MATERIALS AND METHODS

The present study was undertaken to assess the prevalence of different risk factors and their basic relationship and impact on the occurrence of the cardiovascular disease and the familial aggregation of blood pressure in the Ramadasia, a scheduled caste population of North-West Punjab in India.

3.1. Peoples of Punjab

“Punjab” is the land of five rivers (Sutlej, Beas, Ravi, Chenab and Jhelum) and had a significant ancient heritage. Punjab is a cultural region straddling the border between Punjab (Pakistan) and Punjab (India). Punjab on the globe is at 29° 30’ N to 32° 32’ N latitude and 73° 55’ E to 76° 50’ E longitude and has an area of 50, 362 km². People of Punjab are called Punjabis and their language is also called Punjabi. Agriculture is the main occupation of Punjab and economic dependence mainly pertains on their agricultural activities. Punjab has been considered original homeland for all Indo-Aryan tribes and they were branched out to different parts of the world. Punjab in India has always been considered the gateway of India for time immemorial. All foreign invaders like Aryans, Persians, Greeks, Scythians, Parthians, Huns, Turks and Mongols came to Indian through Punjab. Before the independence of India, Punjab was a vast territory but it has been divided into many parts after independence.

The rich and fertile land was the attracting ground of different people and races. The people of Punjab descend from different stocks and have heterogenous racial elements. Due to this fact all the foreign invaders came here, settled down and mixed their own culture with the local culture and in process their culture had also significant impact in the culture of Punjab. The population of Punjab being heterogenous consists of various castes, tribes and classes. But, it is enriched by the Sikh population which constitutes the 80% of Punjab population and contributes to 75% of Sikh population to the total Sikh population of India. The majority of the people are divided into Sikh and Hindu religion however, Muslim, Christian and Buddhism are also present.
A caste has been defined as, “endogamous group or collection of sub-groups having a common name and a common tradition, occupation and have common origin and are followers of same deity” (Sekhon 2000). The individual is obliged to marry outside his family but within the caste and sub-caste to which the family belongs. Marriage is prohibited between two individuals having common ancestors.

3.2. Ethnographic Profile and Reasons for Choosing this Target Population

The scheduled caste in Punjab is predominantly rural and semi-urban. The present study was undertaken in Ramadasia group which is a scheduled caste community of Punjab recognized by article 341(1) (2) of the constitution of India.

“Article 341(1): The President may with respect to any other State or Union Territory, and where it is a State after consultation with the Governor thereof, by public notification, specify the casts, races or tribes or parts of or groups within casts, races or tribes which shall for the purpose of this Constitution be deemed to be Schedule casts in relation to that State or Union Territory, as the case may be”.

“Article 341(2): Parliament may by law include in or exclude from the list of Schedule Casts specified in a notification issued under clause (1) any casts, race or tribe or parts of or groups within casts, races or tribes but save as aforesaid a notification issued under the said clause shall not be varied by any subsequent notification”.

There are 37 scheduled caste notified in the state of Punjab and they are classified at the census 2001 and Punjab hold the 1st rank in having scheduled caste population among all the States and Union Territories. The scheduled caste population of Punjab constitutes 28.9% of the total population of the state (Census 2001). The predominant castes of scheduled caste (SC) population in Punjab are known by different names such as Ramadasia, Ad dharmi, Balmiki and Majbi Sikh. Ramadasias are also known by different names in Punjab according to schedule castes and schedule tribes Amendment Act, 1976 (NO. 108 of 1976 dated 18th September, 1976 such as Chamar, Rehgar, Raigar and Ravidasi. Ramadasias originally were Hindu by religion. However, due to the social, cultural and religion influences of Sikhs many scheduled castes including Ramadasia caste have been converted to Sikh religion. However, there were different
thoughts prevailing regarding the religion and professions of Ramadasias. Some authors believed that Ramadasia (weavers) word originated from converted Chamars (cobbler) who became the followers of fourth Guru, Shri Guru Ramdas ji. Some authors believed that the term Ramadasia is an adaptation of Ravidasias, in affiliation to famous saint Shri Guru Ravi Das, a pioneer revival of Vaisnava (the worship of Vishnu in any of his forms or incarnations), but, the name Ravidasias is not familiar in Punjab. They generally belong to socially low class society. Ramadasias now have a mixture of profession and both Hindu and Sikh religion followers. Now they have changed their life styles due to the overall development of society in agriculture, communication, education and health aspects. As they have fairly good literacy rate and some graduates, post-graduates and professionals. Now they are in good government posts and are exposed to urban and western culture. Overall the economic level is more or less uniform. Ramadasia caste’s people have mixed type of diet pattern as they take cereals, pulses and vegetables in their diet. A part of this community also takes the non-vegetarian diet (chicken, meat, fish and egg). They also consume alcohol and have smoking habit. The converted Ramadasia Sikhs don’t smoke but some of them consume alcohol. Ramadasia caste is now made up of heterogeneous group of Hindu and Sikh religion. They do not belong to any one particular religion.

The Ramadasia communities had traditionally been relegated to the most menial labour with negligible possibility of upward mobility and also subject to social disadvantages and exclusion in comparison to the wider community. Therefore, socio-economic status of the families was rated as far below than the average Indian social structure.

More than 75% of total SC and 80% of Ramadasias in Punjab are living in villages. Scheduled caste population of Punjab mainly concentrated in 7 districts out of 22 districts including Gurdaspur, Jalandhar, Amritsar, Hoshiarpur, Ludhiana, Patiala, Sangrur. These districts account for 62.5% of the total SC population of Punjab. Ramadasias community is one of the predominant populations and constitutes 26.2% of the total SC population in Punjab and is considered the second largest group of SC population. The literacy rate of this community is fairly good upto 63.7% (Census, 2001).
The selection of this scheduled caste community from state of Punjab for this study is due to the fact that this huge population has never been examined before. The other reasons for chosen this backward community is the familial structure of the population. Mostly, they were living in joint family with atleast two generations and sharing common household environment, socio-economic, health and genetics. The composition of household consists of father, mother, grand father and mother, a group of brother and their wives and their sons and unmarried daughter. Most of the women came in the family after marriage and shared household of their husbands (Badaruddoza and Brar, 2006; Byard et al., 1989).

Hence, the present study is motivated by the fact that there is hardly any systematic genetic epidemiological study among these scheduled caste population in Punjab. However, the limitation is that results of present study among this community may not representative of the whole Indian population.

3.3. Sampling Design

This study used a stratified multistage cluster random sampling design (Sugathan et al., 2008). The world health organization (WHO) also recommended this kind of method of sampling to estimate the health conditions in a community (Indrayan, 2008). The present sample is supposed to represent Ramadasia, a scheduled caste population of ages 7 years and above including three generation i.e. grandparental, parental and offspring generation. The present survey had special emphasis on parental and offspring generation. The survey was oversampled to produce reliable estimates for these generations especially parental and offspring generations (Katzmarzyak et al., 2001).

Inhabitant patterns of Ramadasia community are very peculiar. They have a tendency to build their houses as a cluster in different pockets and points in the districts due to socio-economic factors. In the several stages, the clusters with aggregation of Ramadasias households have been identified. However, these clusters were also heterogeneous with respect to caste grouping which helps to increase the size of the sample and thus the precision without a corresponding increase in the cost, labor and time. This method of sampling is also very easy to administer and survey of the subjects is very quick due to close proximity which gives a rapid assessment method.
The clusters have been created with the help of households with respect to roads and lanes. The households of Ramadasia community with respect to roads and lanes have been identified by vigorous pilot survey with the help of older people, municipal commissioners, sarpanch, local political leader, voter list through surnames and relative circles of the investigator (who also belongs to the same community). This type of vigorous pilot survey has been completed in four selected districts of Punjab. Each cluster is made up of 6-10 households according to the availability of the location, close proximity and easy handling of clusters for a single investigator.

This obviously helps in reducing the cost, time and labor of the study. Theoretically it is also advisable not to increase the sample size in the clusters in the view of design effect and for a single investigator it is easy to study a large number of small clusters instead of small number of large clusters, which may help in incorporating various prevailing conditions of the population in the sample making it more representative. It is also necessary for the present study to form a smaller cluster with 6-10 households for complete enumeration of the clusters. This may also provide a comparatively more representative sample. Sampling was done in four stages.

