Queen's University M.Sc Epid(Biostatistics) Final Project
Ontario Health Study Practicum

Brianne Wood
5490399
Supervisor: Dr. Dongsheng Tu, NCIC CTG

Submitted: August 13, 2011

Contents

1 Introduction

The Ontario Health Study (OHS) is a population-based, longitudinal cohort study that will include a large geographically and ethnically diverse sample of Ontario residents over the age of 18. By investigating the interaction of social, genetic and environmental risk factors of health and disease, the OHS will provide an in-depth assessment of the long-term health conditions of Ontario adults and their families. The number of Canadians affected by cancer and other chronic health conditions is escalating; with the increasing financial and social costs on the Ontario health care system, a greater understanding of causes of these diseases will lead to better prevention, treatment and overall health care efforts. Furthermore, the OHS will look for genetic associations with the identified risk factors the "causes of the causes", which can be combined with the epidemiological discoveries to understand the pathways of disease. Funded by four dominant government agencies, Cancer Care Ontario, Ontario Institute of Cancer Research, Public Health Ontario, and Canadian Partnership Against Cancer. The OHS actually consists of more than two hundred clinicians in addition to researchers within thirty working groups, who have collaborated in the design and implementation of the study. The close ties with these government agencies will allow for more efficient translation of research discoveries into health care practice and policy. Ultimately, the OHS research will provide an integrated platform for etiological research, translation medicine, improved prevention, diagnosis and treatment efforts, cost-effective health promotion and intervention programs, and overall, a more effective health care system that will benefit all Ontarioans.[1]

The OHS is part of a larger initiative, the Canadian Partnership for Tomorrow Project (CPTP), which involves five regional cohorts: OHS, Atlantic Partnership for Tomorrow's Health, Alberta Tomorrow Project, CartaGene Quebec, BC Generations project. Though initial establishment of these cohorts will require a large investment of time, money and labour, the resulting quality of evidence and the impact of the research far exceed these initial costs. Parallels can be drawn to the Framingham Heart Study, whose findings have revolutionized epidemiological studies and clinical practice. With the implementation of the study taking place more than sixty years ago, the Framingham Heart Study has since had over 1,200 articles published, [2] the majority of which were published in the last decade using information from the original cohort of the 1950's. Similar to Framingham, the OHS combines methods of traditional epidemiology and molecular epidemiology to generate a detailed understanding of the pathways between exposure and disease.

This type of "integrated" study, one which combines genetic information with broader epidemiological variables, requires a balance between accruing large sample size for a sufficiently powered study and still having the ability to obtain high quality, detailed data without significant loss-to-follow-up. The OHS has a sampling frame of approximately 9.5 million adults from Ontario, from which they hope to have at least 100,000 (ideally more) Ontario adults complete the core questionnaire, as well possible follow-up questionnaires and biological samples, such as blood and saliva, in addition to potential linkage with health records and registries. "Mini-clinics" will be set up across Ontario, at work place institutions and other small organizations in several communities to collect these "thin" biological data. From this sample, the OHS plans to collect "dense" data on approximately 100,000 of those participants who completed the online questionnaire; these participants will be invited to an assessment centre in Toronto where more thorough physical measurements will be taken, such as spirometric and anthropometric measurements. In addition, sub-studies may be conducted on this group. For example, one sub-study will conduct Magnetic Resonance Imaging of the head and neck of approximately 5,000 subjects. Independently, the OHS has the potential to understand and address issues with its current clinical practices and health care system related to cancers and other illness. However, when pooled with the other regional projects that compose the CPTP, the potential for scientific discovery is immense. The five studies of the CPTP have harmonized the variables across the studies by incorporating validated, standardized instruments within each of their independent studies. A rigorous pooling of variables across the studies will allow the researchers to capitalize on the scientific information that is available. The harmonization of variables across the five CPTP cohorts will ensure a diverse and sufficiently powered Canadian cohort, which will be used toward the study, understanding and translation of population health in Canada. [3]

As part of the CPTP, the OHS is responsible for acquiring and maintaining ethical approval, enrolling participants in the study, data collection and custody of questionnaire, physical measurements data, and custody of biological samples collection. Consenting participants of the OHS will complete an online questionnaire regarding health and lifestyle behavior and will be asked to donate biological samples. The OHS team will actively follow-up with the participants by inviting them to participate in additional questionnaires or to provide additional biological samples. Administrative health databases and registries which could include cancer registries, provincial health insurance databases and vital statistics databases, will be used for linking study subjects with diagnoses of relevant illnesses. Participants may also receive invitations to join additional research endeavours.

The OHS baseline questionnaire will provide self-reported information on the following subdomains:

- Sociodemographic information - variables include information on ethnicity, education, family and household characteristics, employment status, income and socioeconomic status, religion and spirituality.
- General health status - variables include information on current health conditions, psychosocial and emotional well-being, reproductive history (specific to sex), and personal medical history.
- Family medical history - variables include information on previous or existing medical conditions within the family.
- Health Behaviours - variables include information on physical activity, recreation activity, tobacco use and alcohol use.
- Environmental characteristics - variables include information on sleep patterns, sunlight exposure, environmental tobacco exposure, environmental chemical exposures, housing characteristics and residential history, cell phone use.
- Dietary characteristics - variables include information on food frequency data and meal patterns.
- Occupational History - variables include information on employment history, job type, and job-related health characteristics.
- Community-level characteristics - variables include information on transportation methods, the neighbourhood built-environment, and social roles (e.g., voting, volunteering, etc), all measured at the individual level.

A core set of physical measurements will be collected from OHS participants that will support studies on the etiology of cancer and other chronic diseases. Non-invasive and efficient measurements will maximize participation rates while being simple and relatively inexpensive so that trained field staff could perform these tests accurately; the following baseline measurements will be evaluated at a field assessment centre: ankle brachial index, arm span, bone density, bioelectrical impedance, blood pressure, grip strength, hip and waist circumference, sitting and standing height, spirometry (lung function), and weight.

The last stage of baseline data collection involves obtaining a biospecimen sample from the participants. Blood and urine samples are the main products that will be sampled and stored in a central biorepository. Blood is a source of DNA, RNA and biomarkers; urine is another source of biomarkers. Triglycerides and cortisol are two examples of biomarkers that may analyzed at this stage. These biologic markers will be used to examine the role of genes in health outcomes, as well as the interaction effects of genes with the environment and lifestyle. [3]

Prior to the provincial launch of the OHS, the OHS Pilot study ran from March 2009 until July 2010, and involved the recruitment of 8,235 individuals into the study. Men and women between the ages of 35 and 70 from three distinct regions in Ontario were enrolled regardless of their health history or prevalent conditions. Three assessment centres were established in Ontario, one each in Mississauga, Owen Sound and Sudbury, representing urban, rural and northern Ontario communities. Study participants were invited to attend the local assessment centre where they completed a touch-screen self-questionnaire - an early version of the current online questionnaire, a nurse-administered questionnaire, provided the appropriate physical measurements, as well as provided blood and urine samples. The Pilot study was essential to assess the effectiveness of the baseline questionnaire, the recruitment strategy, and the logistics and acceptability of data and biospecimen collection. The OHS team was also able to evaluate the functionality and security of the information technology (IT) protocols. For the Pilot phase, the OHS received approval from the University of Toronto Research Ethics Board.

2 Objectives

This summer project consisted of three primary objectives:

1. To learn data cleaning and documentation procedures. In particular, the OHS Pilot data required cleaning as per the Data Quality protocol and concatenation rules so that the information may be analyzed.

2. To perform a literature review on currently used measures of socioeconomic status and ethnicity, with a focus on derived scores, in the context of current Canadian research.

3. To use analytic skills to perform epidemiological analyses of the OHS Pilot data, building models to investigate the association between socioeconomic status, ethnicity and various measures of anthropometry and lung function.

3 Data Cleaning

Before the OHS Pilot data can be described or analyzed, the raw input from the touchscreen Questionnaire, physical measurements and the nurse's questionnaire as entered into the assessment centre databases needs to be organized for ease of analysis. The OHS team had constructed a data quality protocol that outlined the requirements of an analytic dataset and the steps to follow to create clean variables while documenting the process.

The first section of the data quality protocol listed the general requirements of variables and data objects in the OHS Pilot data set. These properties demand that the variables have consistent representations, are unique, logical, and follow definitions from the data dictionary. The data dictionary had been constructed a priori, before any data collection had begun. In this document, the following data attributes are provided: variable names, definitions, valid range, data type, field length, unit of measurement, level of measurement, invalid value notation, and derivation methods. The second section of the data quality protocol requires modification of the data dictionary or data object if the data item's properties do not align with the standard.

The third component of the data cleaning protocol was data error detection with explicit instructions to obtaining clean data. Duplicate records or variables names were to be corrected or removed, skip patterns required identification and validation that missing values were properly defined. Questions with multiple responses required concatenation, which were identified in the protocol as well. For the continuous variables, a univariate analysis was to be performed for outlier identification and to check normality assumptions. The final portion of the data quality protocol explained logical constraints and dependency checks to ensure that the data is accurate and consistent, particularly important for the touchscreen questionnaire (e.g., the reproductive health questions were dependent on participants' responses to the gender question).

While clean data was required for conducting a specific analysis on socioeconomic status and lung function, the cleaned data sets will also be accessed by other working groups who wish to study this cohort. A collection of sociodemographic and self-reported health questions that were particularly important for deriving a socioeconomic status measure were collected in the touchscreen questionnaire, and the spirometric measurements were recorded in the physical measures data. Because of the time constraints for the summer project and the magnitude of the nurse's questionnaire, only the variables related to this project have been cleaned. The remaining data will be cleaned in the same manner following the conclusion of the summer project.

3.1 Touchscreen Questionnaire

With a total of 693 variables in the raw data set, most of the data cleaning processed involved recoding the character categorical variables into numerical factors and concatenating the multiple response question variables as per the data dictionary, or renaming the variable from the input raw data into the names defined by the data dictionary.

Often, data items were stored in a way that was not readily accessible for analysis: multiple columns were required to represent the response for one variable to ensure that logical rules and appropriate skip patterns were followed. For example, instead of the question "What is your gender?" corresponding to one variable GENDER, two Boolean variables GENDER.MALE and GENDER.FEMALE were stored and needed to be merged into a single variable SEX. Other merging instances involved time-related variables, with minutes and hours or days and weeks entered as separate columns to later be combined into one TIME_DAY variable; and, some weight- related input would need conversion into metric units (e.g., pounds to kilograms). Consequently, a large portion of the data cleaning code involved merging several related input columns into a single comprehensive variable with consistent units. Reporting to the OHS Scientific Associates, who are responsible for data quality and epidemiological methods, was imperative throughout the data cleaning process to ensure that the data cleaning was consistent with their objectives. Multiple discussion took place regarding the derivation of summary variables for multiple response questions, flagging suspect or erroneous values and interpreting response patterns.

The OHS Touchscreen questionnaire had several components that were sections of previously validated scales including the Center for Epidemiologic Studies Depression Scale (CESD Scale), Generalized Anxiety Disorder 7-item scale (GAD-7), Stressful Life Events Scale, the Lubben Social Network Scale and the International Physical Activity Long-Form Questionnaire (IPAQ). Derivation and interpretation of the composite scores were explicitly outlined in the protocols defined by the instrument developers, which simplified the coding process these scales. Writing R script for these summary variables was relatively basic; however, scoring the IPAQ long-form section for the OHS Pilot data presented several epidemiological and statistical issues. Several individuals responded "Don't know" or "Prefer not to answer" to one of the ten domains which resulted in their final physical activity measure being set to missing, by the scoring protocol of the IPAQ. This problem has not been encountered previously because previous versions of the IPAQ have not included these questions as a response. This situation is described in the following section.

3.2 The International Physical Activity Questionnaire (Long Form)

The IPAQ long-form is an instrument developed to evaluate physical activity in adults between the ages of 15 and 69. The IPAQ long-form assesses physical activity in four sub-domains: leisure time, domestic, work-related, and transport-related. Within each of these sub-domains, the questionnaire is structured to provide separate scores for walking, moderate-intensity and vigorous intensity activity. Computation of these final scores requires summation of the duration (in minutes) and frequency (in days per week) for all of the activities in all of the domains. Domain- and activity- specific scores can be calculated, and these objects can be weighted by energy expenditure defined as metabolic equivalents (MET). METs are multiples of the resting heart rate (as determined by a 60 kg person), and MET-minutes are determined by multiplying the activity specific MET score with the number of minutes performed. A total physical activity score can be computed as a continuous variable in MET-minutes/week. Categorization of activity levels involves classifying the population based on the defined cutpoints into "slow", "moderate" and "high" groups. Explicit data cleaning directions included conversion instruction, truncation rules, and removal of cases that didn't know or refused responses, or otherwise had missing values for days or time. After examining the descriptive statistics of the created global physical activity score and IPAQ categorical variable for the OHS Pilot sample, it was clear that there were several problems with these physical activity variables:

1. Of 8,235 cases in the complete OHS Pilot data set, 2,041 cases were set to "missing" - approximately one-quarter of the cohort had incomplete data and were consequently assigned to missing according to the IPAQ protocol. This large proportion of missing values suggests that this portion of the OHS Touchscreen questionnaire may be subject to bias; those who correctly completed all fields of the IPAQ long-form may be systematically different than those who were assigned missing values. For example, some of the individuals who did not provide a time estimate for a specific domain in the questionnaire may not be very active and do not want to report their low levels of activity. On the other hand, perhaps some individuals are extremely active in their weekly schedule but could not provide an estimate which they thought represented their activity levels accurately.
2. Of those individuals that had complete IPAQ information, 3,993 were classified by IPAQ categorization rules as "highly active", 1,626 were identified as "moderately active", and 575 individuals fit the criteria for "low activity". Considering the recent published results of the Canadian Health Measures Survey (CHMS) that only 15 percent of adults achieve the moderate-vigorous activity levels described by Canadian guidelines[4], these numbers may not accurately portray the truth.

These issues were presented to OHS science officers and potential reasons for these biases were discussed and problems identified. The extraordinary number of missing values for the final physical activity scores can likely be attributed to a couple reasons. The programming design of the Touchscreen questionnaire required that each question of the IPAQ long-form offer "Don't know" or "Prefer not to answer" as potential responses. If participants selected one of these options for any subdomain, their entire set of physical activity information was nullified, according to the IPAQ long-form protocol; in the original form of the IPAQ, "Don't know" and "Prefer not to answer" were not options and were consequently not issues for past investigators. The seemingly-skewed distribution of activity levels in the study population is a common issue with the IPAQ. A review of IPAQ validation studies and a quality control of the classification script confirmed that this skew is likely due to self-report bias that is inherent to the questionnaire itself, not only observed within the OHS Pilot cohort. In a study [5] that spanned 20 countries including Canada, the proportion of the population who were classified as "highly active" ranged from 21 percent to 63 percent, suggesting that the classification results of the OHS Pilot sample was not unusual.

To address the issue of missing data, I contacted other working groups and researchers within the CPTP for advice in handling missing data, particularly with the IPAQ long form. In particular, a data curator from CartaGene in Montreal, a senior researcher at the Canadian Fitness and Lifestyle Research Institute in Ottawa who has published several papers related to the IPAQ, and a research manager from the Tomorrow Project in Calgary provided their feedback on their experience with the IPAQ.

One suggestion for working with the large amounts of missing data involves looking at the descriptive statistics (e.g., mean, median, maximum, minimum, frequencies) of other characteristics: age, sex, body mass index (BMI), smoking status, level of education, diet variables, job type, ethnicity, etc. - of those with missing values in the IPAQ to compare them with those who responded completely. Then a decision would need to be made to determine if the "missing data" group differed from the complete response group in a significant way. If it was concluded that there was no notable difference between these two sets of people, then the final results of the analysis on the complete cases may be an accurate depiction of the truth.

If the subjects with missing or unknown responses were set to zero, and analysis proceeded as planned, large amounts of bias would be introduced into the data; this method would assume that the responders are similar to the non-responders, which may not be accurate. This technique was proposed in the Global Physical Activity Questionnaire (GPAQ) cleaning protocol [6], a survey which mirrors the questions, outcome measurements, and intent of the IPAQ.

The researchers contacted were in agreement that the OHS Touchscreen questionnaire's version of the IPAQ long form was problematic as the standardized and validated version of the IPAQ did not contain "Don't know" or "Prefer not to answer" as response options. Imputation - the substitution of a value where one was missing - is a possible solution to cases with only one non-response. The imputation process would require using a model that was constructed taking into account the differences between the response and non-response groups based on related characteristics, particularly those variables initially examined for differences; this model should also take into account random variation if used for multiple imputation. Deterministic imputation using regression predictions for imputed values can be conducted in R, and random error can be added to the imputed values to take into account prediction uncertainty.

Further investigation is required before continuing analysis with the IPAQ long form responses in the OHS Pilot data. The OHS team is aware of the problems with the physical activity responses from the OHS Pilot study, and may perform imputation or further missing data analysis at a later date. The current OHS online questionnaire has been adapted to accommodate issues that came to light with the OHS Pilot study, including switching from the IPAQ long form section into the IPAQ short form questionnaire. The short form questionnaire does not include domain-specific information (e.g., travel, work, yard, leisure time), but generally follows the same format with fewer questions and is more convenient for OHS participants. Because of the nature of the missing values in the OHS Pilot physical activity responses, it would be interesting to examine these cases in depth, especially with physical activity levels such a prominent theme in modern Canadian research.

