				X	
- *GSTP1*					X

3.7 Statistical analysis

Data entry and data analysis was done using Statistical Programme for Social Science (SPSS) Version 12.0. Analysis was done according to the specific objectives of the study. Data were first checked and cleaned. The distribution and frequencies were examined. Categories with small sample size and skewed distribution were noted. Descriptive statistics were used to describe all the dependent and independent variables. Mean and standard deviation or median and inter quartile range were calculated for all continuous variables. Categorical variables were calculated as frequency and percentage. Characteristics differences in continuous variables between cases and controls were compared by independent t-tests while chi square tests were used for categorical variables. The level of significance was set at 0.05.

The specific objectives basically were to determine the association between dietary ITC intake, *GSTs* polymorphisms, *ITC-GSTs* polymorphisms interaction with oral cancer risk. To achieve these objectives, logistic regression was employed. Logistic regression analysis, yielding odds ratio (OR), is a widely used technique to measure the association between a risk factor of interest and disease outcome in epidemiological studies while controlling the effect of other potential risk factors. Under the rule that OR closely approximates the relative risk (RR) when the occurrence of disease in the study population is less than 10%, the interpretation of OR in this study would be similar to the relative risk which was the risk of having the disease among exposure than that among non-exposure. Data analysis was done in two stages. First, at univariate level, simple logistic regression analysis was used to screen and select the variables. Simple logistic regression analysis

would yield crude OR of the association. Meaningful combinations of categories for categorical independent variables were done if it was indicated.

In this study, dietary ITC was one of the exposures of interest. For the purpose of making the dietary ITC intake more meaningful, it was categorized into high and low ITC intake based on the median value for the entire case-control study. Low (or high) ITC levels were defined as lower (or higher/equals to) than the median dietary ITC intake among all case-control subjects.

Table 3.11: Levels of dietary ITC intake among all case-control subjects

Level	Indications
Low	Lower than the median dietary ITC intake
High	Equals to or higher than the median dietary ITC intake

For analysis of genotype, the null genotypes of *GSTM1* and *GSTT1* was compared with non-null while the *GSTP1* polymorphism genotypes were compared against the genotype expected to have the highest activity (wild-type). All the non-null and wild-type genotypes were used as the reference group.

Table 3.12: Expressions for the GST genotypes

GST genotypes	Expressions
GSTM1	
Non-null	*GSTM1* gene was present
Null	*GSTM1* gene was absent
GSTT1	
Non-null	*GSTT1* gene was present
Null	*GSTT1* gene was absent
GSTP1	
Wild-type	Consists of ile/ile genotypes with highest activity
Polymorphism	Consists of ile/val and val/val genotypes with low activity

For data analysis involving the combinations of more than one GST genotype, the following expressions were used. All the wild type for the combination genotypes were used as the reference group.

Table 3.13: Expressions for the combination of GST genotypes

GST genotypes	Expressions
GSTM1/GSTT1	
Wild type	Both *GSTM1* and *GSTT1* gene were present
Polymorphism	Either one or both *GSTM1* and *GSTT1* gene were absent
GSTM1/GSTT1/GSTP1	
Wild type	All *GSTM1, GSTT1* and *GSTP1* gene were present or have high activity
Polymorphism	Either one or all *GSTM1*, *GSTT1* and *GSTP1* gene were absent or have low activity

Logistic regression analysis was then done at multivariate level to identify the significant associated factors, where the adjusted OR was derived after adjusting for the other factors. 95% Confidence Interval (CI) of the ORs, Wald statistic, chi-square, and p-value of the association were obtained in order to make inferences to the study population. However, doing complete multiple logistic regression for risk modeling was still subjected to whether our exposures of interest were significant in the multivariate level.

Finally, to achieve the third objective, Cochrane-Mantel Haenszel stratified analysis by genotypes was done to look into association between dietary ITC and oral cancer risk when *GSTs* polymorphisms were controlled. Mantel Haenszel analysis would yield common OR which indicated that the OR for the association between dietary ITC and oral cancer when the particular *GST* was controlled. If the OR for each level such as *GST* non-null/wild-type was similar to null/polymorphism genotypes but differ from

common or pooled OR, then it suggested that the association between the dietary ITC and oral cancer was confounded by the particular *GST*. In case where at least one of the stratified OR differ from common OR, and they also differ from one another, it indicated that an interaction has occurred between the particular *GST* and dietary ITC. However, these indications would be valid to interpret when the association is significant.