### 3.3.1. First Stage: Stratification of Districts

In the first stage four districts (strata) Gurdaspur, Jalandhar, Amritsar and Hoshiarpur among total twenty two districts of Punjab were selected according to comparative concentration of Ramadasia inhabitants. The first stage units i.e. four districts (Gurdaspur, Jalandhar, Amritsar and Hoshiarpur) are called primary sampling unit as shown in figure 3.1.

![Figure 3.1: Map of Punjab highlighting the primary sampling units](image-url)
3.3.2. Second Stage: Sample Size Determination

The appropriate sample size for a population based study is determined generally by three factors:

i. The estimated prevalence of the variable of interest-cardiovascular disease in this study

ii. The desired level of confidence and precision, and

iii. The acceptable margin of error

The total sample size has been calculated according to the following formula:

\[
 n = \frac{t^2 \times p(1-p)}{m^2}
\]

Where, 

- \(n\) = required sample size
- \(t\) = confidence interval at 95% (standard value of 1.96)
- \(p\) = estimated prevalence of cardiovascular disease (hypertension) in this study
- \(m\) = margin of error

3.3.2.1. Calculation: It has been reported that overall 25% (0.25) of Punjabi adult population suffer from hypertension (Sidhu et al., 2005). Therefore, \(p=0.25\). To achieve the high precision of the study, the relative error has been set in this study to be 3% (0.03), though standard value is 5% (0.05)

\[
 n = \frac{1.96^2 \times 0.25(1-0.25)}{(0.03)^2}
 = 800.33
 \approx 800
\]

3.3.2.2. Design effect: The method of survey is designed as a cluster random sample. It has a disadvantage that the elements within cluster tend to be similar to one another and produce a clustering effect. This is also called a design effect. Hence, to correct the difference in design, the sample size is multiplied by the design effect, required to compensate a larger sample relative to simple random sample. The design effect is generally assumed to be 2 for anthropometric/nutrition surveys using cluster sampling methodology. Therefore, the present sample further increase as

\[800 \times 2 = 1600\]
3.3.2.3. **Contingency**: The sample size is further increased by 10% to account for contingencies such as non-response/recording error/ out of station / illness / death.

Therefore, the sample size in this stage: \(1600 \times 1.10 = 1760\)

3.3.2.4. **Final sample size**: Finally, the calculation result is rounded up to the closest number that matches well with the number of clusters i.e. at least 60 in the present study to be surveyed. Therefore, the final sample size was further increased and rounded up to 1800 plus.

\[ \Phi = \frac{1800}{60} = 30 \text{ individuals per cluster} \]

Statistical power has a kind of direct relationship with the level of significance. It has been decided through statistical power calculation for present study that 1800 plus samples are required to detect specified differences. It required 600 households (assumed through pilot survey calculation) which also required at least 60 cluster. In the present study one household is assumed to be one family. All individuals (who were eligible through inclusion criteria) in the households of selected clusters were included for the study.

Therefore, by cluster sampling method (Indrayan, 2008) a total of 61 clusters from four districts according to their Ramadasia population concentration have been selected.

In this stage, all households of selected 61 clusters were introduced to the investigator with the help of respective local leader/aged person/sarpanch/known individuals/local government officials to create the awareness among the individuals of selected households. The response rate is more than 95%. Therefore, a complex multistage probability design was used that gives differential probabilities of selection to various units.

**3.3.3 Third Stage: Questionnaire designing and methods of measurements**

All the informations such as personal, socio-demographic, medical history, family history of CVD, physiometric, anthropometric and lifestyle variables of subjects were collected through pre-tested self-designed questionnaire. The questionnaire was in English language (questionnaire is attached along Annexure-I). Before the data collection the entire questionnaire was explained in local Punjabi language to the
subjects along with the aims and objectives of the study and the procedure for the data collection. An informed consent was duly signed by the subject taken. In case of the offspring (≤18) the entire procedure was explained to their parents or any elder person in the family and his/her signature was taken on offspring’s questionnaire.

The present data was cross-sectional descriptive study and interview method was adopted as it provides an opportunity to the interviewer to extract the appropriate information by having a face to face contact with the subject in his/her residence. The interview of the subject comprised of three parts according to the questionnaire. The questionnaire was grossly divided in three parts:

i) **Socio-economic lifestyle variables:** The standard questions on demographic and socioeconomic information such as age, sex, family status, education level and family income and health (physical activity at work and during leisure time, smoking, alcohol, diet and disease, if any).

ii) **Physiometric measurements:** Physiometric measurements included following parameters; Systolic Blood Pressure (SBP), Diastolic Blood Pressure (DBP), Mean arterial Blood Pressure (MBP) Pulse Rate (PR), Pulse Pressure (PP).

iii) **Anthropometric Measurements:** The anthropometric measurements comprised of the following parameters; height, weight, waist circumference, hip circumference, arm circumference, calf circumference, biceps skinfold, triceps skinfold, waist to hip ratio (WHR) and body mass index (BMI).

### 3.3.3.1. Socio-demographic variables

#### 3.4.3.1.1. Personal information:

i) **Age:** Age refers to the length of the time during which a being has existed. In the present study, the number of years rounded to the nearest whole number the respondent lived since birth at the time of interview was taken as a measure of age.

ii) **Sex:** The members of many species of living things are divided into two or more categories called sexes (or loosely speaking, genders). These refer to complementary groups that combine genetic material in order to reproduce. This process is called sexual reproduction. Typically, a species will have two sexes: Male or Female.
iii) **Marital Status**: The concept of marital status applies to the conjugal arrangements of a person. It includes persons who are living together as husband and wife, regardless of whether they are legally married or in a common law relationship. Persons living in a conjugal relationship are identified as spouses. Spouses may be legally married spouses or common-law partners. In the present study, the procedure of collecting information on marital status of the respondent or respondent’s family was as follows: (a) Un-married (b) Married (c) Widow/widower; No of children : a) Male  b) Female.

3.4.3.1.2. **Socio-economic variables**: The present scoring patterns to quantify the socio-economic variables have been modified according to present senerio from the scoring methods followed by many investigators (Prasad, 1961, 1968, 1970; Kumar, 1993; Sengupta and Karmakar, 2007; Deb and Dasgupta, 2008; Dhargupta, 2009). However, the present modified scale mainly based on Agarwal (2008) on the basis of all India price index (AIWPI) and the cost of living index (COLI). A hypothetical value (0.53) has been calculated since 1993-94 when AIWPI started which value is calculated from COLI that pertains to whole sale price of India (WPI). It is a simple method of multiplying the multiplication factor with income limits and rounding off the value of nearest Rupee. The multiplication factor is calculated from value of AIWPI at the time of study by multiplying it with the hypothetical value.

The multiplication factor=\text{value of AIWPI}\times0.50

He also added one more parameter i.e. below poverty line to Prasad’s classification (Agarwal, 2008)

i) **Education**: Education is the act or process of imparting or acquiring general knowledge, developing the powers of reasoning and judgment, and generally of preparing oneself or others intellectually for mature life. There are different levels of education which represents a broad section of the education “ladder”, that is, the progression from very elementary to more complicated learning experience, embracing all fields and programme groups that may occur at that particular stage of the progression. Pattern adopted to record the educational status of the respondents included five categories: illiterate (person having no education), primary education (first 5–
7 years of formal or structured education), secondary/higher secondary education (7-12th year of schooling), graduate/post graduate (13 or more years of education) and professional (academic degree designed to prepare the holder for a particular profession). The scoring pattern of these categories is as follows: illiterate=1; primary=2; secondary/higher secondary=3; graduate/post graduate=4; professional=5 (Adapted and modified from Dhargupta et al., 2009).

ii) **Occupation:** The occupation of a person is an important indicator to determine the economic status of that person in a society. The respondent’s occupations were divided into four categories namely, idle (no work)/student/housewife/homely (had no occupation), labour, independent occupation and service. The scores for different categories of occupation were as follows: idle (no work)/student/housewife/homely=1; labour =2; independent occupation= 3; service= 4 (Adapted and modified from Dhargupta et al., 2009).

iii) **Income:** Income is the amount of money or its equivalent received during a period of time in exchange for labor or services, from the sale of goods or property, or as profit from financial investments. The income range of the present Ramadasia community was classified into 6 categories such as no income group (represented by housewife/student/no work), poor group (whose income range is 500-1499), lower middle income group (range from 1500-2999), upper middle income group (between 3000-4999), high income group (range 5000-9999), upper high group (10000 and above). In the present study the scoring pattern was as follows: no income=1; poor=2; lower middle=3; upper middle = 4; high=5; upper high=6 (Adapted and modified from Agarwal, 2008).