3.3 Physical Measures Data

At the three assessment centres, twelve distinct physical measurement stations were set up. OHS Pilot study participants were initially questioned by a trained medical professional to evaluate potential contraindications before proceeding to the measurement phase. Following the contraindications evaluation, OHS participants were invited to provide ankle brachial blood pressure measurements, arm span measurements (for individuals who were contraindicated to providing standing height measurements), electrical bioimpedance measurements, normal blood pressure measurements, bone density measures, grip strength, waist and hip measurements, sitting and standing height, spirometry measurements, and weight (for individuals who had contraindications to bioimpedance). The data cleaning process for the physical measures data involved identification of potential outliers and erroneous values. Using descriptive statistic summaries (e.g., maximum, minimum, mean, median, 0.05 percent and 99.95 percent quantiles), box plots and histograms, cases that appeared to present suspect values were flagged and their information was documented in a data quality log, as per the Data Quality protocol. The measurements for these cases may be looked at further for future analysis. Most of the data cleaning for the physical measures data required preparing a final data set with consistent information in a logical format.

The anthropometric measurements were considered reliable the absolute difference between the anthropometric measurements taken at different stations was always less than one. The consistency checking led to one surprising result: only 1 subject had their gender incorrectly recorded during the physical assessment, but 14 physical observations contradicted the Touchscreen questionnaire gender question. Science officers from the OHS were able to search the history of these participants, and resolved to follow the gender response as per the

Touchscreen questionnaire.

It is challenging to identify erroneous physical measures values; with a sample of over 8000 participants, to individually investigate suspect values would demand a great deal of resources and may not actually be fruitful in its efforts. Consequently, the OHS team has decided to include two versions of "cleaned" data with the OHS Pilot analytic data set. The first data frame will include all the variables defined in the data dictionary for the Touchscreen questionnaire, physical measurements and Nurse's questionnaire, such that all logic and skip rules are followed and that all observations are presented in a clear and concise manner (e.g., necessary variables merged, technical survey variables removed, etc). The second data frame will be composed of the same number of variables, but with "extreme values" removed. Extreme values have been defined by the OHS as observations which are located outside 3 standard deviations from the mean. This "cleaned" data will accommodate the possible contaminants within each continuous variable's distribution. This cleaning process will primarily apply to the physical measures data. The choice of 3 standard deviations as a cutoff is an accepted method in health research as approximately 99.73 percent of the observations should be contained within 3 standard deviations of the mean under a normal distribution. Because the OHS Pilot physical measurements are assumed to be relatively robust to the normal assumption, it is likely that observations outside this interval are inaccurate. However, by providing the original data in addition to the observations with extreme values removed, future investigators can use their discretion and particular research question to determine an appropriate outlier detection method.

4 Socioeconomic Status: Literature Review

The health of a population can be attributed to several physical and social determinants and their interaction. Socioeconomic status (SES) has recently been defined as "a broad concept that refers to the placement of persons, families, households and census tracts or other aggregates with respect to the capacity to create or consume goods that are valued in our society". [7] SES, as it relates to health status and outcomes, is ultimately driven by social factors which represent an individual or group's access to resources essential for achieving and maintaining good health. [8] Income variables and SES are commonly used interchangeably, though SES captures a much broader concept of overall social position within a population; SES is generally a function of income, education, occupation and occupational prestige, neighbourhood and housing characteristics, and occasionally societal involvement. SES can be measured at the individual level, using individual income, occupation, and level of education accomplished, as well as at the area level, which can take into account property prices, housing availability, access to health-care. While there are no universally accepted measures of SES, there is ample evidence to suggest that a lower social rank or SES is directly tied to poorer health status, independent of how SES is defined.

A literature review was conducted using Summon and Google Scholar using search terms "socioeconomic status", "deprivation index", "poverty", "derived measure" against the terms "lung function", "Canada health" and "Ontario health". Further SES papers were identified by searching Canadian Institute of Health Information (CIHI), Health Canada and Statistics Canada websites using the aforementioned search terms. Papers that derived an SES index or used multiple components related to SES as a primary analysis, particularly articles from North American sources or large sample studies, were examined further. Six of the papers reviewed are summarized in Appendix II, identifying the key variables which contributed to the SES measure, comparable variables from the OHS Pilot data and analysis methodology.

In summary, different authors and organizations had diverse ideas of which components contribute to overall social rank. Statistics Canada and Canadian Institute of Health Information (CIHI) used indices that were first described by Robert Pampalon and Guy Raymond. [9] The *material* and *social* deprivation indices are closely linked with public health and welfare, and the indicators which make up the scores (e.g., income, education, employment, persons living alone, marital status and belonging to single-parent family) were selected because of their relation to a large number of health and welfare issues, their association with material or social deprivation, as well as their availability in Canadian census data. Statistics Canada and CIHI used principal component analysis to identify the linear combinations of the aforementioned predictors of deprivation. These deprivation indices were constructed such that they could be calculated at an individual level and at a regional level by using Canadian census information. These composite scores could then be divided into quintiles to classify individuals into a “high”(highest quintile), “moderate”(intermediate three quintiles), or “low”(lowest quintile) social class. [10]

In a separate CIHI study that specifically looked at SES and all-cause, suicide and motor vehicle mortality in rural communities, the potential association between area of residence and mortality used Poisson regression to account for the effect of SES health determinants. In addition to geographic location, population- level information on education, household income, number of individuals per household, five-year mobility status, population change, proportion of population who is Aboriginal, proportion of population who have immigrated to Canada, housing data, and occupational data will represent the sociodemographic and economic variables used in the analysis. However, because this study used Census data based on specific geographic location, they were not able to control for other important health aspects, like smoking, inactivity and poor nutrition. [11]

In a study that examined the data from the 1990 Ontario Health Survey, Pomerleau et al. looked at the relationship between SES and four health behaviour in Ontario adults: smoking, fat intake, alcohol consumption and physical activity levels. Instead of constructing a composite measure, this study used four socioeconomic predictors independently in a multivariate analysis, while adjusting for age, gender and marital status. The four SES measures used in the multiple logistic and linear regression models were education level, household income, source of household income, and occupational prestige. [12]

The Chief Public Health Officer’s Report on Public Health, released in 2008, explored the most significant characteristics of SES as they relate to health inequalities in Canada. Using the statistics gathered by Statistics Canada, Health Canada, the Canadian Mortgage and Housing working group and Environment Canada, this report illustrates how income, employment and working conditions, food security, the built environment, education, social networks and access to health care, independently affect healthiness. [13]

Singh-Manoux et al. used Phase Five data of the famous Whitehall study - a cohort study in Britain that examined how varying levels of occupation are associated with health outcomes - to investigate how an individual’s perception of their health reflects and influences their health status. Sixteen variables were identified as measures that impact how individuals rate their social position. [15]

There were several methodological issues to consider before selecting an appropriate SES measure for the OHS Pilot data. As with any study of SES and health, some of these mechanistic and analytical issues include:[8]

- Lack of precision and reliability of measures
- Difficulty with the collection of individual data (e.g., high rates of non-response)
- Acquisition of longitudinal SES measurements
- Classification of women, children, retired and unemployed persons (e.g., derived differences in SES indicators, like income, may not reflect true difference in SES)
- Poor correlation of individual SES measures among some groups (e.g., some previous studies have shown that income and education were not closely associated and varied by ethnicity)
- Misleading interpretation of study results

In determining the most effective SES measure for the OHS analysis, the compositional approach and the contextual approach were considered. The compositional measures of SES refer to individual social, economic and behavioural characteristics, information that the OHS Pilot touchscreen questionnaire could provide. The contextual method involves area-based measures that can represent environmental variables such as access to goods and services, the built environment, social norms and other community level variables, information that could be acquired through Statistics Canada Community and Health profiles, and matching this data with OHS participants' postal codes.

Prior to beginning analysis on SES and lung function in the OHS Pilot data set, we decided that there were two options for including SES in a regression model:

1. SES could be quantified using the social and material deprivation indices that were derived by Statistics Canada using principal component analysis. This measure has been validated in multiple Canadian studies, appears to incorporate many of the main themes of SES and includes both compositional and contextual SES information. However, the OHS Pilot data set includes more information, such as psychological well-being and occupation, which could be reflective of true SES and is not included in this measure.
2. An SES variable could be derived using discriminate analysis like principal components or factor analysis, similar to Statistics Canada, but could include all potential variables related to SES as collected by the OHS Pilot study. This method would take advantage of all of the available information but lacks validation studies and may be too specific with information requirements. However, a derived variable could be used for future SES analysis with the OHS. If deriving a unique SES score, univariate models should be examined to determine whether a particular predictor is significantly associated with the outcome, in this case, lung function.

5 Spirometry and Lung Function

Social determinants have been shown to have a large impact on asthma and other pulmonary conditions, such as chronic obstructive pulmonary disease (COPD). The prevalence of asthma and other respiratory illnesses has risen in Canada over the past few decades, though the etiology and risk factors of the disease are not well understood. Though death due to a respiratory impairment is uncommon, understanding these conditions further will help improve the quality of life of affected individuals and reduce the burden on the Canadian healthcare system. [16]

Individuals in a lower SES group have demonstrated decreased pulmonary function, regardless of how SES is defined; lower SES can exacerbate respiratory function because of higher exposure to indoor pollutants (e.g., tobacco exposure, overcrowded housing) and outdoor pollution (e.g., environmental chemicals, radiation).

The most commonly used measures of respiratory function are spirometric indices, like those considered in the analysis of lung function for the OHS Pilot data: :

- The forced expiratory volume in the first second (FEV1); this quantity is defined as the volume of gas expired during the first second of a forced expiration following a full inspiration, capturing airway size. This is the dynamic portion of the spirometry testing.
- The forced vital capacity (FVC); this quantity is defined as the volume change of the lung between a full inhalation to total lung capacity and a maximal expiration to residual volume.
- The percent predicted values of FEV1 divided by FVC (FEV1/FVC). The ratio FEV1/FVC is clinically significant as it captures obstruction or restriction in the airway,

> representing what percentage of the total FVC was expelled from the lungs during the first second of exhalation. The National Heart, Lung, and Blood Institute/World Health Organization Global Initiative for Chronic Obstructive Lung Disease Workshop summary and the American Thoracic Society(ATS)/European Respriatory Society position paper) define an FEV1 to FVC ratio less than 75 percent as “obstructive.” The current ATS recommendations suggest that the statistically derived lower limit of normal (LLN), a value calculated with specific reference equations such that any observation below this value is considered “abnormal.” In this analysis, the percent predicted takes the observed ratios divided by the ratios of predicted values. Predicted values based on American Thoracic Society’s reference equations, as determined from Hankinson’s paper, are calculated with respect to gender, ethnicity, age and height. Percent predicted values greater than 80 percent are considered normal. [16], [17]

The above measurements, along with other common spirometric indices, were typically measured three times by trained professionals and recorded in the OHS Pilot study database. For the final analysis, the “best” measurements, equivalent to the maximum attempt for each spirometric index, will be used as the outcome values for that particular index. Because body size is associated with lung size, standing height, gender, age, and ethnicity are highly correlated with spirometric measures and should be adjusted for when modeling pulmonary function. It is also common to consider the aforementioned spirometric measures as percent predicted values. The American Thoracic Society (ATS) and the National Health and Nutrition Survey (NHANES) III have provided reference equations for prediction of FEV1 and FVC as a function of age and height, for different genders and ethnicities. [16] Using the NHANES III prediction equations, observed FEV1 and FVC will be modeled in gender-stratified regressions. Decreased spirometric indices tend to indicate impaired pulmonary function, the determinants of which are influenced by multiple genetic, social and environmental factors. Using the OHS Pilot data, FEV1 and FVC, and FEV1/FVC percent predicted will be regressed against measures of SES, after adjusting for potential confounders, such as age, gender and ethnicity.

6 Exploratory Analysis

Before constructing models that will investigate the relationship between SES and lung function, measured through observed FEV1, FVC, FEV1/FVC percent predicted, the OHS Pilot data set should be explored through descriptive statistical analysis. Being able to summarize the demographic and clinical characteristics of the sample studied is important for applying and comprehending results of inferential analysis. Social, demogaphic, cultural and economic characteristics of the OHS Pilot sample that have provided the relevant information can be summarized in the following tables. A chi-squared test using a 2 x 2 correction was used with the categorical variables to test for differences between males and females (an approximation was used for cells that had less than 5 individuals). A t-test was used to test for differences in means between males and females with the continuous variables.

Table 1: Univariate Descriptive Table of OHS Pilot Data Categorical Variables

	Total (n=8100)	F (n=4398)	M (n=3702)	
Variable	% [n]	% [n]	% [n]	P-Value
ACTIVE_WORK	n=4675	n=2488	n=2187	<0.0001
- Sedentary or office job	90.439% [4228]	95.82% [2384]	84.316% [1844]	
- Active job	9.561% [447]	4.18% [104]	15.684% [343]	
Admin.Participant.siteNo	n=8100	n=4398	n=3702	<0.0001
- MISSIS	63.889% [5175]	58.936% [2592]	69.773% [2583]	
- OWENS	16.556% [1341]	20.077% [883]	12.372% [458]	
- SUDBUR	19.556% [1584]	20.987% [923]	17.855% [661]	
ASTHMA_OCCURRENCE	n=8090	n=4391	n=3699	<0.0001
- No asthma	89.555% [7245]	88.066% [3867]	91.322% [3378]	
- Diagnosed with asthma	10.445% [845]	11.934% [524]	8.678% [321]	
CURRENT_SITUATION	n=8084	n=4389	n=3695	<0.0001
- Full time	49.171% [3975]	44.703% [1962]	54.479% [2013]	
- Part time	11.244% [909]	14.878% [653]	6.928% [256]	
- Unable to work	2.474% [200]	2.734% [120]	2.165% [80]	
- Looking after family	4.119% [333]	7.223% [317]	0.433% [16]	
- Student	0% [0]	0% [0]	0% [0]	
- Retired	29.119% [2354]	26.202% [1150]	32.585% [1204]	
- Unemployed	2.66% [215]	2.552% [112]	2.788% [103]	
- Unpaid work	1.212% [98]	1.709% [75]	0.622% [23]	
ETHNICITY_COUNT	n=8070	n=4382	n=3688	<0.0001
- Aboriginal	0.347% [28]	0.479% [21]	0.19% [7]	
- Anglo-Indian	0.05% [4]	0.046% [2]	0.054% [2]	
- Arab	0.533% [43]	0.342% [15]	0.759% [28]	
- Black	2.441% [197]	2.784% [122]	2.034% [75]	
- Black and White	0.025% [2]	0.023% [1]	0.027% [1]	
- East Asian	2.751% [222]	2.145% [94]	3.471% [128]	
- Eurasian	0.124% [10]	0.114% [5]	0.136% [5]	
- Filipino	0.818% [66]	0.822% [36]	0.813% [30]	
- Hispanic	1.264% [102]	1.073% [47]	1.491% [55]	
- Jewish	0.397% [32]	0.434% [19]	0.352% [13]	
- Multiples	2.069% [167]	2.282% [100]	1.817% [67]	
- Other	1.066% [86]	1.095% [48]	1.03% [38]	
- South Asian	2.788% [225]	1.848% [81]	3.905% [144]	
- Southeast Asian	2.652% [214]	1.552% [68]	3.959% [146]	
- West Asian	0.248% [20]	0.16% [7]	0.352% [13]	
- White	82.429% [6652]	84.801% [3716]	79.61% [2936]	