3.4.3.1.3. **Lifestyle variables:** Lifestyle is a term to describe the way a person lives, which was originally coined by Austrian psychologist Alfred Adler in 1929. It includes variables such as food habit, lifestyle, exercise, stress, smoking and drinking behavior.

i) **Food habits:** Dietary habits are the habitual decisions of a person influenced by culture in choosing what foods to eat. Many cultures hold some food preferences and some food taboos. Dietary choices can also define cultures and play an important role in
religion. These are broadly classified as vegetarian or non-vegetarian. Vegetarianism encompasses the practice of following plant-based diets (fruits, vegetables, etc.), with or without the inclusion of dairy products or eggs, and with the exclusion of meat (red meat, poultry, and seafood). Abstention from by-products of animal slaughter, such as animal-derived rennet and gelatin, may also be practiced. In the present study food habits was classified into two groups vegetarian and non-vegetarian. The scoring pattern is as follows: vegetarian=1 and non-vegetarian=2.

ii) Physical activity: Inactive or sedentary lifestyle refers to no or irregular physical activity. Sedentary activities included sitting, reading, watching television and computer use for much of the day with little or no vigorous physical exercise. A sedentary lifestyle can contribute in developing many chronic lifestyle diseases like type 2 diabetes and CVD. Here physical activity was classified as sedentary or active. The scoring pattern was: sedentary=1 and active=2 (Adapted and modified from Rastogi et al., 2004b; Deb and Dasgupta, 2008).

iii) Stress: It is the non-specific response of the body to any demand of change that requires a physical, mental or emotional adjustment or response. It has been categorized into no stress (student/children), physical stress, psychological stress and psychosocial stress and the scoring pattern was: no stress (student/children), physical stress=1; psychological stress=2; psychosocial stress=3.

iv) Exercise: It is any bodily activity that enhances or maintains physical fitness and overall health and wellness. Subjects were categorized with respect to the regularity of exercise performed by them into three groups, never exercise, occasional exercise and regular exercise. The scoring pattern was as follows: never exercise=1; occasional exercise=2; regular exercise =3 (Adapted and modified from Rastogi et al., 2004b).

v) Smoking: Smoking is the inhalation of the gases and hydrocarbon vapors generated by slowly burning tobacco in cigarettes. A smoker is defined as a person who had ever smoked at least 100 cigarettes in his life time and currently smokes every day or some days (McGruder et al, 2004). The present respondents were broadly classified into three categories: never smoked (they did not smoke at any time), former smoker (quit during at least past two years) and current smoker which is further divided into
three categories like light smoker (<5 cigarette), medium smoker (5-10 cigarette) and heavy smoker (>10). The scoring pattern is as follows: never smoked = 1; former smoker = 2; current light smoker = 3; current medium smoker = 4 and current heavy smoker = 5 (Adapted and modified from Nilsen et al., 2000).

vi) Alcohol drinking: Alcoholism is generally used to mean compulsive and uncontrolled consumption of alcoholic beverages, usually to the detriment of the drinker’s health, personal relationships and social standing. WHO (2001) has estimated that there are 140 million people with alcoholism worldwide. The present respondents were broadly classified into three categories: never taken (they did not drink at any time), former drinker (quit during at least past two years) and current drinker which is further divided into three categories like light drinker (<50 ml), medium drinker (50-100 ml) and heavy drinker (>100). The scoring pattern is as follows never taken = 1; former drinker = 2; current light drinker = 3; current medium drinker = 4; current heavy drinker = 5 (Adapted and modified from Nilsen et al., 2000).

3.4.3.1.4. Physiometric measurements: Physiometric variables included the measurement of systolic blood pressure (SBP), diastolic blood pressure (DBP) and pulse rate (PR). There are two calculated physiometric variables: pulse pressure (PP) and mean arterial blood pressure (MBP).

i) Blood pressure: The blood pressure is the pressure of the blood within the arteries. It is produced primarily by the contraction of the heart muscle. Blood pressure is measured by the recording of two numbers. The first (systolic blood pressure) is measured after the heart contracts and is highest. The second (diastolic blood pressure) is measured before the heart contracts and is lowest. A blood pressure cuff is used to measure the pressure. Elevation of blood pressure is called "hypertension". The measurements were taken with the help of mercury sphygmomanometer, after the subject had rested for five minutes in the supine position, with right forearm placed horizontal on the table. An appropriate sized cuff was used on the arm of the subject and was inflated to about 20 mmHg above the point at which the radial pulse disappeared. The pressure within the cuff was then released at the rate of 2 mmHg/sec (approx.), while osculating with stethoscope placed over the brachial artery. The onset
of sound (Korotokoff-phase I) was taken as indicative of systolic blood pressure (SBP) and the disappearance of sound (Korotokoff-phase V) was taken as indicative of diastolic blood pressure (DBP). Korotokoff phase was taken as recommended by American Heart association (1981) and others. At least two readings of blood pressure were taken. But, when the difference between the two readings was found to be more than 5 mmHg, then the third reading was taken and the average of the three measurements was used as the estimate of SBP and DBP in this study. All the efforts were made to minimize the factors like anxiety, fear, stress, laughter and recent activity which might affect blood pressure. The units of measurement taken were mmHg (American Heart Association, 1981).

**Instrument used:** Sphygmomanometer (Diamond: regular # BPMR112), Stethoscope (3M Littmann #3128).

**Cuff size:** (13×26 cm for adults and 9×19 cm for children)

**Note:** Prior to the sample collection, the investigator had been trained for two months in recording blood pressure from Batala Hospital, Batala, District Gurdaspur (Punjab).

Based on the SBP and DBP recorded the subjects were classified into normotensives, pre-hypertensives and hypertensives.

ii) **Pulse Rate:** The alternate expansion and recoil of elastic arteries after each systole of the left ventricle create a traveling pressure wave that is called pulse (Tortora and Grabowski, 1996). It is strongest in the artery closest to the heart. It becomes weaker in the arteriole and disappears altogether in the capillaries. The radial artery at the wrist is most commonly used to feel the pulse by pressing it with the index and middle fingers. It was counted over a minute.

3.4.3.1.5. **Anthropometric variables:** All the anthropometric measurements were taken on each individual using standard anthropometric measurement techniques (Singh and Bhasin, 1968; Weiner and Lourie, 1981). The detailed procedures of different anthropometric measurements are given below:

i) **Height (cm):** Height was measured without shoes with subject standing fully erect on a flat surface, with heels, buttocks and shoulders flat to the walls, and the
subject looking straight ahead. Then height was measured as vertical distance from vertex to floor. The reading was then recorded to the nearest 0.1 cm.

**Landmark:** Vertex to floor.

**Instrument used:** Anthropometer (Biocraft # 3AN-1)

ii) **Body weight (Kg):** The body weight of the subject was taken in kilogram while the subject was wearing usual light indoor clothes and no shoes on a weighing machine. The reading was then recorded to the nearest 0.1 kg.


iii) **Waist circumference (cm):** The subject was asked to stand erect with feet 25-30 cm apart, weight evenly distributed and with his abdomen relaxed. The measurement was taken mid-way between the inferior margin of the last rib and the crest of the ileum in horizontal plane in such a way that the tape was fitted snugly without compressing the soft tissue. The measurement was taken to the nearest 0.1 cm.

**Landmark:** Mid-way between the inferior margin of the last rib and the crest of the ileum in horizontal plane.

**Instrument used:** Steel tape (Markson measuring tape)

iv) **Hip circumference (cm):** Hip circumference was taken around the pelvis at the point of maximal protrusion of buttocks while the subject was standing with his/her feet close to each other. The measurement was taken to the nearest 0.1 cm.

**Landmark:** Around the pelvis at the point of maximal protrusion of buttocks while the subject’s feet are close to each other.

**Instrument used:** Steel tape (Markson measuring tape)

v) **Arm circumference (cm):** Arm circumference was measured as maximum circumference of upper arm taken horizontally i.e. where the biceps muscles were most developed. The measurement of arm was taken from the left side of the body. The measurement was taken to the nearest 0.1 cm.

**Landmark:** Maximum circumference of upper arm.
Instrument used: Steel tape (Markson measuring tape)

vi) Calf circumference (cm): Calf circumference was measured as the maximum circumference of lower leg when the calf muscle was relaxed and the subject was asked to sit so that his knee was bent at right angles to the lower leg. The measurement was taken at the right angles to the axis of lower leg where it was maximum developed. The measurement was taken to the nearest 0.1 cm.