Table 2: Univariate Descriptive Table of OHS Pilot Data Categorical Variables

	Total (n=8100)	F (n=4398)	M (n=3702)	
Variable	% [n]	% [n]	% [n]	P-Value
FOOD_INSECURITY	n=8050	n=4376	n=3674	0.021
- Haven't experienced	97.988% [7888]	97.646% [4273]	98.394% [3615]	
- Experienced	2.012% [162]	2.354% [103]	1.606% [59]	
FREQUENCY_RELIGIOUS_PRACTICE	n=7874	n=4276	n=3598	<0.0001
- None	27.267% [2147]	23.293% [996]	31.99% [1151]	
- Once a year	5.486% [432]	4.888% [209]	6.198% [223]	
- 3-4 times a year	11.506% [906]	10.664% [456]	12.507% [450]	
- Once a month	5.69% [448]	5.683% [243]	5.698% [205]	
- Once a week	16.497% [1299]	16.23% [694]	16.815% [605]	
- Daily or almost daily	33.553% [2642]	39.242% [1678]	26.793% [964]	
HIGHEST_LEVEL_COMPLETED	n=8066	n=4378	n=3688	<0.0001
- None	0% [0]	0% [0]	0% [0]	
- Elementary	2.306% [186]	1.804% [79]	2.901% [107]	
- High School	18.46% [1489]	20.466% [896]	16.079% [593]	
- Technical Certificate	8.331% [672]	5.025% [220]	12.256% [452]	
- College	22.824% [1841]	28.095% [1230]	16.567% [611]	
- University Certificate	4.525% [365]	4.5% [197]	4.555% [168]	
- Bachelor's Degree	28.651% [2311]	28.392% [1243]	28.959% [1068]	
- Graduate Degree	14.902% [1202]	11.718% [513]	18.682% [689]	
HOUSE_INCOME_LAST_YEAR	n=7673	n=4099	n=3574	<0.0001
- High	19.536% [1499]	18.492% [758]	20.733% [741]	
- Low	18.806% [1443]	21.42% [878]	15.809% [565]	
- Medium	61.658% [4731]	60.088% [2463]	63.458% [2268]	
LAST_ROUTINE_MEDICAL_EXAM	n=8076	n=4387	n=3689	<0.0001
- Never	34.51% [2787]	33.804% [1483]	35.348% [1304]	
- <6 months ago	0.161% [13]	0.068% [3]	0.271% [10]	
- 6 months - 1 year	32.293% [2608]	32.733% [1436]	31.77% [1172]	
- 1 year - 2 years	22.363% [1806]	23.684% [1039]	20.792% [767]	
- 2 years - 3 years	5.349% [432]	5.334% [234]	5.367% [198]	
- More than 3 years	5.324% [430]	4.377% [192]	6.452% [238]	
LAST_VISIT_DENTIST	n=8085	n=4394	n=3691	0.003
- Never	69.4% [5611]	69.345% [3047]	69.466% [2564]	
- <6 months ago	0.272% [22]	0.182% [8]	0.379% [14]	
- 6 months - 1 year	19.604% [1585]	20.687% [909]	18.315% [676]	
- 1 year - 2 years	5.504% [445]	5.166% [227]	5.906% [218]	
- 2 years - 3 years	1.979% [160]	1.684% [74]	2.33% [86]	
- More than 3 years	3.241% [262]	2.936% [129]	3.603% [133]	
MARITAL_STATUS	n=8085	n=4386	n=3699	<0.0001
- Married	81.126% [6559]	73.78% [3236]	89.835% [3323]	
- Divorced, widowed or separated	14.335% [1159]	20.315% [891]	7.245% [268]	
- Single, never married	4.539% [367]	5.905% [259]	2.92% [108]	
SMOKE_STATUS	n=8066	n=4381	n=3685	<0.0001
- Never smoked	55.728% [4495]	59.895% [2624]	50.773% [1871]	
- Ex-smoker	38.173% [3079]	34.49% [1511]	42.551% [1568]	
- Current smoker	6.1% [492]	5.615% [246]	6.676% [246]	

Table 3: Univariate Descriptive Table of OHS Pilot Data Continuous Variables

	Total (n=8100)	F (n=4398)	M (n=3702)	
Variable	Mean (CI)[n]	Mean (CI)[n]	Mean (CI)[n]	P-Value
AGE	55.7 (55.5,55.8)[7950]	54.5 (54.3,54.8)[4314]	57.0 (56.7,57.3)[3636]	<0.0001
HEIGHT	168.7 (168.4,168.8)[7158]	162.8 (162.6,163.0)[3876]	175.5 (175.3,175.7)[3282]	<0.0001
BMI	27.3 (27.2,27.4)[7702]	26.7 (26.6,26.9)[4219]	28.0 (27.9,28.1)[3483]	<0.0001
WEIGHT	78.0 (77.7,78.4)[7732]	71.5 (71.1,72.0)[4280]	86.1 (85.7,86.5)[3452]	<0.0001
	% [n]	% [n]	% [n]	

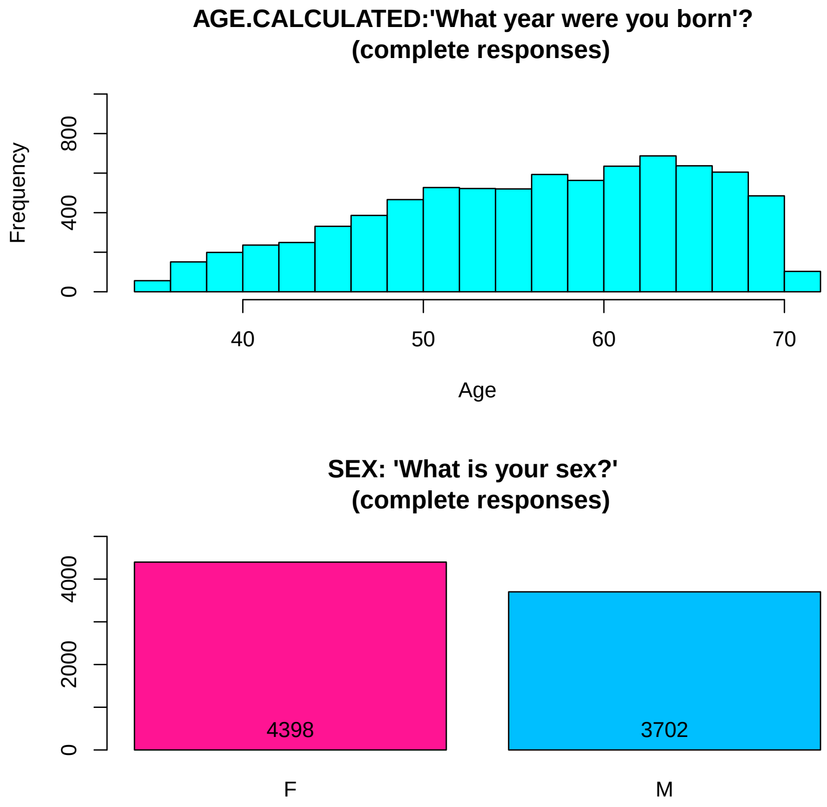


Figure 1: Age and Sex of the OHS Pilot

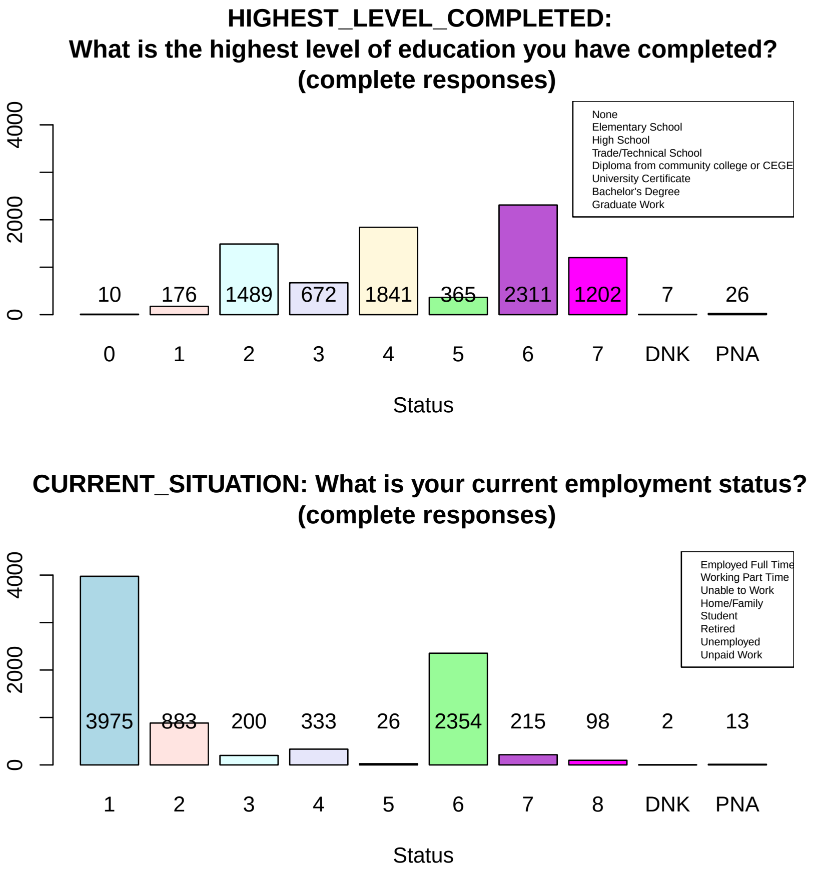


Figure 2: Education and Employment Status

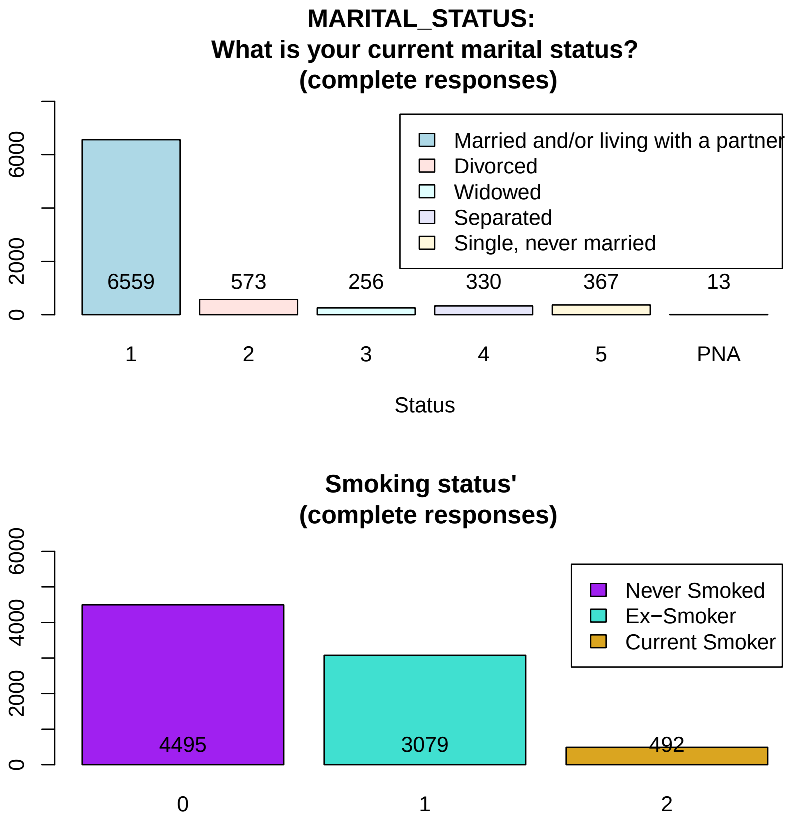
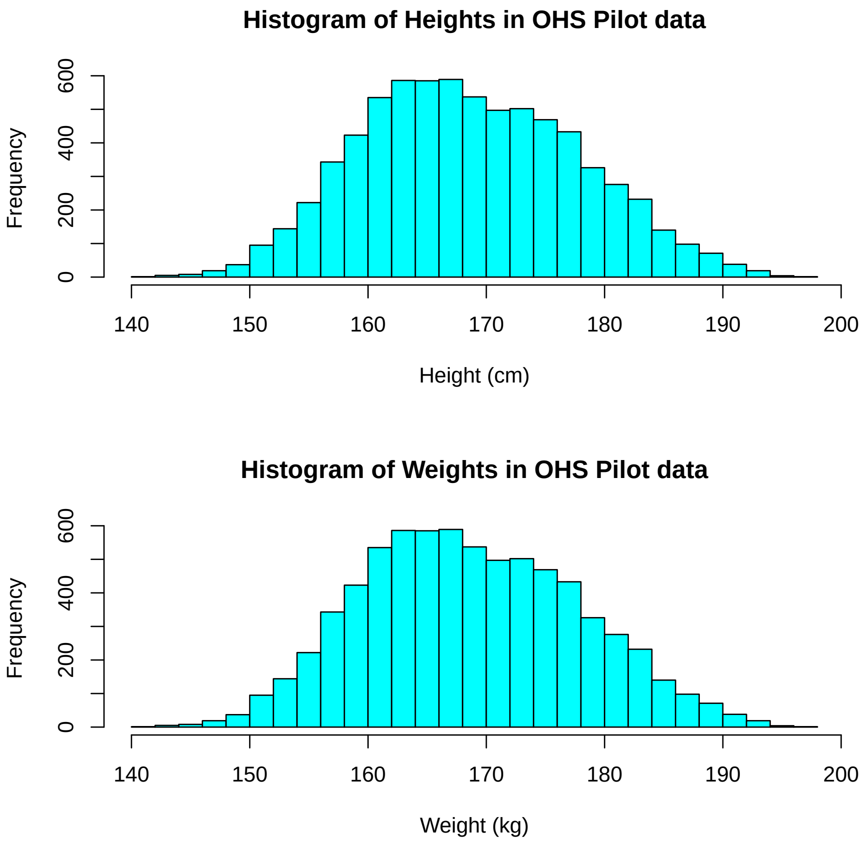


Figure 3: Marital Status and Smoking Status of the OHS Pilot



Histogram of Heights in OHS Pilot data

Histogram of Weights in OHS Pilot data

Figure 4: Height and Weight distributions of the OHS Pilot

These plots verify that the OHS Pilot sample are between the ages of 35 and 70, and there appear to be higher frequencies of older adults. There is a slightly higher proportion of females than males in the sample. The majority of participants have at least some post-secondary education (approximately 80 %), with almost half of the cohort having at least a Bachelor's degree (almost 45%); a smaller proportion of the Ontario population are University educated (25%). The high proportion of retirees in the sample make it hard to compare the working populations, though the OHS sample has a much lower unemployment rate(2.66% versus 6.4%) and the OHS has a larger proportion of married individuals (81% versus 74%). It is important to note that the Census has gathered information from individuals 15 and older, which means that age is probably playing a big role in the discrepancies between these proportions. The largest proportion of people in the OHS Pilot study were sampled from the Mississauga (e.g., the "urban" region), though this proportion is actually less than the proportion of Ontarioans who reside in a metropolitan setting; according to the 2006 Census from Statistics Canada, approximately 85 percent of Ontarioans live in an urban area. This reflects the choice by the OHS to "over-sample" the rural and northern communities to obtain a better grasp on their health status, to be able to describe that population more clearly. The height and weight histograms reflect what should be seen in a sample of both males and females - the apparent "double peak" in the anthropometric histograms represents the peak of each distinct distribution for males and females. Similar to the results of the 2006 Census, the OHS Pilot data (82%) reflects the predominantly white Ontario population (77%). [18]

Studying the relationship between SES and lung function requires that ethnicity be taken into account. To accommodate confounding by ethnicity, epidemiological models typically use multivariate analysis or stratification. As demonstrated above, OHS Pilot sample is predominantly white and because the remaining individuals are divided among seventeen other ethnici groups, to stratify by ethnicity would result in a large loss of power as would adjusting for it in a model. Because lung function has not been well researched in all ethnic origins, spirometric reference equations for determining predicted values have only been rigourously defined for white, black and Hispanic ethnic groups.. To capitalize on the current data available for lung function, the analysis in this paper considers only the proportion of cases who have identified themselves as "white". Further analysis will be required to determine the true impact of different cultural backgrounds on lung function, as there is a gap in current research, though the OHS Pilot data does not provide a heterogenous sample of ethnicities.

The characteristics described above are crucial for understanding the relationship between lung function and SES. Age, height, weight and ethnic background influence work through different biological mechanisms to influence an individual's respiratory potential, though the relationship between SES and lung function is less clear.

To understand the nature of the outcome measures used in spirometry, in particular from the OHS Pilot data, the following graphs illustrate the ranges of FEV1, FVC and FEV1/FVC percent predicted, average values and how they are distributed within the OHS Pilot cohort. Note that these plots are stratified by gender, and only include the spirometric measures of individuals who classified themselves as "white". Upon initial glance, it appears that FEV1 and FVC are normally distributed though the percent of predicted FEV1/FVC is not as clearly defined.