Landmark: The axis of lower leg where it is maximum developed.

Instrument used: Steel tape (Markson measuring tape)

vii) Biceps skinfold (mm): The biceps skinfold was measured over the biceps muscle in the middle of the upper arm. Picked the skin and subcutaneous tissue fold over the biceps muscle about one centimeter above the marked level (mid-point of the distance between the inferior border of the acromion process and the external superior border of the head of the radius in the line of capital fossa). Applied the jaws of Lange caliper at the marked level. Precaution was taken to pick up all the subcutaneous adipose tissue. The measurement was noted two seconds after applying the full pressure of caliper. The measurement was taken to the nearest 0.2 mm.

Landmark: Mid-point of the distance between the inferior border of the acromion process and the external superior border of the head of radius in the line of capital fossa.

Instrument used: Lange caliper (Beta technology Santa Cruz PAT no: 3.008.239)

viii) Triceps skinfold (mm): The triceps skinfold was measured over the triceps muscles in the middle of arm at the level of upper arm circumference in the line of olecranon process. Marked the midpoint on the upper arm and picked up skinfold about one cm above the marked level. Applied the jaws of Lange caliper to the fold at the marked level and noted down the value after two seconds. The measurement was taken to the nearest 0.2 mm.

Landmark: Halfway between the acromion (shoulder) and olecranon process (elbow).

Instrument used: Lange caliper (Beta technology Santa Cruz PAT no: 3.008.239)
3.4.4. **Fourth stage: Data Collection and Field Operation**

The survey combines face to face interview and physical examination method for eliciting data. The interview includes demographic, socio-economic, dietary and health related questions. The examination comprises physiological measurements such as SBP, DBP and pulse rate and battery of anthropometric measurements such as height, weight, waist circumference, hip circumference, arm circumference, calf circumference, biceps skinfold and triceps skinfold. All these measurements were conducted at respective households through subsequent visits. The subjects were provided complete transcripts of the examination results. However, all information has been kept fully confidential and privacy was maintained and protected.

A cross-sectional survey was done in the four districts of Punjab namely, Gurdaspur, Jalandhar, Amritsar and Hoshiarpur. Houses were selected randomly from the particular locality. Mostly houses are situated in linear row fashion, one after another. Each house was visited twice during collection of data. During first visit all the information was gathered about subject’s personal and social information (age, gender, number of individuals in family including number of children in case of the parents and grandparents, education, occupation, income), anthropometric measurements (height, weight, waist circumference, hip circumference, arm circumference, calf circumference, biceps skinfold and triceps skinfold) and physiometric measurements (blood pressure, pulse rate). Only those subjects were included in the study those were healthy and had not taken medicine prior two weeks. A second visit was made after two weeks. During this visit physiometric measurements were taken only from those subjects who were previously included in the first visit. No new enrollments of subjects were made this time and some subjects were also excluded from the study due to unavailability or some kind of illness (medicine taken). However, for those subjects that were not available in the second visit for physiometric measurements, a third visit was made with prior information. Furthermore, the subjects who were again absent in the third visit due to illness or any other reasons were excluded from the study.
3.5. Inclusion and Exclusion Criteria

i) **Inclusion criteria:** Healthy individuals were selected from four districts of Punjab (Gurdaspur, Jalandhar, Amritsar and Hoshiarpur) and only those individuals who had not taken any medication 2 weeks prior to study were chosen.

ii) **Exclusion criteria:** Unwillingness, unavailability in the first, second and third visits, illness, taken medicine in the prior two weeks and pregnancy

3.6. Ethics

The study is ethically approved by the ethical research committee of Guru Nanak Dev University on 30\textsuperscript{th} October, 2007. The subjects were recruited in the study were pre-informed and had given their written consent.

3.7. Total Samples:

Total number of samples taken at first visit was 1923, which included 971 males and 952 females. The exclusion of subjects after second visit reduced the total samples studied to 1827, including 911 males and 916 females.

| Table 3.1. The distribution of selected clusters and households of four districts |
|---|---|---|---|
| Districts | No. of clusters | No. of households | No. of subjects |
| Gurdaspur | 28 | 279 | 846 |
| Jalandhar | 13 | 129 | 372 |
| Amritsar | 11 | 108 | 351 |
| Hoshiarpur | 9 | 84 | 258 |
| **Total** | **61** | **600** | **1827** |
3.8. Calculated variables

3.8.1. Mean Arterial Blood Pressure (MBP): The mean arterial blood pressure (MBP) is a term used to describe an average blood pressure in an individual. It is defined as the average arterial pressure during a single cardiac cycle. It was calculated for each of the three readings taken for SBP and DBP by using the formula.

\[ MBP = DBP + \frac{(SBP - DBP)}{3} \] (Perusse et al., 1989).

3.8.2. Pulse Pressure (PP): It is the difference between systolic blood pressure and diastolic blood pressure. Pulse Pressure is calculated by given formula:

\[ PP = SBP - DBP \]

3.8.3. Body Mass Index (BMI): Body Mass Index (BMI) is a number calculated from a person's weight and height. BMI provides a reliable indicator of body fatness for most people and is used as a measure of obesity. BMI was calculated by dividing weight of the subject in kilogram by square of his height in meter.

\[ BMI = \frac{\text{weight in kilogram}}{(\text{height in meter})^2} \]

3.8.4. Waist-to-Hip Ratio (WHR): The waist to hip ratio (WHR) has been used as an indicator or measure of the health of a person, and the risk of developing serious health conditions. It is a simple and useful measure of fat.

WHR is calculated by the formula given below:

\[ WHR = \frac{\text{Waist circumference in cm}}{\text{Hip circumference in cm}} \]
3.9. Classifications

Table 3.2. Classification of blood pressure phenotypes in adults (JNC 7, 2004)

<table>
<thead>
<tr>
<th>Classification</th>
<th>SBP(mm Hg)</th>
<th>DBP(mm Hg)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>&lt;120</td>
<td>&lt;80</td>
</tr>
<tr>
<td>Pre-hypertension</td>
<td>120-139</td>
<td>80-89</td>
</tr>
<tr>
<td>Hypertension Stage I</td>
<td>140-159</td>
<td>90-99</td>
</tr>
<tr>
<td>Hypertension Stage II</td>
<td>≥160</td>
<td>≥100</td>
</tr>
</tbody>
</table>

Table 3.3. Classification of blood pressure for children (age <18 years) (Brookes, 2004)

<table>
<thead>
<tr>
<th>Classification</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>&lt;90th percentile</td>
</tr>
<tr>
<td>Pre-hypertension</td>
<td>90th-95th percentile or 120/80 mm Hg</td>
</tr>
<tr>
<td>Hypertension Stage I</td>
<td>95th-99th percentile + 5 mm Hg</td>
</tr>
<tr>
<td>Hypertension Stage II</td>
<td>&gt;99th percentile + 5 mm Hg</td>
</tr>
</tbody>
</table>

Table 3.4. Classification of Body Mass Index (WHO, 2000)

<table>
<thead>
<tr>
<th>Classification</th>
<th>BMI(Kg/m²)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Underweight</td>
<td>≤18.5</td>
</tr>
<tr>
<td>Normal</td>
<td>18.5-24.9</td>
</tr>
<tr>
<td>Overweight</td>
<td>25.0-29.9</td>
</tr>
<tr>
<td>Obese I</td>
<td>30.0-34.9</td>
</tr>
<tr>
<td>Obese II</td>
<td>34.0-39.9</td>
</tr>
<tr>
<td>Obese III</td>
<td>≥40.0</td>
</tr>
</tbody>
</table>

The age group below 18 years of age is not specified in WHO criteria. So to overcome this problem the present study used the international standard parameters of obesity and for thinness (Normal: <25.0 kg/m², Overweight: 25.0-29.9 kg/ kg/m², obese: ≥30.0 kg/ kg/m²) (Cole et. al., 2000).
### Table 3.5: Classification of Waist Circumference (Yalcin *et al.*, 2005)

<table>
<thead>
<tr>
<th>Classification</th>
<th>Waist circumference</th>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>No risk</td>
<td>&lt;94</td>
<td></td>
<td>&lt;80</td>
</tr>
<tr>
<td>Medium risk</td>
<td>94-101</td>
<td></td>
<td>80-87</td>
</tr>
<tr>
<td>High risk</td>
<td>&gt;101</td>
<td></td>
<td>&gt;87</td>
</tr>
</tbody>
</table>

### Table 3.6. Classification of Waist to Hip Ratio (Yalcin *et al.*, 2005)

<table>
<thead>
<tr>
<th>Classification</th>
<th>WHR</th>
<th>Men</th>
<th>Women</th>
</tr>
</thead>
<tbody>
<tr>
<td>No risk</td>
<td>&lt;0.90</td>
<td></td>
<td>&lt;0.80</td>
</tr>
<tr>
<td>Medium risk</td>
<td>0.90-1.0</td>
<td></td>
<td>0.81-0.85</td>
</tr>
<tr>
<td>High risk</td>
<td>&gt;1.0</td>
<td></td>
<td>&gt;0.85</td>
</tr>
</tbody>
</table>
3.10. Districtwise collection of data

3.10.1. Gurdaspur District

The Gurdaspur district (Figure 3.2) is the northern most district of Punjab state. It falls in the Jalandhar division and is sandwiched between river Ravi and Beas. The district lies between north-latitude 31° 36' and 32° 34' and east longitude 74° 56' and 75° 24' and shares common boundaries with Kathua district of Jammu and Kashmir state in the north, Chamba and Kangra districts of Himachal Pradesh in the north-east, Hoshiarpur district in the south-east, Kapurthala district in the south, Amritsar district in the south west and Pakistan in the north west.