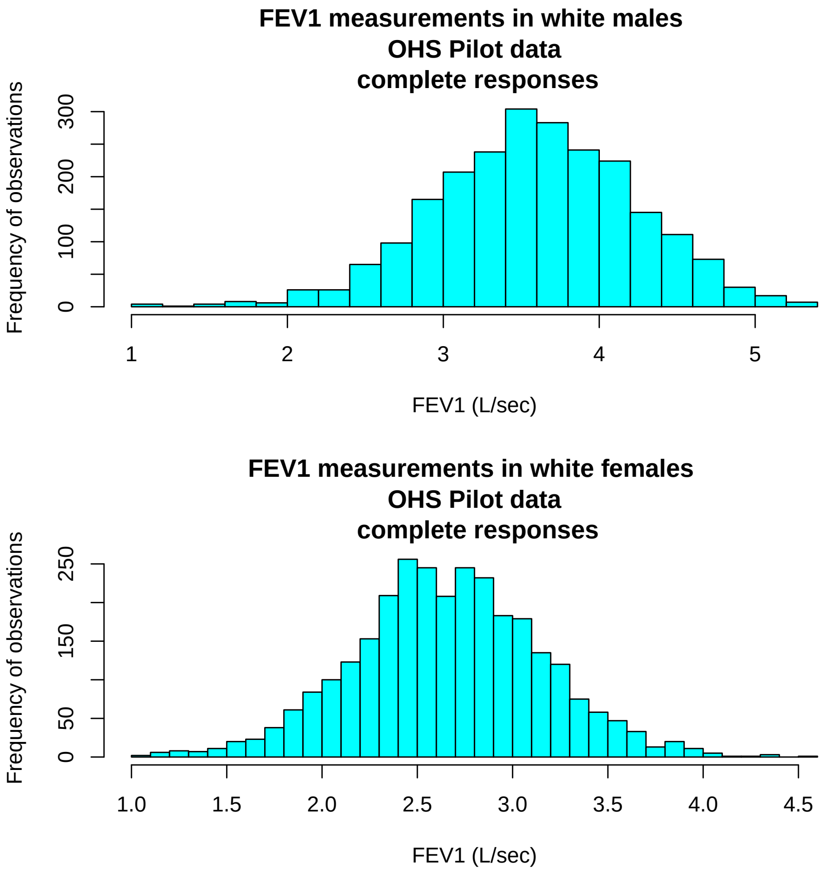


Figure 5: FEV1 in males and females

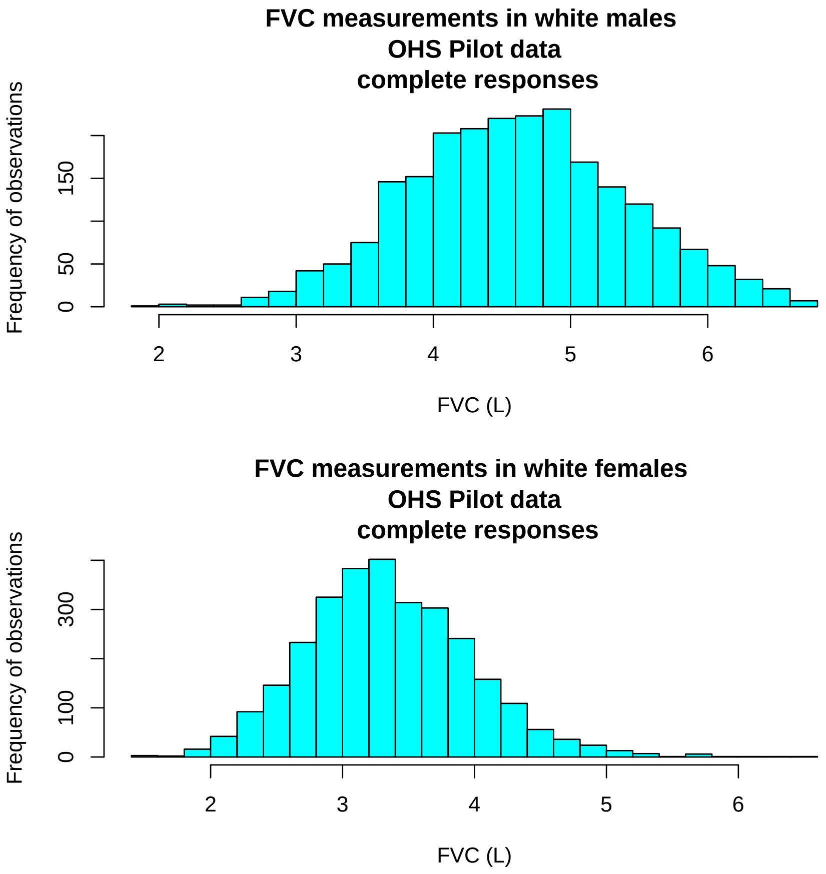


Figure 6: FVC in males and females

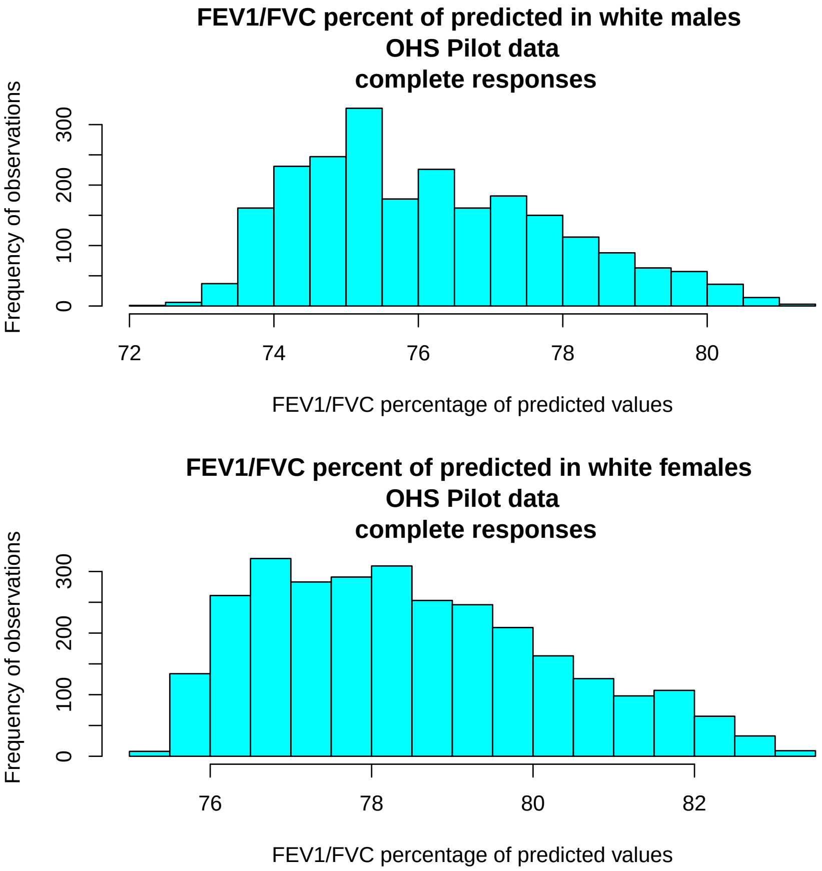


Figure 7: FEV1/FVC % predicted males and females

6.1 Bivariate Analysis

To determine which SES variables should be included in the multivariate models for lung function, univariate analysis will examine the independent influence of the potential SES predictosr on each of the measured outcomes. The factors which have a significant impact on the outcomes individually will then be included in a multivariate linear model that adjusts for age and height. After these multivariate models have been created, the SES structures which affect lung function in white Ontarioans can be described with greater detail.

Two variables needed to be created using questions from the Touchscreen questionnaire to attain the appropriate information. The Touchscreen questionnaire asked participants if they *ever* smoked (0=No; 1=Yes) and if they *currently* smoke(0=Not in the past 30 days; 1= Occasionally - more than 1 cigarrette in the past 30 days but not every day; 2=Every day). For the spirometric analysis, it was decided to recode these two questions into one categorical variable that would allocate participants to “0=Never smoked”, “1=Ex-Smoker”, “2=Current Smoker”. Similarly, the self-reported occupation at time of assessment - a 27-category question - was recoded into an indicator variable that reflected the activity level of the participant’s self-reported occupation: “0=Sedentary”, “1=Active”. The occupational domains which include Legislators, Senior Officials, Managers, Professionals, Technicians, Clerks and Service and Shop workers were classified as sedentary; Agricultural and Fishery workers, Trades workers,Plant and Machine Operators, Elementary Occupation workers and those in the Armed Forces were classified as active.

```
> spiro_clean$SMOKE_STATUS <- rep(NA, nrow(spiro_clean))
> spiro_clean$SMOKE_STATUS[which(spiro_clean$CURRENTLY_SMOKE ==
+     1 | spiro_clean$CURRENTLY_SMOKE == 2)] <- 2
> spiro_clean$SMOKE_STATUS[which(spiro_clean$EVER_SMOKE == 1) &
+     spiro_clean$CURRENTLY_SMOKE == 0] <- 1
> spiro_clean$SMOKE_STATUS[which(spiro_clean$EVER_SMOKE == 0)] <- 0
> spiro_clean$ACTIVE_WORK <- rep(NA, nrow(spiro_clean))
> spiro_clean$ACTIVE_WORK[which(spiro_clean$CURRENT_WORK_ISIC1 ==
+     1 | spiro_clean$CURRENT_WORK_ISIC1 == 2 | spiro_clean$CURRENT_WORK_ISIC1 ==
+     3 | spiro_clean$CURRENT_WORK_ISIC1 == 4 | spiro_clean$CURRENT_WORK_ISIC1 ==
+     5)] <- 0
> spiro_clean$ACTIVE_WORK[which(spiro_clean$CURRENT_WORK_ISIC1 ==
+     6 | spiro_clean$CURRENT_WORK_ISIC1 == 7 | spiro_clean$CURRENT_WORK_ISIC1 ==
+     8 | spiro_clean$CURRENT_WORK_ISIC1 == 9 | spiro_clean$CURRENT_WORK_ISIC1 ==
+     10)] <- 1
```

The univariate analysis has been conducted for eighteen variables believed to affect lung function:

- **AGE.CALCULATED**: continuous variable represents age of individual; in the range of 34 to 71.
- **HIGHEST_LEVEL_COMPLETED**: 7-category variable represents the highest level of education completed; 0=None, 1=Elementary School, 2=High School, 3=Trade School, 4=Diploma from community college, or non-university certificate, 5=University certificate below Bachelor’s degree, 6=Bachelor’s degree, 7=Graduate degree.
- **CURRENT_SITUATION**: 8-category variable represents current employment status: 1=Working full-time, 2=working part-time, 3=Unable to work because of disability or sickness, 4=Looking after family, 5=Student, 6=Retired, 7=Unemployed, 8=Unpaid work. In the multivariate analysis, 8 individuals who had classified themselves as “Students” were recoded to “Working part-time” because the small sample size would make it hard to detect differences between this group.

- **MARITAL_STATUS**: 5-category question represents current marital status: 1=Married or living with partner, 2=Divorced, 3=Widowed, 4=Separated, 5=Single, never married. For analytic purposes, this category was recoded into 3 levels to condense the information: 1=Married or living with partner, 2=Divorced, widowed or separated, 3=Single, never married. This categorization corresponds to the Statistics Canada classification of marital status.

- **HOUSE_INCOME_LAST_YEAR**: 8-category question reflects the approximate total household income before taxes: 1=Less than $10,000, 2=$10,000 to $24,999, 3= $25,000 to $49,999, 4=$50,000 to $74,999, 5=$75,000 to $99,999, 6=$100,000 to $149,999, 7=$150,000 to $199,999, 8=$200,000 or more. This was recoded for the analysis as 1='Low' (corresponding to the first 3 factors), 2='Medium' (corresponding to the intermediate 3 factors) and 3='High' (corresponding to the top 2 factors). This recoding was done to condense the information in a logical way after exploring the trends of lung function across the income categories.

- **NUMBER_SUPPORTED_BY_INCOME**: value represents the total number of people that were supported by this income, maximum value was 16, though most observations were between 1 and 3.

- **FREQUENCY_RELIGIOUS_PRACTICE**: 6-category question reflects how often a participant practiced spirituality: 0=Not at all, 1= Once a year, 2=3 to 4 times a month, 3= Once a month, 4= Once a week, 5= Daily or almost daily.

- **LAST_ROUTINE_MEDICAL_EXAM**: 6-category question acts as an indicator of access to healthcare by representing the last medical check-up a subject had: 0=Never had a check-up, 1= Less than 6 months ago, 2= 6 months to 1 year ago, 3= 1 year to less than 2 years ago, 4=2 years ago to less than 3 years ago, 5= More than 3 years ago . Because of small response numbers in the "0=Never" group, the reference category was releveled to "1".

- **LAST_VISIT_DENTIST**: 6-category question acts as an indicator of access to healthcare, by reflecting the last dental exam a subject had: 0=Never had a check-up, 1= Less than 6 months ago, 2= 6 months to 1 year ago, 3= 1 year to less than 2 years ago, 4=2 years ago to less than 3 years ago, 5= More than 3 years ago. Because of small response numbers in the "0=Never" group, the reference category was releveled to "1".

- **FOOD_INSECURITY**: Indicator variable representing whether the participant or anyone in the household had too little food to eat due to lack of money in the past 12 months: 0= No, 1=Yes.

- **INPUT_PART_HEIGHT_SP**: Physical measure that represents the height in centimetres measured at the assessment centre. Extreme values were removed (outside 3 standard deviations from the mean).

- **RES_WEIGHT_BIO**: Physical measure which represents the weight in kilograms as measured at the assessment centre. Extreme values were removed (outside 3 standard deviations from the mean).

- **RES_BODY_MASS_INDEX**: Physical measure which represent the body mass index as measured at the assessment centre, in kilograms per metre-squared. Extreme values were removed (outside 3 standard deviations from the mean).

- **SMOKE_STATUS**: 3-category variable that represents participants' smoking status: 0= Never smoked, 1= Ex-smoker, 2= Current smoker.

- **ASTHMA_OCCURENCE**: Indicator variable from the Nurse's interview at the assessment centre that represents whether a participant has doctor-diagnosed asthma. 0= Never been diagnosed with asthma, 1= Diagnosed with asthma.

- **Admin.Participant.siteNo**: 3-category variable represents the location of the assessment centre that the participants attended. The three assessment centres were Mississauga, Owen Sound and Sudbury, which represent urban, rural and northern communities, respectively. This variable can be interpreted as an area-level indicator for SES, as it will capture regional sociodemographic and economic variations.
- **ACTIVE_WORK**: Indicator variable which represents the activity required in the participants' self-reported occupations. 0= sedentary, 1= active.
- **IPAQ_SCORE**: 3-category score variable from the International Physical Activity Questionnaire, long-form, as it was answered in the Touchscreen questionnaire, classified by metabolic equivalent minutes per week. 1= Low Activity levels, 2= Moderate Activity levels, 3= Vigorous Activity levels. Due to the current frequency of missing values, as discussed earlier in this report, the influence of physical activity using the IPAQ categories will be measured with a sensitivity analysis at a later date if it is found to be significant. Future sensitivity analyses of the lung function and SES models will compare the subset of individuals who have complete IPAQ scores with the model including the entire cohort studied. The IPAQ responses will not be used in this particular analysis for a couple reasons: the large proportion of missing data for the physical activity measures induces a need for an accompanying sensitivity analysis and further missing data investigation; and, the influence of physical activity on lung function and socioeconomic position requires an in-depth analysis to accommodate the cloudy understanding of its directionality (e.g., someone who with low exercise levels will likely report a poorer lung function than expected, however a person with poor lung function may be less likely to exercise for longer).

In the bivariate analysis, the outcome spirometric indices FEV1, FVC, and FEV1/FVC percent predicted are regressed onto these variables. Because the outcomes are continous variables, the *lm* function in R is used to generate appropriate linear models and summaries, which will be used to evaluate the significance of each predictor independently. The following excerpt from the R script used to run these models for the forced expiratory volume in 1 second outcome in white males:

```
> mwFEV1_age <- lm(malew$FEV1_final ~ malew$AGE.CALCULATED, malew)
> summary(mwFEV1_age)


Call:
lm(formula = malew$FEV1_final ~ malew$AGE.CALCULATED, data = malew)

Residuals:
     Min       1Q   Median       3Q      Max
-2.50236 -0.36502  0.00299  0.35598  1.64099

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           5.571078   0.080179   69.48   <2e-16 ***
malew$AGE.CALCULATED -0.034668   0.001382  -25.09   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5663 on 2281 degrees of freedom
Multiple R-squared: 0.2162,	Adjusted R-squared: 0.2159
F-statistic: 629.3 on 1 and 2281 DF,  p-value: < 2.2e-16
```

The summary produced for this univariate model shows that the residuals for this covariate are concentrated around 0 (as would be expected for a normal model), and that the age predictor alone explains a significant portion of the variation (21.59%) in the outcome for FEV1 (p <

0.05). The negative coefficient value -0.035110 confirms the widely accepted concept that as age increases, airway size may begin to decrease. The remaining predictors were analyzed in the same manner:

```
> mwFEV1_edu <- lm(malew$FEV1_final ~ malew$HIGHEST_LEVEL_COMPLETED,
+     malew)
> mwFEV1_work <- lm(malew$FEV1_final ~ malew$CURRENT_SITUATION,
+     malew)
> mwFEV1_mari <- lm(malew$FEV1_final ~ malew$MARITAL_STATUS, malew)
> mwFEV1_inc <- lm(malew$FEV1_final ~ malew$HOUSE_INCOME_LAST_YEAR,
+     malew)
> mwFEV1_num_inc <- lm(malew$FEV1_final ~ malew$NUMBER_SUPPORTED_BY_INCOME,
+     malew)
> mwFEV1_rel <- lm(malew$FEV1_final ~ malew$FREQUENCY_RELIGIOUS_PRACTICE,
+     malew)
> mwFEV1_med <- lm(malew$FEV1_final ~ malew$LAST_ROUTINE_MEDICAL_EXAM,
+     malew)
> mwFEV1_dent <- lm(malew$FEV1_final ~ malew$LAST_VISIT_DENTIST,
+     malew)
> mwFEV1_food <- lm(malew$FEV1_final ~ malew$FOOD_INSECURITY, malew)
> mwFEV1_height <- lm(malew$FEV1_final ~ malew$INPUT_PART_HEIGHT_SP,
+     malew)
> mwFEV1_weight <- lm(malew$FEV1_final ~ malew$RES_WEIGHT_BIO,
+     malew)
> mwFEV1_BMI <- lm(malew$FEV1_final ~ malew$RES_BODY_MASS_INDEX,
+     malew)
> mwFEV1_smoke <- lm(malew$FEV1_final ~ malew$SMOKE_STATUS, malew)
> mwFEV1_asthma <- lm(malew$FEV1_final ~ malew$ASTHMA_OCCURRENCE,
+     malew)
> mwFEV1_IPAQ <- lm(malew$FEV1_final ~ malew$IPAQ_SCORE, malew)
> mwFEV1_ac <- lm(malew$FEV1_final ~ malew$Admin.Participant.siteNo,
+     malew)
> mwFEV1_act_work <- lm(malew$FEV1_final ~ malew$ACTIVE_WORK, malew)
```

The results of the bivariate analysis for FEV1, in addition to the remaining two outcomes (FVC and FEV1/FVC percent predicted) for white males are summarized in the tables below. The Analysis of Variance (ANOVA) for each univariate model are provided in the tables below, where p <0.05 (“Pr(> F)”) is considered signicificant, and $0.05 \leq p \leq 0.10$ is considered borderline significant and will be tested in the multivariate model.