**Figure 3.2:** Map of Gurdaspur district highlighting sampled areas
**Data Distribution:** The study sample consisted of 279 families with a total of 846 individuals comprising 414 males and 432 females. The distribution of the families in different sub areas of Gurdaspur was as following (Figure 3.3):

![Pie chart showing distribution of families in Gurdaspur district](image)

**Figure 3.3:** Pie chart showing distribution of families in Gurdaspur district
Jalandhar (Figure 3.4) was the capital of Punjab since India's independence (1947) until Chandigarh was built in 1953. Jalandhar is situated at 71° 31’ East and 30° 33’ North at a distance of 146 kms from state capital Chandigarh. It is at a distance of 350 kms from Delhi on Delhi-Amritsar highway. It is surrounded by Ludhiana district in east, Kapurthala in west, Hoshiarpur in north and Firozpur in south. It is well connected by road and train. Nearest Airport is Raja-Sansi Airport, Amritsar at a distance of 90 kms. Jalandhar is located on the intensively irrigated plain between the Beas and Sutlej rivers. The city, which has major road and rail connections, is a market for agricultural products. Manufactures include textiles, leather goods, wood products, and sporting goods.
**Data Distribution:** The study was conducted on 129 families comprising a total of 372 individuals. The latter included 201 males and 171 females. Distribution of families from different areas was as following (Figure 3.5):

![Pie chart showing distribution of families in Jalandhar district](image)

**Figure 3.5:** Pie chart showing distribution of families in Jalandhar district
3.10.3. Amritsar District

Amritsar district (Figure 3.6), situated in northern Punjab state of northwestern India, lies about 15 miles (25 km) east of the India-Pakistan border. It is situated at 31° 37´ N, 74° 55´ E with an average elevation of 234 metres (768 ft). Amritsar is an important city in Punjab and is a major commercial, cultural, and transportation centre. It is also the centre of Sikhism and the site of the Sikh’s principal place of worship (Golden temple).
**Data Distribution:** A total of 108 families were taken as study sample and constituted 351 individuals comprising of 174 males and 177 females. The distribution of the families in different sub areas of Amritsar was as following (Figure 3.7):

![Pie chart showing distribution of families in Amritsar district.](image)
3.10.4. Hoshiarpur District

Hoshiarpur district (Figure 3.8) falls in the Jalandhar Revenue Division and is situated in the Bist Doab, Doaba region of the State. The district is sub-mountainous and stretches of river Beas in the north-west. It is located at north latitude 31° 32’ and east longitude 75° 57’. It shares common boundaries with Kangra and Una districts of Himachal Pardesh in the north east, Jalandhar and Kapurthala districts (interspersed) in south-west and Gurdaspur district in the north-west.

Figure 3.8: Map of Hoshiarpur district showing highlighted sampled areas
**Data distribution:** A total of 258 individuals of 84 families were taken for the study. The individuals included 122 males and 136 females. Distribution of families from different areas was as following (Figure 3.9):

![Pie chart showing distribution of families in Hoshiarpur district](image)

**Figure 3.9:** Pie chart showing distribution of families in Hoshiarpur district
3.11. **Distribution of total data through family-wise**

The total samples can be categorized on the basis of total number of families having various number of individuals (1-8) in the four districts (Figure 3.10)

**Table 3.7: The family-wise distribution of the subjects**

<table>
<thead>
<tr>
<th>No. of individuals</th>
<th>No. of Families</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Gurdaspur</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>97</td>
</tr>
<tr>
<td>3</td>
<td>97</td>
</tr>
<tr>
<td>4</td>
<td>57</td>
</tr>
<tr>
<td>5</td>
<td>17</td>
</tr>
<tr>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
</tr>
<tr>
<td>Total Families</td>
<td><strong>279</strong></td>
</tr>
</tbody>
</table>

**Figure 3.10.** Graphical representation of the district-wise distribution of the subjects
3.12. **Generation-wise Distribution of Data**

As the study is familial in nature, the total sampled individuals have been categorized into three generations; offspring, parental and granparental. The three generations have been further divided into males and females. The distribution of the subjects based on the number of individuals in each generation is as follows:

![Diagram showing generation-wise distribution of subjects]

**Figure 3.11**: Generation-wise distribution of subjects (figure in parenthesis indicates the number of individuals)
3.13. **Overall distribution of total samples**

The total data collected has been summarized in the following figure:

**Figure 3.12:** Overall distribution of total data
3.14 STATISTICAL ANALYSIS

At least six basic statistical books have been consulted for present statistical data analysis (Snedecor and Cocharan, 1994; Indrayan, 2008; Daniel, 2009; Kothari, 2010; Sundaram et al., 2010; Sokal and Rohlf, 2012).

Statistics help to understand a phenomenon by confirming or rejecting a hypothesis and to be able to analyze the data sensibly, therefore the raw data needs to be processed. Different statistical methods have been employed in the present work to understand the data better and more accurately.

3.14.1 Method of scoring:

The simplest linear scoring method has been applied to quantify the different risk factors of cardiovascular disease like ‘0’ for no risk, ‘1’ for mild risk, ‘2’ for moderate risk, ‘3’ for severe risk and so on. The scoring system to measure the risk factors is very complicated and very few studies have been carried out to investigate various alternatives like geometric scoring (0,1,2,3,4…… or 0,3,6,…… and so on) and thus nothing can be stated with confidence. Nevertheless, the 0,1,2,3,4…… scoring types remain the most widely used scoring method because of their simplicity and therefore no explanation is generally required when such simple scores have been used.

For the results of many lifestyle indicators for cardiovascular diseases such as food habits, behavioral lifestyle, exercise, stress, smoking and alcohol the gold standards are not available. Therefore, the present scoring system is worth to assess the severity of risk factors to develop cardiovascular diseases and it is also assumed that this scoring system gives results that are consistent with general hypothesis related to lifestyle factors to develop cardiovascular diseases.

3.14.2 Descriptive statistics:

Descriptive statistics are used to describe the basic features of the data in a study. This provides simple summary about the sample and the measures. These are used to present quantitative data in a manageable form.

Mean: A value nearly in the centre or commonly observed could be considered as a representative of the central tendency. Most popular, measure of central tendency is arithmetic mean.
Mean, \( \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \)

The symbol \( \Sigma \) denotes summation of \( X_i \) values, \( i \) ranging from 1 to \( n \). The mean of the \( n \) values of \( x \) is denoted by \( \bar{x} \).

### 3.14.3 Variance and standard deviation:

Its magnitude depends on the extent of difference between a value and that of others. Therefore, it is convenient to compute the difference of each from the central value i.e. mean. The difference \( (x_i - \bar{x}) \) where, \( i = 1,2,\ldots,n \) is called the deviation from the mean. Some of these deviations would be positive and some would be negative. The average of squared deviations is called variance, which is also useful measure of dispersion. Algebraically, standard deviation (\( S \)) is

\[
S = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x - \bar{x})^2}
\]

### 3.14.4 Quantiles:

Quantiles in ungrouped data are obtained:

\( P^{th} \) s-tile = \((P \times n)/s\) the value in ascending order of magnitude

Where, \( n \) is the total number of the subjects.

### 3.14.5 Skewness:

It defines the departure from symmetry of the distribution. In normal distribution, which is symmetric, 50% of the values will be below the mean and remaining will be above the median. The tendency of the distribution to depart from symmetry is called skewness. When more raw values are concentrated above the mean then it is called negative skewness and if more raw values are concentrated below the mean then it is called positive skewness. If the values are symmetrically distributed on either sides of the mean/median then there is no skewness and it is called normally distributed.

**Measurement:** Two quartiles \( Q_1 \) (first quartile) and \( Q_3 \) (third quartile) will be equidistant from the median in the case of normal distribution. Any deviation from this is taken as a relative measure of skewness.
Coefficient of skewness (Sk)

\[ Sk = \frac{(Q_3 - M) - (M - Q_1)}{(Q_3 - Q_1)} \]

Where, M is the median, if this value is equal to zero, then it is assumed symmetric distribution and there is no skewness, and if this value is positive then the distribution is called positively skewed and if it is negative then it is called negatively skewed. The extent of skewness will be high or low depending upon the magnitude of this value.

3.14.6 Kurtosis

The other kind of deviation from normal distribution is called kurtosis. It gives the peakedness of the curve. It gives three types of peaks. If the peakedness is high as compared to normal curve, then it is called leptokurtic and if it is low then it is called platykurtic. If the peakedness is moderate then it is called normokurtic or mesokurtic which is also normal curve.

**Measurements:** A simple measure of kurtosis (K)

\[ K = \frac{Q_3 - Q_1}{2(P_{90} - P_{10})} \]

Where, \( Q_3 \) = third quartile; \( Q_1 \) = first quartile; \( P_{90} \) = 90th percentile; \( P_{10} \) = 10th percentile. If K is equal to 0.263, the curve is normokurtic, if it is greater than 0.263, it is platykurtic and if it is less than 0.263, it is leptokurtic. The extent of kurtosis is assumed based on the magnitude of this statistics.

3.14.7 Pie chart

Pie (П) is a mathematical constant defined as the ratio of the circumference of a circle to its diameter. In the pie chart, a circle (total 360°) is divided into sectors with areas proportional to the frequencies or the relative frequencies of the categories of the variables.

3.14.8 Bar diagram

Different categorical data have been represented by bar diagram. In this diagram, the categories of the variables on the X-axis and frequencies on the Y-axis have been presented.
3.14.9 Spot map
Maps are shown for spatial distribution of sample collection area in different districts. Concentration of dots in certain areas indicates that the concentration of sampled subjects of Ramdasia community.

3.14.10 Significance level
The decision as to which values go into the rejection region and which ones go into the non-rejection region is made on the basis of desired level of significance, designated by $p$. The term level of significance reflects the fact that hypothesis tests are called significance test and computed values of test statistics that fall in a rejection region is said to be significant. Therefore, the level of significance, $p$ is the probability and, in fact, is the probability of rejecting a true null hypothesis. The probability values less than or equal to 0.05 (two-tailed) were considered to be significant in the present thesis. In the present analysis, although the statistical software itself provided given $p$-values with respect to the test, the more frequently encountered values are 0.05, 0.01 and 0.001.

3.14.11 Correlation:
To understand the dimension of relationship of blood pressure phenotypes (SBP, DBP and MBP) as the dependent variables with the set of anthropometric variables Karl Pearson’s product moment correlation coefficients have been calculated. The correlation coefficient is denoted by ‘$r$’ and is expressed as follows:

$$(\text{denote two variables as } x \text{ and } y)$$

$$r = \frac{\sum_{i=1}^{n} x_i y - \bar{x} \bar{y}}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

**Interpretation of $r$:** A value close to +1 indicates a perfect or near perfect positive relationship between the two quantitative variables and a value close to -1 indicates perfect negative relationship between them. A value close to zero indicates that the two quantitative variables are not linearly related.

**Square of the correlation coefficient ($r^2$):** The square of the correlation coefficient ($r^2$) is the proportion of variance in one variable can be explained by the other. If $r^2$ is close
to one, it implies that the total variance in one variable can be explained by other, for example if \( r^2 = 0.328 \) for SBP with age, which implies that 32.8% of variations in SBP can be explained by age.

**Statistical significance of \( r \):** A correlation coefficient \( r \) is called statistically significant when the probability of it being zero in the population is less than 0.05 or any predetermed level of significance. This is tested by referring the criteria

\[
t = \frac{r \sqrt{n-2}}{\sqrt{1-r^2}} \quad \text{with (n-2)df}
\]

**Interclass/familial correlation:** Relatives within the family or extended family are likely to be more similar to one another than individuals belonging to different families. This phenomenon is known as familial aggregation. This familial correlation coefficient is the measure of the degree of consistency or conformity between members belonging to the same subgroup. The different familial correlations with respect to different cardiovascular risk factors such as BMI, WHR, biceps skinfold, pulse pressure, SBP, DBP and MBP in between spouse, father-male offspring, father-female offspring, mother-male offspring, mother-female offspring, brother-sister, grandfather-male offspring, grandfather-female offspring, grandmother-male offspring and grandmother-female offspring have been calculated.

### 3.14.12 Heritability Estimates

The total phenotypic variance of a given trait can be partitioned into genetic and environmental components. The proportion of phenotypic variance in a trait that is attributable to additive genetic factors is referred to as heritability \( (h^2) \) of the trait (Falconer, 1989). Heritability can be estimated from the components of variance using the genetic information contained in the pedigrees (Hopper and Matthews, 1982; Lange and Boehnke, 1983; Arya et al. 2002).

The estimation of heritability has been done from the degree of resemblance between relatives. The regression of offspring on parents and correlation has been expressed in the terms of heritability. Therefore, the relationship between these two is as follows:

\[
h^2 = b/r
\]

where \( r \) is the coefficient of the additive variance in the co-variance and \( b \) is the regression coefficient of offspring on parents.
3.14.15 Regression Analysis

To quantify the association between dependent variables (SBP, DBP and MBP) and sets of independent anthropometric and lifestyle variables have been calculated through regression analysis. However, the variables of cardiovascular disease risk factors such as SBP and DBP can be considered the ‘predicted variable’ and anthropometric and lifestyle variables can be considered as ‘predictor variables’, respectively.

**Linear regression:** More precisely, if x and y are two related variables, then linear regression analysis helps us to predict the value of y for a given value of x or vice versa. By linear regression, we mean models with just one independent and one dependent variable. The variable whose value is to be predicted is known as the dependent variable and the one whose known value is used for prediction is known as the independent variable.

In this study how SBP, DBP and MBP are affected by different anthropometric and lifestyle variables separately likewise one dependent (SBP/DBP/MBP) and one independent (age/height/weight/smoking…). Mathematically, the equation for straight line is,

\[ y = a + bx \]

Where, \( y \) is the value of the dependent variable (SBP/DBP) and \( x \) is the value of the independent (anthropometric or lifestyle) variables, ‘\( b \)’ is the slope of the line and ‘\( a \)’ is the point where the line intercepts the vertical axis. The slope is called the regression coefficient. The difference between the observed value of the dependent variable \( y \) and its expected value by the regression model is called the residual. For the data points falling above the regression line then the residual is positive and for those falling below the line, then the residuals will be negative. The expression of the slope (\( b \)) and the intercept (\( a \)) are obtained as:

\[
\begin{align*}
    b &= \frac{\text{sum of } xy - \{(\text{sum } x)(\text{sum } y)n\}}{\text{sum } x^2(\text{sum } x)^2/n} \\
    a &= [(\text{sum } y)/n] - [b(\text{sum } x)/n]
\end{align*}
\]

\[
\begin{align*}
    b &= \frac{\sum_{i=1}^{n} x_i y_i - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)/n}{\sum_{i=1}^{n} x_i^2(\sum_{i=1}^{n} x_i)^2/n}
\end{align*}
\]
Multiple linear regression:
It is also studied that how SBP, DBP and MBP have been affected by age, BMI, waist circumference, WHR, lifestyle variables, etc together. This relationship is called multiple regression, in this setup the number of independents is more than one. Therefore, in this analysis, the regression of (dependent variables SBP/DBP) on a set of independent variables \((x_1, x_2, \ldots, x_n)\) like age+BMI+WHR\(\ldots\ldots\ldots\) is studied. This analysis has helped in finding out whether the different anthropometric and lifestyle variables are related to SBP/DBP or not and if related then which all of them are significantly related indicating the order of their importance. Furthermore, this analysis also helps to predict the dependent variable (SBP/DBP) based on the significantly contributing independent variables (anthropometric and lifestyle), indicating the error observed in the prediction.