Table 4: ANOVA of Bivariate Linear Analyses for FEV1 for males

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
malew$AGE.CALCULATED	1	201.81	201.81	629.28	0.0000
malew$HIGHEST_LEVEL_COMPLETED	6	24.36	4.06	10.17	0.0000
malew$CURRENT_SITUATION	6	93.99	15.67	42.48	0.0000
malew$MARITAL_STATUS	2	0.29	0.15	0.36	0.6983
malew$HOUSE_INCOME_LAST_YEAR	2	37.99	18.99	48.37	0.0000
malew$NUMBER_SUPPORTED_BY_INCOME	1	57.66	57.66	150.20	0.0000
malew$FREQUENCY_RELIGIOUS_PRACTICE	5	2.05	0.41	1.01	0.4095
malew$LAST_ROUTINE_MEDICAL_EXAM	5	21.71	4.34	10.85	0.0000
malew$LAST_VISIT_DENTIST	5	6.41	1.28	3.15	0.0077
malew$FOOD_INSECURITY	1	0.66	0.66	1.62	0.2038
malew$INPUT_PART_HEIGHT_SP	1	159.37	159.37	469.68	0.0000
malew$RES_WEIGHT_BIO	1	3.84	3.84	9.43	0.0022
malew$RES_BODY_MASS_INDEX	1	22.30	22.30	55.83	0.0000
malew$SMOKE_STATUS	1	22.90	22.90	57.37	0.0000
malew$ASTHMA_OCCURRENCE	1	10.28	10.28	25.41	0.0000
malew$IPAQ_SCORE	1	0.13	0.13	0.34	0.5625
malew$Admin.Participant.siteNo	2	0.03	0.02	0.04	0.9605
malew$ACTIVE_WORK	1	0.01	0.01	0.02	0.8926

Table 5: ANOVA of Bivariate Linear Analyses for FVC for males

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
malew$AGE.CALCULATED	1	237.37	237.37	454.22	0.0000
malew$HIGHEST_LEVEL_COMPLETED	6	39.36	6.56	10.74	0.0000
malew$CURRENT_SITUATION	6	116.76	19.46	33.74	0.0000
malew$MARITAL_STATUS	2	0.96	0.48	0.77	0.4635
malew$HOUSE_INCOME_LAST_YEAR	2	51.55	25.78	42.65	0.0000
malew$NUMBER_SUPPORTED_BY_INCOME	1	72.27	72.27	121.47	0.0000
malew$FREQUENCY_RELIGIOUS_PRACTICE	5	2.88	0.58	0.93	0.4626
malew$LAST_ROUTINE_MEDICAL_EXAM	5	29.20	5.84	9.50	0.0000
malew$LAST_VISIT_DENTIST	5	9.98	2.00	3.20	0.0070
malew$FOOD_INSECURITY	1	0.35	0.35	0.56	0.4533
malew$INPUT_PART_HEIGHT_SP	1	367.08	367.08	788.17	0.0000
malew$RES_WEIGHT_BIO	1	3.12	3.12	4.99	0.0256
malew$RES_BODY_MASS_INDEX	1	71.16	71.16	119.51	0.0000
malew$SMOKE_STATUS	1	24.97	24.97	40.55	0.0000
malew$ASTHMA_OCCURRENCE	1	4.05	4.05	6.48	0.0110
malew$IPAQ_SCORE	1	0.91	0.91	1.50	0.2216
malew$Admin.Participant.siteNo	2	1.35	0.68	1.08	0.3397
malew$ACTIVE_WORK	1	0.02	0.02	0.03	0.8682

Table 6: ANOVA of Bivariate Linear Analyses for FEV1/FVC percent predicted for males

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
malew$AGE.CALCULATED	1	0.18	0.18	87.90	0.0000
malew$HIGHEST_LEVEL_COMPLETED	6	0.02	0.00	1.29	0.2591
malew$CURRENT_SITUATION	6	0.09	0.01	7.24	0.0000
malew$MARITAL_STATUS	2	0.01	0.00	1.26	0.2851
malew$HOUSE_INCOME_LAST_YEAR	2	0.01	0.01	3.34	0.0355
malew$NUMBER_SUPPORTED_BY_INCOME	1	0.04	0.04	19.23	0.0000
malew$FREQUENCY_RELIGIOUS_PRACTICE	5	0.00	0.00	0.29	0.9189
malew$LAST_ROUTINE_MEDICAL_EXAM	5	0.01	0.00	1.31	0.2566
malew$LAST_VISIT_DENTIST	5	0.01	0.00	0.61	0.6950
malew$FOOD_INSECURITY	1	0.00	0.00	0.47	0.4922
malew$INPUT_PART_HEIGHT_SP	1	0.07	0.07	35.03	0.0000
malew$RES_WEIGHT_BIO	1	0.01	0.01	2.90	0.0888
malew$RES_BODY_MASS_INDEX	1	0.05	0.05	25.28	0.0000
malew$SMOKE_STATUS	1	0.04	0.04	18.18	0.0000
malew$ASTHMA_OCCURRENCE	1	0.01	0.01	4.35	0.0371
malew$IPAQ_SCORE	1	0.01	0.01	3.58	0.0586
malew$Admin.Participant.siteNo	2	0.01	0.00	2.01	0.1341
malew$ACTIVE_WORK	1	0.00	0.00	0.09	0.7636

Table 7: ANOVA of Bivariate Linear Analyses for FEV1 for females

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
femalew$AGE.CALCULATED	1	196.77	196.77	1124.49	0.0000
femalew$HIGHEST_LEVEL_COMPLETED	6	25.67	4.28	18.27	0.0000
femalew$CURRENT_SITUATION	6	80.24	13.37	62.10	0.0000
femalew$MARITAL_STATUS	2	10.51	5.25	21.98	0.0000
femalew$HOUSE_INCOME_LAST_YEAR	2	43.86	21.93	96.39	0.0000
femalew$NUMBER_SUPPORTED_BY_INCOME	1	72.33	72.33	332.26	0.0000
femalew$FREQUENCY_RELIGIOUS_PRACTICE	5	8.60	1.72	7.19	0.0000
femalew$LAST_ROUTINE_MEDICAL_EXAM	5	2.11	0.42	1.74	0.1213
femalew$LAST_VISIT_DENTIST	5	2.75	0.55	2.28	0.0446
femalew$FOOD_INSECURITY	1	1.51	1.51	6.23	0.0126
femalew$INPUT_PART_HEIGHT_SP	1	129.35	129.35	652.92	0.0000
femalew$RES_WEIGHT_BIO	1	1.12	1.12	4.64	0.0313
femalew$RES_BODY_MASS_INDEX	1	14.20	14.20	59.76	0.0000
femalew$SMOKE_STATUS	1	0.06	0.06	0.25	0.6164
femalew$ASTHMA_OCCURRENCE	1	2.38	2.38	9.83	0.0017
femalew$IPAQ_SCORE	1	0.57	0.57	2.37	0.1236
femalew$Admin.Participant.siteNo	2	4.79	2.39	9.93	0.0001
femalew$ACTIVE_WORK	1	0.01	0.01	0.04	0.8427

Table 8: ANOVA of Bivariate Linear Analyses for FVC for females

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
femalew$AGE.CALCULATED	1	278.01	278.01	937.30	0.0000
femalew$HIGHEST_LEVEL_COMPLETED	6	46.98	7.83	20.80	0.0000
femalew$CURRENT_SITUATION	6	114.96	19.16	54.25	0.0000
femalew$MARITAL_STATUS	2	14.46	7.23	18.68	0.0000
femalew$HOUSE_INCOME_LAST_YEAR	2	70.77	35.39	96.20	0.0000
femalew$NUMBER_SUPPORTED_BY_INCOME	1	108.65	108.65	306.30	0.0000
femalew$FREQUENCY_RELIGIOUS_PRACTICE	5	12.11	2.42	6.27	0.0000
femalew$LAST_ROUTINE_MEDICAL_EXAM	5	2.35	0.47	1.20	0.3054
femalew$LAST_VISIT_DENTIST	5	4.72	0.94	2.41	0.0341
femalew$FOOD_INSECURITY	1	2.78	2.78	7.12	0.0077
femalew$INPUT_PART_HEIGHT_SP	1	284.73	284.73	967.51	0.0000
femalew$RES_WEIGHT_BIO	1	0.02	0.02	0.05	0.8307
femalew$RES_BODY_MASS_INDEX	1	53.94	53.94	144.42	0.0000
femalew$SMOKE_STATUS	1	0.10	0.10	0.26	0.6125
femalew$ASTHMA_OCCURRENCE	1	0.37	0.37	0.95	0.3303
femalew$IPAQ_SCORE	1	1.55	1.55	4.00	0.0455
femalew$Admin.Participant.siteNo	2	13.87	6.93	17.90	0.0000
femalew$ACTIVE_WORK	1	0.02	0.02	0.06	0.8033

Table 9: ANOVA of Bivariate Linear Analyses for FEV1/FVC percent predicted for females

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
femalew$AGE.CALCULATED	1	9053.60	9053.60	221792.01	0.0000
femalew$HIGHEST_LEVEL_COMPLETED	6	375.25	62.54	20.68	0.0000
femalew$CURRENT_SITUATION	6	3413.42	568.90	287.36	0.0000
femalew$MARITAL_STATUS	2	293.47	146.74	48.14	0.0000
femalew$HOUSE_INCOME_LAST_YEAR	2	736.75	368.38	127.21	0.0000
femalew$NUMBER_SUPPORTED_BY_INCOME	1	2192.37	2192.37	915.25	0.0000
femalew$FREQUENCY_RELIGIOUS_PRACTICE	5	206.54	41.31	13.34	0.0000
femalew$LAST_ROUTINE_MEDICAL_EXAM	5	50.04	10.01	3.19	0.0071
femalew$LAST_VISIT_DENTIST	5	40.92	8.18	2.61	0.0232
femalew$FOOD_INSECURITY	1	0.94	0.94	0.30	0.5852
femalew$INPUT_PART_HEIGHT_SP	1	234.70	234.70	76.52	0.0000
femalew$RES_WEIGHT_BIO	1	8.16	8.16	2.59	0.1074
femalew$RES_BODY_MASS_INDEX	1	93.50	93.50	30.01	0.0000
femalew$SMOKE_STATUS	1	6.25	6.25	1.99	0.1587
femalew$ASTHMA_OCCURRENCE	1	78.51	78.51	25.16	0.0000
femalew$IPAQ_SCORE	1	6.23	6.23	1.98	0.1595
femalew$Admin.Participant.siteNo	2	39.52	19.76	6.30	0.0019
femalew$ACTIVE_WORK	1	2.52	2.52	1.01	0.3146

White male FEV1 bivariate analysis: As expected,the physiological characteristics age, height, weight (similarly, body mass index), smoking status and asthma presence are significantly associated with the lung function measured by FEV1. Of the SES variables, the conventional measures of SES - income, education, and working status - have all been detected as significant predictors, in addition to the number of individuals supported by income, marital status, and the last visit to the dentist and doctor. These bivariate relationships will be explored and described further in the following section.

White male FVC bivariate analysis: Mirroring the bivariate findings of the FEV1 analysis, FVC in white males is similarly influenced by age, height, weight, smoking status, asthma diagnoses, household income, education level, working status, number of individuals supported by the household income, marital status and the access to health care aggregate questions. Though there are minor differences in the ANOVA output (e.g., marital status is borderline significant, as is the last visit to the dentist, and the IPAQ physical activity measure could be considered a covariate), generally the selected variables demonstrate similar relationships, so the same variables will be used in the multivariate analysis.

White male FEV1:FVC percent predicted bivariate analysis: By definition, predicted values for spirometric indices incorporate gender, age and height in their equations, therefore it is expected that there are strong bivariate associations between these measurements and FEV1/FVC predicted value. Every physical measure (age, height, weight, body mass index) and most of the social demographic variables were significantly associated with the outcome, with the exception of the IPAQ score, assessment centre and the active work indicator.

White female FEV1 bivariate analysis: Considering forced expiratory volume in 1 second, white females appear to have the same physical and social influences as males, though frequency of spiritual practice, food insecurity, physical activity (e.g., IPAQ score), and demographic location also demonstrate a significant contribution to this spirometric index. For the same reasons mentioned previously, the influence of physical activity will be examined in a future analysis, though the other variables will be included in a multivariate model. It is interesting to note that weight plays a smaller role in explaining variability in lung function than body mass index for females than males.

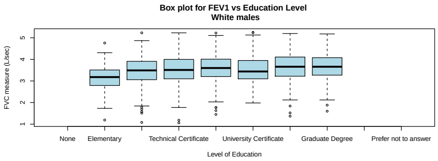
White female FVC bivariate analysis: As expected, most of the biological measurements and indicators are independently related to FVC in white females. There is less emphasis on the access to health care SES variables (e.g., medical and dental visit frequencies), though similar to the female FEV1 analysis, religious practices and assessment centre seem to significantly affect this spirometric index.

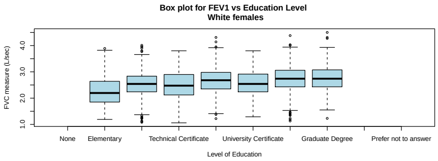
White female FEV1/FVC percent predicted: Fewer bivariate relationships are noted for FEV1/FVC percent predicted in the female cohort; food insecurity, physical activity measures and work activity levels were not associated with the outcome, though the remaining socioeconomic factors will be included in further analysis. Though somewhat surprising that weight and smoking status don't demonstrate a significant association with the FEV1/FVC predicted value in females, a multivariate analysis will adjust for potential confounders that may be disguising the true relationship.

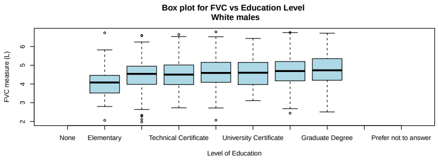
The influence of age, education, income, working status, number of people supported by income, height, smoking status, asthma diagnoses and body mass index are consistent across FEV1 and FVC in the different strata. Multivariate analysis will clarify the impact of the additional social variables, and will allow for a better interpretation of the predictors lung function and the interaction between these factors.

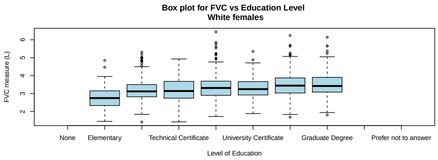
6.2 Graphical Description of Bivariate Relationships

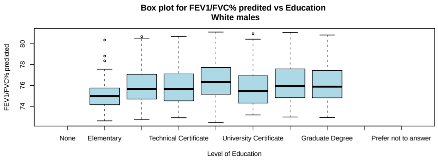
While the ANOVA and summary tables of the bivariate models have identified the several significant associations between SES and FEV1, FVC and FEV1/FVC percent predicted measurements, a closer examination of the key players can help to illustrate the mechanisms by which they work. The following graphical analysis will also help to identify effect modification in the categorical SES variables, as the average means for each factor level are presented in relation to each other. To see summary tables for the means and standard deviations of these important SES predictors, see Appendix I, section 11. Continuous physiological measurements such age, height and body mass index can be modeled in a simple scatterplot to demonstrate the trends of FEV1, FVC and FEV1/FVC percent predicted over the range of the corresponding variable.