In the multivariate linear regression analysis, the model is

\[
y = a + b_1x_1 + b_2x_2 + \ldots + b_nx_n
\]

Where,\(x_1, x_2, \ldots, x_n\) are the independent variables; \(b_1, b_2, \ldots, b_n\) are called the regression coefficients of \(x_1, x_2, \ldots, x_n\), respectively. \(x_1, x_2, \ldots, x_n\) are also called as predictor variables and \(y\) is also referred to as explained or predicted variables.

**Interpretation of slope and intercept:** The slope or regression coefficient \(b\) determined the average change in the dependent variable for a unit change in the independent variable. The intercept is a constant value that is unrelated to the value of the independent variables. Statistical significance is a reasonable assurance that the regression coefficient is not zero and the corresponding regressor have some contribution to determine the value of the dependent variable.

**Goodness of fit and \(R^2\) (coefficient of determination):** The statistical significance of regression model indicates the model is not useless. It may be a good fit. The ‘goodness of fit’ is denoted by \(R^2\):

\[
R^2 = 1 - \frac{SSE(\text{sum of squares due to error})}{SST(\text{total sum of squares})}
\]

The quantity \(R^2\) is interpreted as the proportion sum of squares of dependent variable explained by the regression. The larger \(R^2\) is the better fit. \(R^2\) is also used to compare one model to other model (stepwise regression analysis). The model with higher \(R^2\) is
considered a better fit. Therefore, a model is considered to be good when it contained a small number of regressor but gives a maximum large value.

**Stepwise regression analysis:**
The most widely used strategy for selecting independent variables for a multiple regression model is the stepwise procedure. The procedure consists of a series of steps. At each step of the procedure each variable then in the model is evaluated to see if, according to specified criteria, it should remain in the model. The nature of the stepwise procedure is such that, although a variable may be deleted from the model in one step, it is evaluated for possible re-entry into the model in subsequent steps.

Stepwise regression analysis has been done in the present study to specify the regressors only those that contribute significantly among large number of anthropometric and lifestyle regressors. Maximum 5 model for anthropometric regressor (age, BMI, waist circumference, WHR and pulse rate) and 8 models for lifestyle regressors (occupation, income, education, food habits, behavioral lifestyle, exercise, smoking and alcohol) have been included. In this procedure variables were entered continued to remain in the model and a variable once deleted remains excluded. The utility of variables also entered and deletes those that become insignificant after addition of one of the other explanatory variables or enters one of the already deleted variables. The procedure stops when neither deletion nor addition significantly alters $R^2$.

**Logistic Regression Analysis:**
Logistic regression is typically suitable in the situation where the response or dependent variable is dichotomously observed i.e. response is binary (yes/no) in all the cases the dependent variable ($y$) is given a value zero for negative response or 1 for positive response and nothing in between. Statistically, the dependent is the proportion or the probability of subjects providing a positive response. The functional form of the relationship between logit (logistic integral transformation) and its possible predictor is called logistic model. The equation is

$$y = b_0 + b_1 x_1 + b_2 x_2 + \ldots \ldots b_k x_k$$
Where, \( k \) is the predictor and the equation is called logistic regression equation. The values of \( b_0, b_1, b_2, \ldots, b_k \) are the estimates of corresponding parameters which is also called logistic coefficients.

The Logistic Regression model is widely used in epidemiological research. This logistic model is frequently used to identify the probability (interpreted as the risk) that an individual will acquire a disease during some specified time period during which he or she is exposed to a condition (called a risk factor) known to be or suspected of being associated with the disease.

### 3.14.16 Wald chi square statistics

Wald test has been used in great variety of different models of dichotomous variables. Whenever a relationship within or between data items can be expressed as statistical model with parameters to be estimated from a sample then the wald test can be used to test the true value of the parameters based on sample estimate. The wald statistics in the present logistic analysis has been used through maximum likelihood ratios of the parameters.

### 3.14.17 Student’s t-test in independent sample setup

Homogeneity of variances in the two samples is assumed and accepted

\[
t = \frac{|\bar{x}_1 - x_2|}{S \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}
\]

Where, \( S \) is pooled variance given by,

\[
S = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 1}}
\]

Where, \( s_1 \) and \( s_2 \) are the standard deviation of two samples (\( n_1 \) and \( n_2 \)). \( \bar{x}_1 \) and \( \bar{x}_2 \) are the respective means of the two samples (\( n_1 \) and \( n_2 \)), degrees of freedom = \( n_1 + n_2 - 1 \)

### 3.14.18 Odds ratio

The odds ratio is used in the present study to show the strength of association of pre-hypertension and hypertension with different anthropometric and lifestyle indicators in different generations. The structure of study for odds ratio (OR) in independent samples is given below (Indrayan, 2008)
### Materials and Methods

#### Antecedent

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Present</th>
<th>Absent</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Present (cases)</td>
<td>$\pi_{11}$</td>
<td>$\pi_{12}$</td>
<td>$n_1(1)$</td>
</tr>
<tr>
<td>Absent</td>
<td>$\pi_{21}$</td>
<td>$\pi_{22}$</td>
<td>$n_2(1)$</td>
</tr>
<tr>
<td>Total</td>
<td>$\pi_1$</td>
<td>$\pi_2$</td>
<td>$n$</td>
</tr>
</tbody>
</table>

The odds of the presence of an antecedent among cases are $\pi_{11}/\pi_{12}$ and among controls are $\pi_{21}/\pi_{22}$. Thus, the odds ratio (OR) is

$$R = \frac{\pi_{11}/\pi_{22}}{\pi_{21}/\pi_{22}} = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}}$$

$$OR = \frac{O_{11} \times O_{22}}{O_{12} \times O_{21}}$$

Since the numerator is the product of elements in the leading diagonal in above table and the denominator is that of the elements in the other diagonal, OR is also sometimes called the Cross Product Ratio.

The interpretation of odds ratio is similar to that of relative risk. An OR = 2 means that the presence of the antecedent is twice as common among the cases as in the controls.

It can be shown that the odds ratio approximates the relative risk fairly well when the outcome of interest is rare, say less than 5% in the target population.

Confidence Interval (CI) for OR (independent sample) (Indrayan, 2008):

OR is a ratio and its natural logarithm (ln) becomes a linear function. The distribution of OR can be shown to be highly skewed but lnOR has a nearly Gaussian pattern for large $n$. It has been established for large $n$ that the estimated

$$SE(lnOR) = \sqrt{\frac{1}{O_{11}} + \frac{1}{O_{12}} + \frac{1}{O_{21}} + \frac{1}{O_{22}}}$$

95% CI for OR: $\exp[lnOR \pm 1.96 \times (lnOR)]$

Where exp is the exponent on Naperian base e.
3.14.19 Confidence limit (CI) interval estimation

In CI estimation a lower limit and upper limit have been computed from sample value and it can be said with certain amount of confidence that the true value of the parameter will lie within these limits. These limits are called CI or interval estimates. These limits are computed from the estimated value and its standard error (SE) value such as mean/true value – C*SE and mean/true value + C*SE, where C=confidence coefficient; SE=SD/SE = SD/√n; n=sample size

The present study has been concentrated on 95% CI level, in this case C=1.96

3.14.20 Sensitivity, specificity and likelihood ratio

A valid diagnostic indicator could correctly detect the presence as well as absence of the disease. Some indicators are more valid than others. Sensitivity and specificity are the two components of the validity of the test that measures its inherent goodness. In the present study sensitivity, specificity and likelihood ratios for different anthropometric and lifestyle indicators such as BMI, waist circumference, WHR, food habits, behavioral lifestyle, exercise, smoking and alcohol to detect pre-hypertension and hypertension has been calculated to determine the best indicator among them in different generations. Both sensitivity and specificity can be converted to percentage by multiplying with 100 for best illustration.

Schematic structure for calculation: These tests can be summarized in following 2×2 contingency table

<table>
<thead>
<tr>
<th>Prediction by the diagnostic test</th>
<th>Truth</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Disease present</td>
<td>Disease absent</td>
</tr>
<tr>
<td>Disease present</td>
<td>a</td>
<td>b</td>
</tr>
<tr>
<td>Disease absent</td>
<td>c</td>
<td>d</td>
</tr>
<tr>
<td>Total</td>
<td>a+c</td>
<td>b+d</td>
</tr>
</tbody>
</table>

The test can lead to the conclusion that the disease is present when the disease is truly present as well as when it is truly absent. Similarly, the test result can lead to a negative
Materials and Methods

collection when there is truly no disease and also when the disease is present. In above

above table frequency $a$ indicates the number of diseased persons who had a positive result by

the diagnostic test and, therefore, the remaining, $c$, diseased persons test results are

negative. Similarly, $d$ indicates the number of normal (no disease) individuals who

showed a negative result by the test and $b$ is the number of normal persons showing a

positive test result.