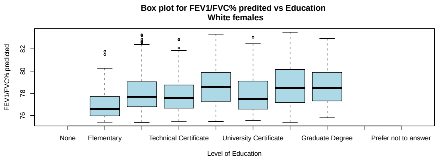

Box plot for FEV1 vs Education Level
White males
FVC measure (L/sec)
1
2
3
4
5
None
Elementary
Technical Certificate
University Certificate
Graduate Degree
Prefer not to answer
Level of Education

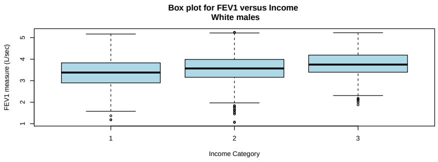

Box plot for FEV1 vs Education Level
White females
FVC measure (L/sec)
1.0
2.0
3.0
4.0
None
Elementary
Technical Certificate
University Certificate
Graduate Degree
Prefer not to answer
Level of Education


Box plot for FVC vs Education Level
White males
FVC measure (L)
2
3
4
5
6
None
Elementary
Technical Certificate
University Certificate
Graduate Degree
Prefer not to answer
Level of Education

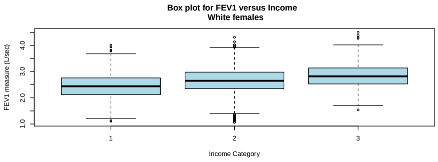

Box plot for FVC vs Education Level
White females
FVC measure (L)
2
3
4
5
6
None
Elementary
Technical Certificate
University Certificate
Graduate Degree
Prefer not to answer
Level of Education


Box plot for FEV1/FVC% predited vs Education
White males
FEV1/FVC% predicted
74
76
78
80
None
Elementary
Technical Certificate
University Certificate
Graduate Degree
Prefer not to answer
Level of Education

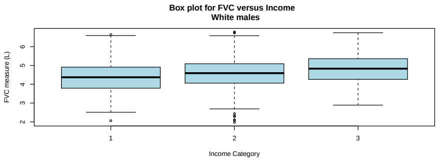

Box plot for FEV1/FVC% predited vs Education
White females
FEV1/FVC% predicted
76
78
80
82
None
Elementary
Technical Certificate
University Certificate
Graduate Degree
Prefer not to answer
Level of Education



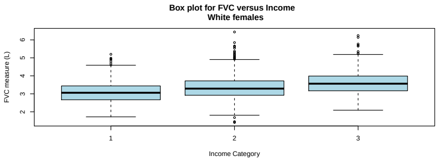
Box plot for FEV1 versus Income
White males



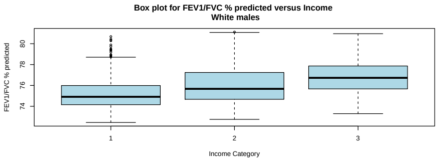
Box plot for FEV1 versus Income
White females



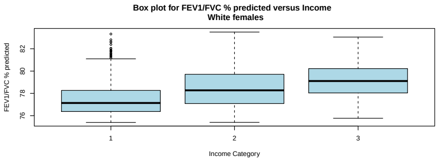
Box plot for FVC versus Income
White males



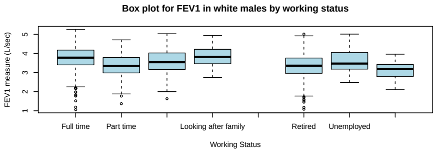
Box plot for FVC versus Income
White females

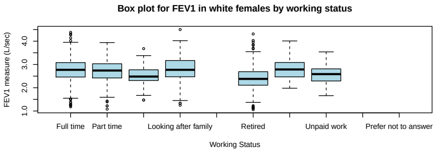


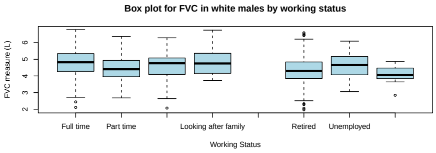
Box plot for FEV1/FVC % predicted versus Income
White males

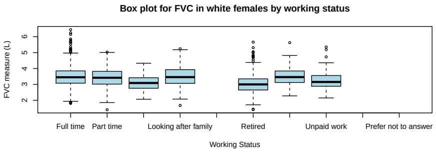


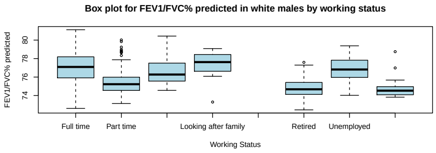
Box plot for FEV1/FVC % predicted versus Income
White females

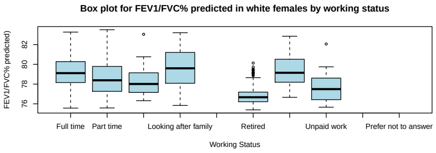

Box plot for FEV1 in white males by working status
FEV1 measure (L/sec)
1
2
3
4
5
Full time
Part time
Looking after family
Retired
Unemployed
Working Status

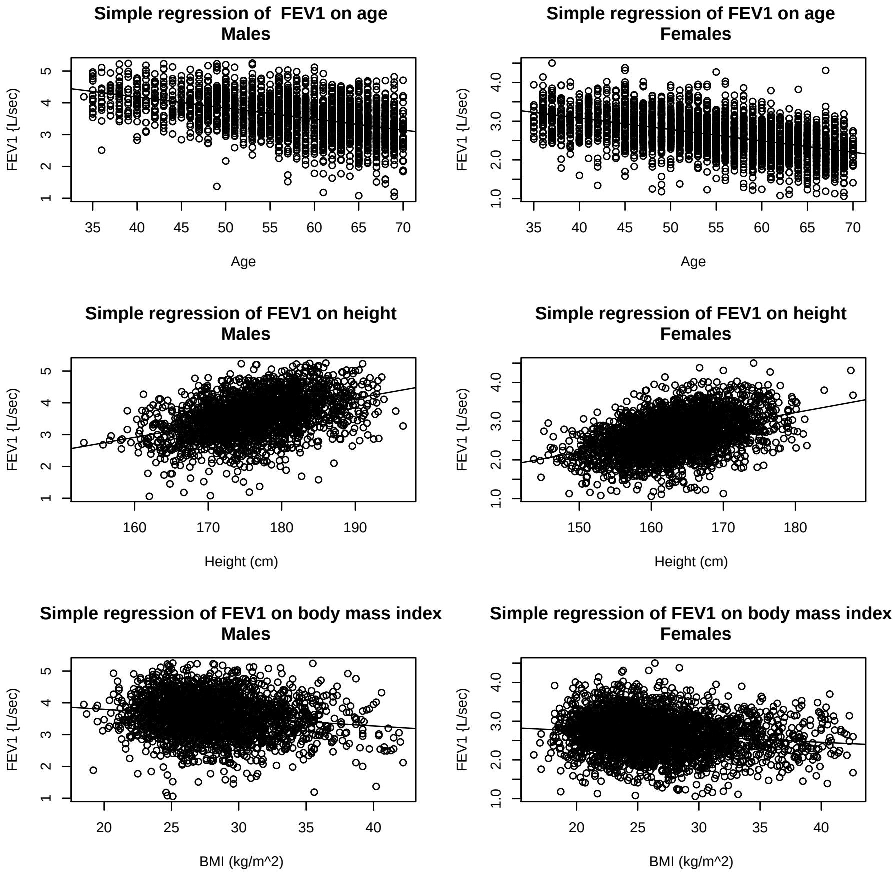

Box plot for FEV1 in white females by working status
FEV1 measure (L/sec)
1.0
2.0
3.0
4.0
Full time
Part time
Looking after family
Retired
Unpaid work
Prefer not to answer
Working Status

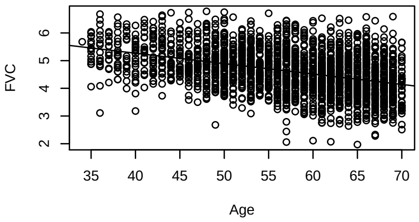

Box plot for FVC in white males by working status
FVC measure (L)
2
3
4
5
6
Full time
Part time
Looking after family
Retired
Unemployed
Working Status


Box plot for FVC in white females by working status
FVC measure (L)
2
3
4
5
6
Full time
Part time
Looking after family
Retired
Unpaid work
Prefer not to answer
Working Status

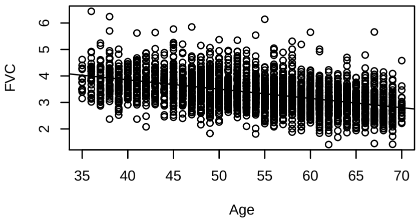

Box plot for FEV1/FVC% predicted in white males by working status
FEV1/FVC% predicted
74
76
78
80
Full time
Part time
Looking after family
Retired
Unemployed
Working Status


Box plot for FEV1/FVC% predicted in white females by working status
FEV1/FVC% predicted)
76
78
80
82
Full time
Part time
Looking after family
Retired
Unpaid work
Prefer not to answer
Working Status

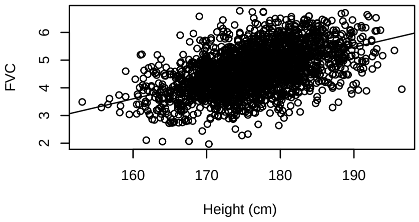

Simple regression of FEV1 on age
Males
FEV1 {L/sec}
1
2
3
4
5
35
40
45
50
55
60
65
70
Age
Simple regression of FEV1 on age
Females
FEV1 {L/sec}
1.0
2.0
3.0
4.0
35
40
45
50
55
60
65
70
Age
Simple regression of FEV1 on height
Males
FEV1 {L/sec}
1
2
3
4
5
160
170
180
190
Height (cm)
Simple regression of FEV1 on height
Females
FEV1 {L/sec}
1.0
2.0
3.0
4.0
150
160
170
180
Height (cm)
Simple regression of FEV1 on body mass index
Males
FEV1 {L/sec}
1
2
3
4
5
20
25
30
35
40
BMI (kg/m^2)
Simple regression of FEV1 on body mass index
Females
FEV1 {L/sec}
1.0
2.0
3.0
4.0
20
25
30
35
40
BMI (kg/m^2)



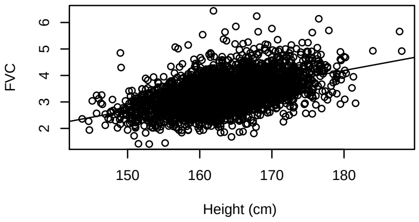
Simple regression of FVC on age
Males



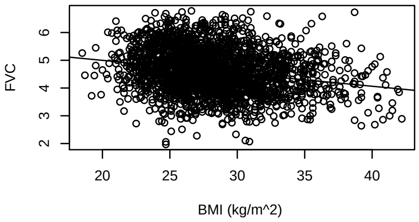
Simple regression of FVC on age
Females



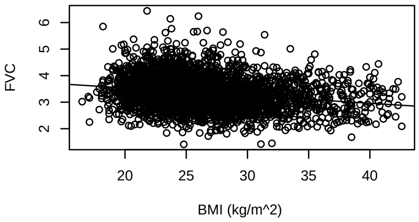
Simple regression of FVC on height
Males



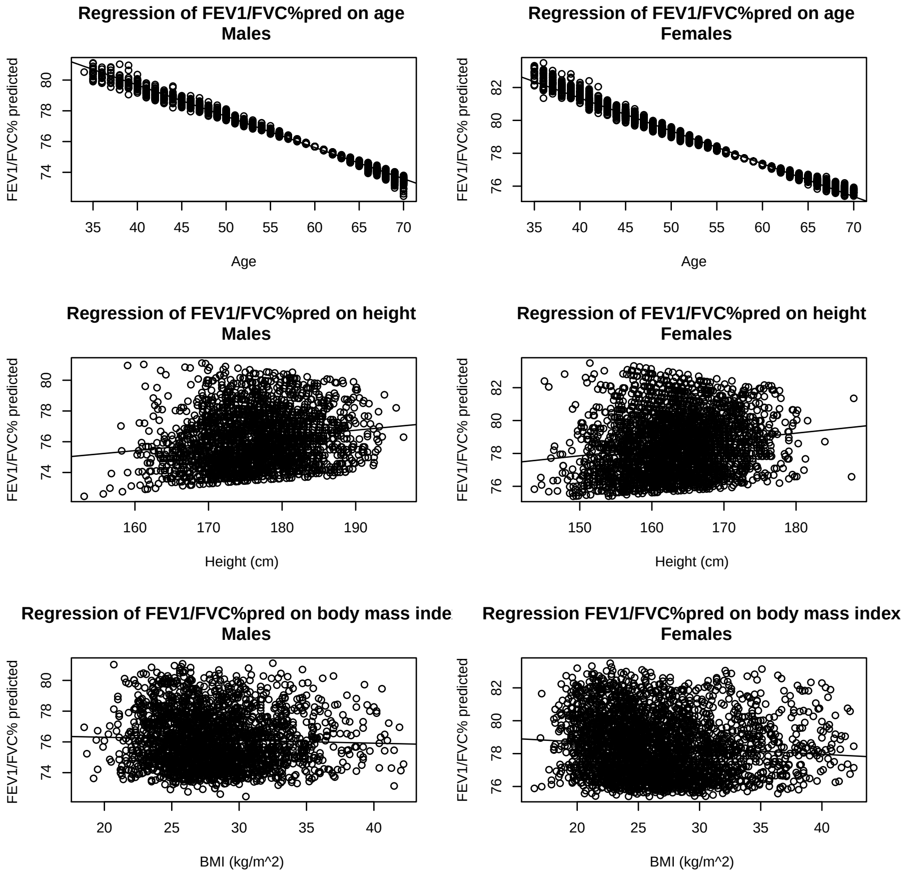
Simple regression of FVC on height
Females



Simple regression of FVC on body mass index
Males



Simple regression of FVC on body mass index
Females


Regression of FEV1/FVC%pred on age
Males
FEV1/FVC% predicted
74
76
78
80
35
40
45
50
55
60
65
70
Age
Regression of FEV1/FVC%pred on age
Females
FEV1/FVC% predicted
76
78
80
82
35
40
45
50
55
60
65
70
Age
Regression of FEV1/FVC%pred on height
Males
FEV1/FVC% predicted
74
76
78
80
160
170
180
190
Height (cm)
Regression of FEV1/FVC%pred on height
Females
FEV1/FVC% predicted
76
78
80
82
150
160
170
180
Height (cm)
Regression of FEV1/FVC%pred on body mass inde
Males
FEV1/FVC% predicted
74
76
78
80
20
25
30
35
40
BMI (kg/m^2)
Regression FEV1/FVC%pred on body mass index
Females
FEV1/FVC% predicted
76
78
80
82
20
25
30
35
40
BMI (kg/m^2)

6.3 Comments on Graphical Exploration

A higher education level is related to better spirometric output, though the relationship is not entirely linear. It seems that the group of individuals with a University certificate had an equal or slightly lower average spirometric reading than those who had a college education. This observation may be a result of confusion with the questionnaire choices and description, with a "University certificate" not necessarily corresponding with more education than those who received a college diploma.

Income category appears to play a role in lung function for both males and females, with a higher income category corresponding to a higher average spirometry measure in FEV1, FVC and FEV1/FVC percent predicted. This trend may be more distinct in females.

The nominal nature of the employment variable makes it more challenging to interpret the box plot, the trend is not as clear. Though the mean respiratory measures for full-time employees tends to be higher, apparently participants who work at home also tend to have some of the highest values of spirometric measures. Particularly in females, it appears as though retirees may produce the lowest spirometric measures, though this relationship is likely related to the age of this particular sample, although males don't show this decrease to the same degree.

It is clear that the same physiological mechanisms contribute to lung function for both white males and females, that older individuals tend to generate lower spirometric readings than younger people, higher body mass index readings typically result in poorer outcomes, and that taller people have greater lung function, on average. When comparing the male and female bivariate relationships, it appears that the fitted lines are relatively similar in slope, which indicates that age, body mass index and height act on lung function in the same way. It is interesting to note the very steep, negative slope for the FEV1/FVC percent predicted regressed on age model - there appears to be a steep decline in this percent predicted value as an individual's age increases.

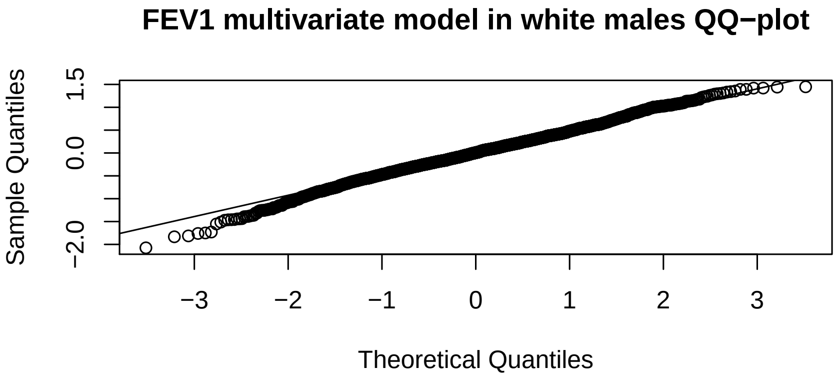
7 Building Epidemiological Models for Lung Function

To continue exploring the relationship between socioeconomic position and lung function, as measured by the spirometric indices FEV1, FVC and the percent of predicted value of FEV1/FVC, multivariate linear regression will be used. Multivariate modeling allows one to investigate the relationship between two independent variables, while accounting for potentially confounding, variables. Constructed separately for males and females, these regressions will include social, economic, demographic and physical variables that will describe the independent influence of each predictor. The bivariate analyses identified age, working status, education level, marital status, household income, number supported by income, height, weight, body mass index, smoking status, most recent visit to the dentist and doctor variables, and presence of diagnosed asthma as a collective of predictors that should be tested in both the male and female models. The bivariate analyses also indicated that assessment centre region play a role in lung function for females, so it will be included as a potential predictor in the female models, though it was not signficant in any of the male bivariate analysis. Quadratic terms of age, weight and height are added to the full model to accommodate natural curvature that is often seen with these physiological measures.

Six models were created in total: the set of models relating FEV1, FVC and the percentage of predicted value of FEV1/FVC to the predictor variables, stratified by sex. A model with all of the potential predictors was first established for each outcome. The continuous variables were centralized by subtracting a rounded mean, which addresses potential multicollinearity between the predictors.ANOVA tables and model summaries were generated for each of the models, to identify non-significant variables ($p < 0.05$). Using the *update* function in R, the covariate with the greatest p-value was dropped from the model, and the resulting model was then summarized. A likelihood ratio test was used to determine whether the updated model was a significant change from the initial model - if the variable that was dropped explained a significant amount of variation, then the differences between the models will be significant. This process continued until all of the predictors were significantly associated with the respective outcome. At this point, variables were added back to the model, one at a time, and tested for significant improvement to the model using the likelihood ratio test. Somewhat analogous to backwards elimination and stepwise regression,this method will ideally lead to the most parsimonious and descriptive model. The results of this model selection with FEV1, FVC and percent predicted FEV1/FVC are shown below, along with the Q-Q plots of the model and a plot of residuals versus the model's fitted values. Discussion of the coefficients and diagnostics is below. See Appendix III for the R script for this model selection.

Table 10: FEV1 in white males from the OHS Pilot

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.5908	0.0721	49.79	0.0000
I(AGE.CALCULATED - 60)	-0.0327	0.0024	-13.45	0.0000
CURRENT_SITUATION2	-0.1371	0.0444	-3.09	0.0020
CURRENT_SITUATION3	-0.1020	0.0888	-1.15	0.2506
CURRENT_SITUATION4	0.1257	0.1468	0.86	0.3919
CURRENT_SITUATION6	-0.0157	0.0309	-0.51	0.6118
CURRENT_SITUATION7	-0.1447	0.0705	-2.05	0.0402
CURRENT_SITUATION8	-0.2289	0.1334	-1.72	0.0863
HIGHEST_LEVEL_COMPLETED2	0.1211	0.0694	1.75	0.0811
HIGHEST_LEVEL_COMPLETED3	0.1702	0.0712	2.39	0.0170
HIGHEST_LEVEL_COMPLETED4	0.1509	0.0700	2.16	0.0312
HIGHEST_LEVEL_COMPLETED5	0.2042	0.0838	2.44	0.0149
HIGHEST_LEVEL_COMPLETED6	0.1863	0.0690	2.70	0.0070
HIGHEST_LEVEL_COMPLETED7	0.1606	0.0710	2.26	0.0238
MARITAL_STATUS2	0.0612	0.0417	1.47	0.1420
MARITAL_STATUS3	-0.0939	0.0624	-1.50	0.1326
HOUSE_INCOME_LAST_YEAR2	-0.0078	0.0340	-0.23	0.8173
HOUSE_INCOME_LAST_YEAR3	0.0165	0.0418	0.40	0.6924
I(INPUT_PART_HEIGHT_SP - 180)	0.0366	0.0019	19.57	0.0000
I(RES_WEIGHT_BIO - 90)	-0.0041	0.0009	-4.43	0.0000
as.factor(SMOKE_STATUS)1	-0.0363	0.0229	-1.59	0.1129
as.factor(SMOKE_STATUS)2	-0.2559	0.0471	-5.43	0.0000
ASTHMA_OCCURRENCE	-0.3444	0.0380	-9.06	0.0000
I((AGE.CALCULATED - 60)^2)	-0.0003	0.0001	-2.01	0.0450
I((RES_WEIGHT_BIO - 90)^2)	-0.0001	0.0000	-3.01	0.0027


FEV1 multivariate model in white males QQ−plot
Sample Quantiles
1.5
0.0
−2.0
−3
−2
−1
0
1
2
3
Theoretical Quantiles

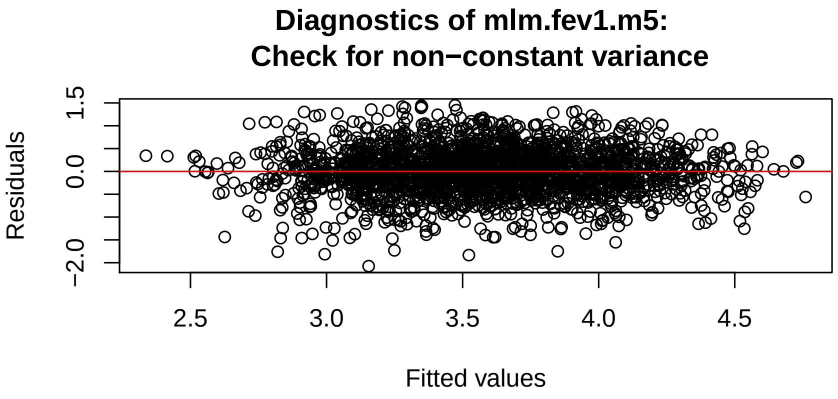
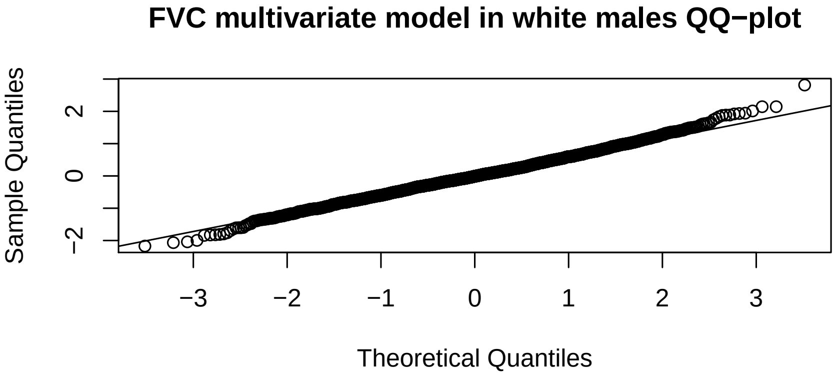

Diagnostics of mlm.fev1.m5:
Check for non−constant variance
Residuals
1.5
0.0
−2.0
2.5
3.0
3.5
4.0
4.5
Fitted values

Table 11: FVC in white males from the OHS Pilot

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.6268	0.0863	53.62	0.0000
I(AGE.CALCULATED - 60)	-0.0271	0.0020	-13.58	0.0000
CURRENT_SITUATION2	-0.1487	0.0536	-2.77	0.0056
CURRENT_SITUATION3	-0.0999	0.1075	-0.93	0.3531
CURRENT_SITUATION4	0.1785	0.1781	1.00	0.3163
CURRENT_SITUATION6	-0.0606	0.0365	-1.66	0.0972
CURRENT_SITUATION7	-0.1614	0.0853	-1.89	0.0588
CURRENT_SITUATION8	-0.3124	0.1615	-1.93	0.0532
HIGHEST_LEVEL_COMPLETED2	0.1752	0.0841	2.08	0.0373
HIGHEST_LEVEL_COMPLETED3	0.2104	0.0864	2.43	0.0150
HIGHEST_LEVEL_COMPLETED4	0.2028	0.0849	2.39	0.0169
HIGHEST_LEVEL_COMPLETED5	0.2892	0.1017	2.84	0.0045
HIGHEST_LEVEL_COMPLETED6	0.2179	0.0837	2.60	0.0093
HIGHEST_LEVEL_COMPLETED7	0.2152	0.0862	2.50	0.0126
MARITAL_STATUS2	0.0790	0.0505	1.56	0.1184
MARITAL_STATUS3	-0.0427	0.0757	-0.56	0.5729
HOUSE_INCOME_LAST_YEAR2	-0.0012	0.0410	-0.03	0.9762
HOUSE_INCOME_LAST_YEAR3	0.0367	0.0503	0.73	0.4657
I(INPUT_PART_HEIGHT_SP - 180)	0.0628	0.0022	28.06	0.0000
I(RES_WEIGHT_BIO - 90)	-0.0101	0.0011	-9.10	0.0000
as.factor(SMOKE_STATUS)1	-0.0422	0.0278	-1.52	0.1291
as.factor(SMOKE_STATUS)2	-0.1755	0.0571	-3.07	0.0022
ASTHMA_OCCURRENCE	-0.2686	0.0461	-5.83	0.0000


FVC multivariate model in white males QQ−plot
Sample Quantiles
−2
0
2
−3
−2
−1
0
1
2
3
Theoretical Quantiles

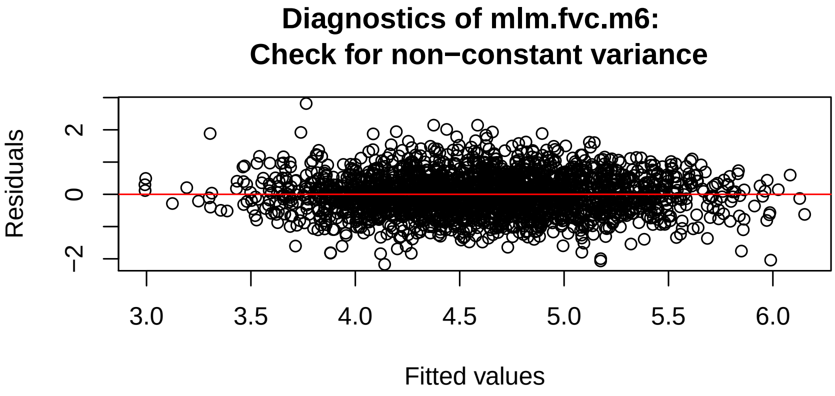
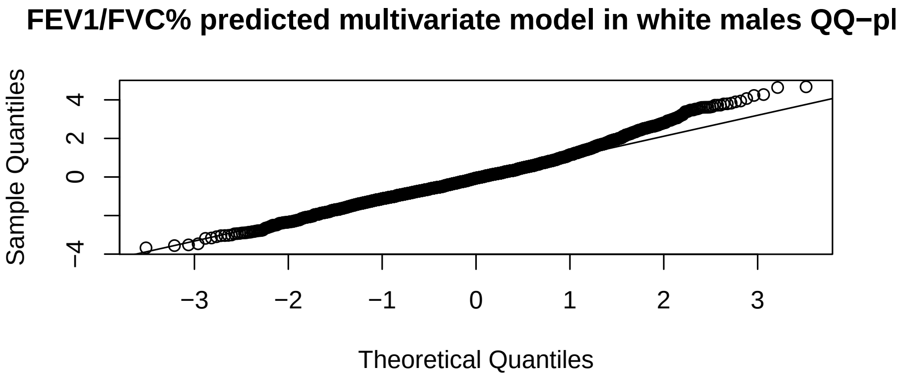

Diagnostics of mlm.fvc.m6:
Check for non−constant variance
Residuals
−2
0
2
3.0
3.5
4.0
4.5
5.0
5.5
6.0
Fitted values

Table 12: FEV1/FVC% predicted in white males from the OHS Pilot

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	75.1443	0.1866	402.75	0.0000
CURRENT_SITUATION2	-1.2241	0.1055	-11.61	0.0000
CURRENT_SITUATION3	-0.1271	0.2171	-0.59	0.5582
CURRENT_SITUATION4	-0.1569	0.3597	-0.44	0.6628
CURRENT_SITUATION6	-1.7456	0.0648	-26.96	0.0000
CURRENT_SITUATION7	-0.1370	0.1725	-0.79	0.4272
CURRENT_SITUATION8	-1.7547	0.3238	-5.42	0.0000
HIGHEST_LEVEL_COMPLETED2	0.2648	0.1697	1.56	0.1188
HIGHEST_LEVEL_COMPLETED3	0.3898	0.1742	2.24	0.0254
HIGHEST_LEVEL_COMPLETED4	0.5203	0.1710	3.04	0.0024
HIGHEST_LEVEL_COMPLETED5	0.0640	0.2053	0.31	0.7552
HIGHEST_LEVEL_COMPLETED6	0.3615	0.1690	2.14	0.0325
HIGHEST_LEVEL_COMPLETED7	0.1797	0.1741	1.03	0.3019
MARITAL_STATUS2	0.3867	0.1049	3.69	0.0002
MARITAL_STATUS3	1.2257	0.1597	7.67	0.0000
HOUSE_INCOME_LAST_YEAR2	0.2348	0.0829	2.83	0.0047
HOUSE_INCOME_LAST_YEAR3	0.2833	0.1022	2.77	0.0056
NUMBER_SUPPORTED_BY_INCOME	0.3711	0.0251	14.80	0.0000
I(RES_WEIGHT_BIO - 90)	0.0138	0.0041	3.33	0.0009
I(RES_BODY_MASS_INDEX - 30)	-0.0492	0.0140	-3.51	0.0005
as.factor(SMOKE_STATUS)1	-0.3339	0.0557	-6.00	0.0000
as.factor(SMOKE_STATUS)2	0.1385	0.1158	1.20	0.2316
LAST_ROUTINE_MEDICAL_EXAM0	0.2522	0.4741	0.53	0.5948
LAST_ROUTINE_MEDICAL_EXAM2	0.1408	0.0646	2.18	0.0295
LAST_ROUTINE_MEDICAL_EXAM3	0.2830	0.0717	3.95	0.0001
LAST_ROUTINE_MEDICAL_EXAM4	0.4958	0.1176	4.22	0.0000
LAST_ROUTINE_MEDICAL_EXAM5	0.7449	0.1089	6.84	0.0000
ASTHMA_OCCURRENCE	0.1549	0.0931	1.66	0.0962


FEV1/FVC% predicted multivariate model in white males QQ−pl
Sample Quantiles
−4
0
2
4
−3
−2
−1
0
1
2
3
Theoretical Quantiles

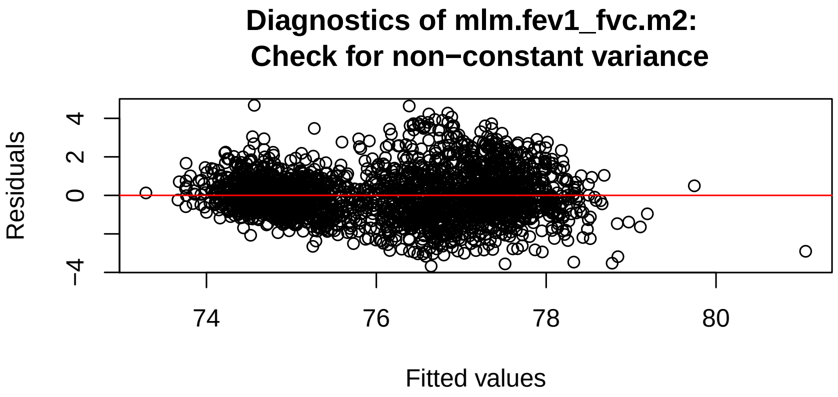
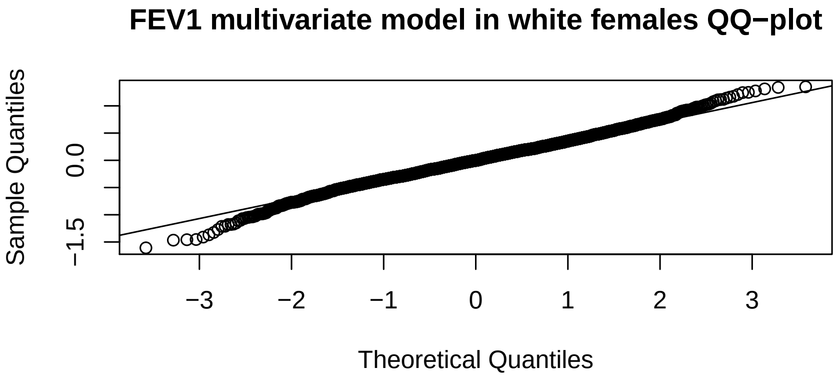
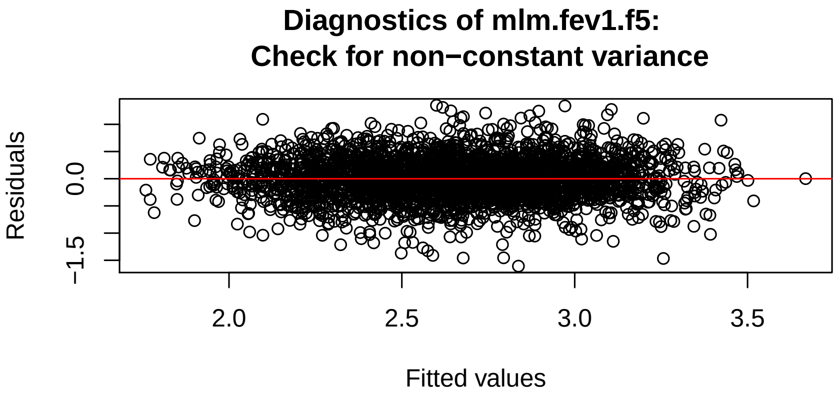

Diagnostics of mlm.fev1_fvc.m2:
Check for non−constant variance
Residuals
−4
0
2
4
74
76
78
80
Fitted values