The proportion $a/(a+c)$ is called the sensitivity which shows how likely a person with

the disease is shown as positive by the indicator test. The proportion $d/(d+b)$ is called

the specificity of the test which shows what is the chance that an individual who does

not have the disease is shown as negative by the test also. Thus, sensitivity of the
diagnostic test is the probability of picking up the disease when it is present and

specificity is probability of showing a normal result when the disease is absent.

The presence and absence of disease is denoted by $D+$ and $D-$ respectively. Similarly, the positive and negative results of an indicator test are denoted by $T+$ and $T-$

respectively.

Sensitivity = probability (test result is positive, given that the disease is truly present)

$= P (T+|D+)$

Sensitivity is the True Positive Fraction (TPF) of the total positives declared by the test and is analogous to the ‘power’ or $(1-\beta)$ of a statistical test of hypothesis (probability of declaring a significant difference between two treatments by the statistical test, when infact a significant difference exists).

Specificity = probability (test result is negative, given that truly there is no disease)

$= P (T-|D-)$

Specificity is the True Negative Fraction (TNF) of the total negatives declared by the test.

In terms of the frequencies,

Sensitivity = $a/(a+c)$

Specificity = $d/(b+d)$
Sensitivity and specificity can be combined to provide another useful measure of the inherent goodness of test. Consider the chance that the test is positive in those who have the concerned disease and the chance that the test is positive in those who do not have the disease. The ratio of these two chances is called likelihood ratio of a positive test result and denoted by LR+ and similarly, likelihood ratio for negative test result is denoted by LR-.

LR of a positive test (LR+) is defined as the ratio of True Positive Fraction (sensitivity) to the False Positive Fraction (1-specificity).

\[
LR^+ = \frac{\text{sensitivity}}{1-\text{specificity}} = \frac{a/(a+c)}{b/(b+d)}
\]

LR of a negative test (LR-) is defined as the ratio of False Negative Fraction (1-sensitivity) to the True negative Fraction (specificity).

\[
LR^- = \frac{1-\text{sensitivity}}{\text{specificity}} = \frac{c/(a+c)}{d/(b+d)}
\]

### 3.14.21 Youden’s Index (J):

A diagnostic test is done based on a continuous measurements, a range of different decisions thresholds or cut-off values may be investigated in order to decide which cutoff points should be used to discriminate between patients according to the outcome. In practice the sensitivity and the specificity may not be regarded as equally important. It is desirable to choose the test that has high values for both sensitivity and specificity. However, if no judgement is made between the two then Youden’s index (J) may be chosen as an appropriate cut-off. The Youden’s index has been calculated through following formula:

\[
J = [\text{sensitivity} + \text{specificity}] - 1
\]

The maximum value of J is 1 when the test is perfect and the minimum value is usually zero when the test has no diagnostic value.

### 3.14.22 Univariate Analysis of Variance F-test ANOVA:

To test the equality of the variances of blood pressure phenotypes with respect to different anthropometric and lifestyle parameters are validated through ANOVA F-test.

\[
F = \frac{MSB \ (mean \ square \ between \ groups)}{MSE \ (mean \ square \ between \ errors)}
\]
The primary purpose of ANOVA in this study is to test the null hypothesis of equality of means of dependent (blood pressure phenotypes) and independent variables (anthropometric and lifestyle). The statistical packages provided the \( p \)-value of F to access the significant differences of variances.

3.1.4.23 Principal Component Factor Analysis (PCFA):

Factor analysis is used to uncover the latent structure (dimensions) of a set of variables. It reduces attribute space from a larger number of variables to a smaller number of factors and as such is a "non-dependent" procedure. Factor analysis could be used for any of the following purposes:

i) To reduce a large number of variables to a smaller number of factors.

ii) To establish that multiple tests measure the same factor, thereby giving justification for administering fewer tests.

iii) To validate a scale or index by demonstrating that its constituent items load on the same factor, and to drop proposed scale items which cross-load on more than one factor.

iv) To select a subset of variables from a larger set based on which original variables have the highest correlations with the principal component factors.

v) To create a set of factors to be treated as uncorrelated variables as one approach to handling multi-collinearity in such procedures as multiple regression.

vi) To identify clusters of cases and/or outliers.

vii) To determine network groups by determining which sets of people cluster.

Factor analysis is part of the general linear model (GLM) family of procedures and makes many of the same assumptions as multiple regressions: linear relationships, interval or near-interval data, untruncated variables, proper specification (relevant variables included, extraneous ones excluded), lack of high multicollinearity, and multivariate normality for purposes of significance testing. Factor analysis generates a table in which the rows are the observed raw indicator variables and the columns are the factors or latent variables which explain as much of the variance in these variables as possible. The cells in this table are factor loadings, and the meaning of the factors must be induced from seeing which variables are most heavily loaded on which factors.
There are several different types of factor analysis, however, with the most common being principal components analysis (PCA), which is preferred for purposes of data reduction in the present study.

**Factor loadings:** The factor loadings, also called component loadings in PCA, are the correlation coefficients between the variables (rows) and factors (columns). Analogous to Pearson's r, the squared factor loading is the percent of variance in that indicator variable explained by the factor. To get the percent of variance in all the variables accounted for by each factor, add the sum of the squared factor loadings for that factor (column) and divide by the number of variables. (Note the number of variables equals the sum of their variances as the variance of a standardized variable is 1.) This is the same as dividing the factor's eigenvalue by the number of variables.

**Interpreting factor loadings:** By one rule of thumb in confirmatory factor analysis, loadings should be 0.7 or higher to confirm that independent variables identified a priori are represented by a particular factor, on the rationale that the 0.7 level corresponds to about half of the variance in the indicator being explained by the factor. However, the 0.7 standard is a high one and real-life data may well not meet this criterion, which is why some researchers, particularly for exploratory purposes, will use a lower level such as 0.4 for the central factor and 0.25 for other factors (Raubenheimer, 2004). Hair et al. (1998) call loadings above 0.6 "high" and those below 0.4 "low". In any event, factor loadings must be interpreted in the light of theory, not by arbitrary cutoff levels.

**Factor, component, pattern, and structure matrices:** In SPSS (Statistical Package for Social Sciences), the factor loadings are found in a matrix labeled Factor Matrix if common factor analysis is requested, or in one labeled Component Matrix if PCA is requested.

**Sum of the squared factor loadings:** The sum of the squared factor loadings for all factors for a given variable (row) is the variance in that variable accounted for by all the factors, and this is called the communality. In a complete PCA, with no factors dropped, this will be 1.0, or 100% of the variance. The ratio of the squared factor loadings for a given variable (row in the factor matrix) shows the relative importance of the different factors in explaining the variance of the given variable. Factor loadings are the basis for imputing a label to the different factors.
Communality ($h^2$): It is the squared multiple correlation for the variable as dependent using the factors as predictors. The communality measures the percent of variance in a given variable explained by all the factors jointly and may be interpreted as the reliability of the indicator.

Low communality:
When an indicator variable has a low communality, the factor model is not working well for that indicator and possibly it should be removed from the model. Low communalities across the set of variables indicate the variables are little related to each other. However, communalities must be interpreted in relation to the interpretability of the factors. A communality of 0.75 seems high.

Eigen values: The eigen value for a given factor measures the variance in all the variables which is accounted for by that factor. The ratio of eigen values is the ratio of explanatory importance of the factors with respect to the variables. If a factor has a low eigen value, then it is contributing little to the explanation of variances in the variables and may be ignored as redundant with more important factors.

Varimax rotation: In the present analysis varimax rotation has been used. It is an orthogonal rotation of the factor axes to maximize the variance of the squared loadings of a factor (column) on all the variables (rows) in a factor matrix, which has the effect of differentiating the original variables by extracted factor. Each factor will tend to have either large or small loadings of any particular variable. A varimax solution yields results which make it as easy as possible to identify each variable with a single factor.

3.14.24 Software used in the study: All the statistical analyses were done by using two softwares, Statistical Package for Social Sciences (SPSS v.17.0, Chicago, IL, USA) and SYSTAT 12 (SYSTAT software, Inc. CA, USA).