Table 13: FEV1 in white females from the OHS Pilot

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.6565	0.0688	38.63	0.0000
I(AGE.CALCULATED - 55)	-0.0271	0.0013	-21.04	0.0000
CURRENT_SITUATION2	0.0298	0.0213	1.40	0.1617
CURRENT_SITUATION3	-0.0185	0.0540	-0.34	0.7317
CURRENT_SITUATION4	-0.0230	0.0307	-0.75	0.4536
CURRENT_SITUATION6	0.0469	0.0232	2.02	0.0437
CURRENT_SITUATION7	0.0483	0.0473	1.02	0.3068
CURRENT_SITUATION8	0.0308	0.0568	0.54	0.5880
HIGHEST_LEVEL_COMPLETED2	0.0444	0.0639	0.70	0.4870
HIGHEST_LEVEL_COMPLETED3	-0.0218	0.0701	-0.31	0.7562
HIGHEST_LEVEL_COMPLETED4	0.0647	0.0637	1.02	0.3096
HIGHEST_LEVEL_COMPLETED5	0.0505	0.0707	0.71	0.4754
HIGHEST_LEVEL_COMPLETED6	0.0687	0.0639	1.07	0.2825
HIGHEST_LEVEL_COMPLETED7	0.0672	0.0663	1.01	0.3107
HOUSE_INCOME_LAST_YEAR2	0.0198	0.0202	0.98	0.3259
HOUSE_INCOME_LAST_YEAR3	0.0793	0.0265	2.99	0.0028
NUMBER_SUPPORTED_BY_INCOME	0.0136	0.0067	2.02	0.0431
I(INPUT_PART_HEIGHT_SP - 165)	0.0068	0.0076	0.90	0.3698
I(RES_WEIGHT_BIO - 70)	0.0230	0.0087	2.62	0.0087
I(RES_BODY_MASS_INDEX - 25)	-0.0621	0.0229	-2.72	0.0066
as.factor(SMOKE_STATUS)1	0.0119	0.0150	0.79	0.4293
as.factor(SMOKE_STATUS)2	-0.0656	0.0331	-1.98	0.0477
ASTHMA_OCCURRENCE	-0.1472	0.0226	-6.51	0.0000
I((AGE.CALCULATED - 55)^2)	-0.0004	0.0001	-3.77	0.0002
I((RES_WEIGHT_BIO - 70)^2)	-0.0001	0.0000	-4.10	0.0000



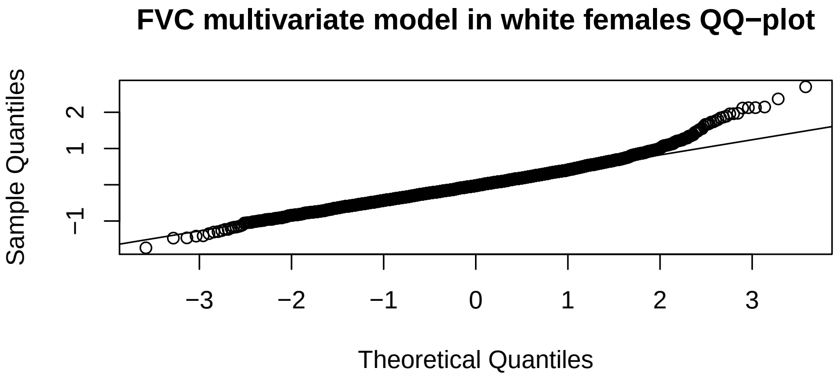
FEV1 multivariate model in white females QQ−plot



Diagnostics of mlm.fev1.f5:
Check for non−constant variance

Table 14: FVC in white females from the OHS Pilot

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.1805	0.0870	36.57	0.0000
I(AGE.CALCULATED - 60)	-0.0281	0.0016	-17.41	0.0000
CURRENT_SITUATION2	0.0212	0.0264	0.80	0.4211
CURRENT_SITUATION3	-0.0630	0.0668	-0.94	0.3455
CURRENT_SITUATION4	-0.0557	0.0382	-1.46	0.1446
CURRENT_SITUATION6	0.0389	0.0290	1.34	0.1801
CURRENT_SITUATION7	0.0318	0.0586	0.54	0.5880
CURRENT_SITUATION8	0.0512	0.0703	0.73	0.4665
HIGHEST_LEVEL_COMPLETED2	0.0599	0.0790	0.76	0.4486
HIGHEST_LEVEL_COMPLETED3	0.0061	0.0867	0.07	0.9439
HIGHEST_LEVEL_COMPLETED4	0.0783	0.0788	0.99	0.3204
HIGHEST_LEVEL_COMPLETED5	0.0952	0.0876	1.09	0.2772
HIGHEST_LEVEL_COMPLETED6	0.1059	0.0792	1.34	0.1813
HIGHEST_LEVEL_COMPLETED7	0.0698	0.0823	0.85	0.3966
as.factor(MARITAL_STATUS)2	0.0223	0.0252	0.88	0.3772
as.factor(MARITAL_STATUS)3	0.1171	0.0417	2.80	0.0051
HOUSE_INCOME_LAST_YEAR2	0.0098	0.0262	0.38	0.7073
HOUSE_INCOME_LAST_YEAR3	0.0876	0.0348	2.52	0.0119
NUMBER_SUPPORTED_BY_INCOME	0.0297	0.0091	3.27	0.0011
I(INPUT_PART_HEIGHT_SP - 165)	0.0262	0.0094	2.80	0.0052
I(RES_WEIGHT_BIO - 70)	0.0174	0.0108	1.61	0.1075
I(RES_BODY_MASS_INDEX - 25)	-0.0580	0.0283	-2.05	0.0401
as.factor(SMOKE_STATUS)1	0.0292	0.0186	1.57	0.1162
as.factor(SMOKE_STATUS)2	-0.0162	0.0411	-0.39	0.6944
ASTHMA_OCCURRENCE	-0.0881	0.0279	-3.15	0.0016
Admin.Participant.siteNoOWENS	-0.0476	0.0232	-2.05	0.0400
Admin.Participant.siteNoSUDBUR	-0.0687	0.0223	-3.08	0.0021
I((AGE.CALCULATED - 55)^2)	-0.0003	0.0001	-2.19	0.0283
I((RES_WEIGHT_BIO - 70)^2)	-0.0001	0.0000	-3.75	0.0002


FVC multivariate model in white females QQ−plot
Sample Quantiles
2
1
−1
−3
−2
−1
0
1
2
3
Theoretical Quantiles

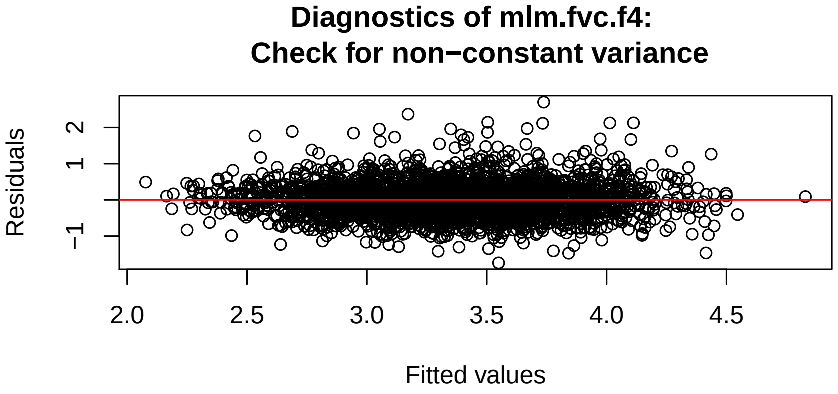
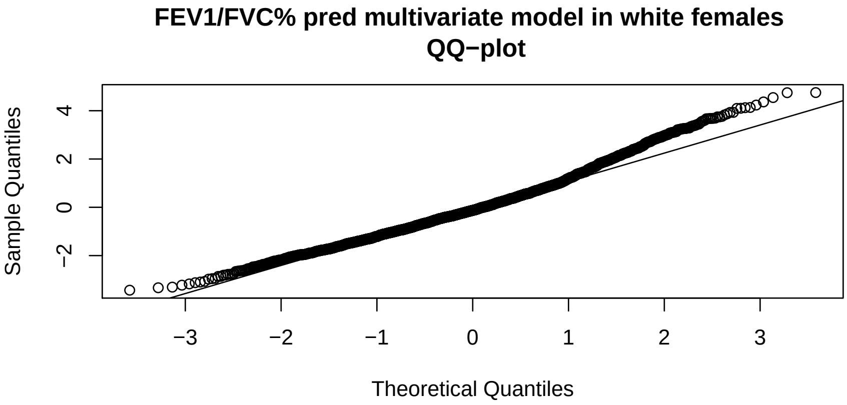
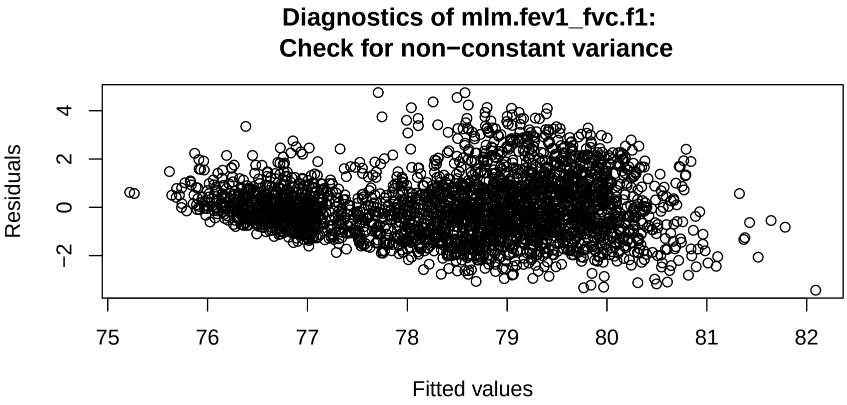

Diagnostics of mlm.fvc.f4:
Check for non−constant variance
Residuals
2
1
−1
2.0
2.5
3.0
3.5
4.0
4.5
Fitted values

Table 15: FEV1/FVC% predicted in white females from the OHS Pilot

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	77.3637	0.2294	337.21	0.0000
CURRENT_SITUATION2	-0.5894	0.0701	-8.40	0.0000
CURRENT_SITUATION3	-0.7032	0.1795	-3.92	0.0001
CURRENT_SITUATION4	-0.0198	0.1026	-0.19	0.8469
CURRENT_SITUATION6	-1.9253	0.0628	-30.65	0.0000
CURRENT_SITUATION7	0.2147	0.1581	1.36	0.1745
CURRENT_SITUATION8	-1.3801	0.1872	-7.37	0.0000
HIGHEST_LEVEL_COMPLETED2	0.2156	0.2136	1.01	0.3129
HIGHEST_LEVEL_COMPLETED3	0.1329	0.2345	0.57	0.5709
HIGHEST_LEVEL_COMPLETED4	0.4125	0.2131	1.94	0.0530
HIGHEST_LEVEL_COMPLETED5	0.2580	0.2371	1.09	0.2767
HIGHEST_LEVEL_COMPLETED6	0.5007	0.2141	2.34	0.0194
HIGHEST_LEVEL_COMPLETED7	0.2473	0.2226	1.11	0.2667
as.factor(MARITAL_STATUS)2	-0.0263	0.0681	-0.39	0.6994
as.factor(MARITAL_STATUS)3	0.9444	0.1112	8.49	0.0000
HOUSE_INCOME_LAST_YEAR2	0.1291	0.0706	1.83	0.0676
HOUSE_INCOME_LAST_YEAR3	0.1340	0.0937	1.43	0.1527
NUMBER_SUPPORTED_BY_INCOME	0.4475	0.0229	19.55	0.0000
I(RES_WEIGHT_BIO - 70)	0.0107	0.0046	2.34	0.0193
I(RES_BODY_MASS_INDEX - 25)	-0.0602	0.0126	-4.79	0.0000
as.factor(SMOKE_STATUS)1	-0.0861	0.0501	-1.72	0.0858
as.factor(SMOKE_STATUS)2	0.1032	0.1112	0.93	0.3538
LAST_VISIT_DENTIST0	-0.7608	0.7323	-1.04	0.2989
LAST_VISIT_DENTIST2	0.1490	0.0585	2.55	0.0109
LAST_VISIT_DENTIST3	0.2796	0.1120	2.50	0.0126
LAST_VISIT_DENTIST4	0.5273	0.1824	2.89	0.0039
LAST_VISIT_DENTIST5	0.2253	0.1670	1.35	0.1774
ASTHMA_OCCURRENCE	0.3772	0.0751	5.03	0.0000
Admin.Participant.siteNoOWENS	-0.0723	0.0626	-1.16	0.2480
Admin.Participant.siteNoSUDBUR	0.1946	0.0599	3.25	0.0012
I((RES_WEIGHT_BIO - 70)^2)	0.0003	0.0001	3.82	0.0001


FEV1/FVC% pred multivariate model in white females
QQ−plot
Sample Quantiles
−2
0
2
4
−3
−2
−1
0
1
2
3
Theoretical Quantiles


Diagnostics of mlm.fev1_fvc.f1:
Check for non−constant variance
Residuals
−2
0
2
4
75
76
77
78
79
80
81
82
Fitted values

8 Interpretation of Models

Though the absolute values of spirometric indices are significantly different between males and females, this result can be attributed primarily to physical differences between the sexes - the bivariate analyses showed that lung function shows a similar response to changes in height, body mass index and age in both sexes. The objective of this analysis included identifying the socioeconomic variables which are significantly associated with spirometric outcomes, and the strength and direction of each effect. By stratifying by males and females, it will be easier to see and describe how the socioeconomic variables affect the outcome, but also to note parallels and differences between the genders.

8.1 FEV1 in white males and females of Ontario

While adjusting for physical differences (e.g., age, height, and weight), it seems that thel social and demographic factors that influence lung function differ between males and females. For males, there appear to be significant differences in FEV1 measures between those who are working full-time and those who are not: on average, males who are not working full-time may have somewhere between a 0.1 to 0.2 litres per second lower FEV1 measure than those who are working (with the other factors held constant), with the exception of individuals who look after the home. Educational level also seems to have a large impact on average FEV1 values in males, with higher levels of education corresponding to a higher FEV1 output. Comparatively, working status and education level do not have as great of an effect in the white female cohort. However, household income and the number supported by this income seems to play a greater role in female FEV1 measure. Common to both genders,as expected, average FEV1 decreases in current smokers and people with an asthma diagnosis, though the strength of the effect of both quantities is greater in males.

8.2 FVC in white males and females of Ontario

After adjusting for age, height, and weight in the multivariate model, similar socioeconomic factors appear to influence FVC measurements as FEV1, which was expected. In males, individuals who were not either working full-time or looking after the family had a lower average FVC measurement, with the most notable difference experienced by individuals who volunteer who have approximately 0.31 litres lower FVC on average than those who work full-time. Higher education levels were significantly correlated with increased FVC measures in males, as with FEV1. Females, on the other hand, do not seem as sensitive to changes in education level and working status, though, as noted in the FEV1 model, reported household income, marital status and area of residence appear to be more important determinants in average FVC values. In the female cohort, a shift to a higher income category would result in a small increase in FVC output (approximately 0.088 L difference between women in the lowest income category and those in the highest). As well, females who were never married experience a significant increase in FVC, than their currently married or once-married counterparts. On the contrary, though not a significant association, single males may actually have a lower average FVC. The assessment centre location is also significantly associated with FVC in this female cohort as well. Subjects from an urban centre (Mississauga) have a higher average FVC output than females in a rural area (Owen Sound) or northern community (Sudbury). Mirroring the FEV1 analysis, diagnosed asthma is significantly associated with a lower FVC measure, and current smokers are also likely to have lower FVC.

8.3 FEV1/FVC percentage of predicted values

It is interesting to note that the socioeconomic parallels between males and females in the model summaries of FEV1/FVC percent of predicted, arguably the most clinically important indicator of lung function. In both the FEV1 and FVC analyses, education and working status played a bigger overall role in determining the outcome in males, while female lung function seemed to be affected more by household income and familial characteristics, such as marital status and location of residence. With the FEV1/FV1 percent predicted value, the defined socioeconomic indicators have similar impact in both sexes. The models included adjustment for weight and BMI, but because the reference equations included age and height terms, these variables were already accounted for the in outcome measures. Individuals who reported working full-time had significantly higher FEV1/FVC percent predicted than any other job classification, with the exception of unemployed females, who experience approximately a 0.21 increase in FEV1/FVC percent predicted value than their working counterparts (this association was not significant, p=0.175). The influence of education remains consistent with the previous analyses for both men and women - any education beyond elementary school will likely lead to increased lung function. Marital status in both males and females seems to affect lung function, with participants who were never married demonstrating a greater FEV1/FVC percent predicted measure. Higher reported household incomes seems to correspond to higher outcome values as well, as well as number supported by income (e.g., the larger the number of people supported by the household income, the greater the expected FEV1/FVC percent of predicted measurement). Ex-smokers in both genders experienced significantly lower average predicted values, though the model doesn't indicate a significant decrease for current smokers which could be due to small sample size in the current smoking category (135 and 146 for males and females, respectively), but should be looked into further because of the obvious relationship that smoking has with lung function. Interestingly, both male and female FEV1/FVC outcome models show significant association with variables that contain information about health care access: LAST_ROUTINE_MEDICAL_EXAM (male model) and LAST_VISIT_DENTIST (female). Males who visit their doctors less frequently have a greater FEV1/FVC outcome measure than those who have had a routine check-up in the past six months. This could be related to other health problems: males may see their doctors or have a check-up more recently if they have known health problems, implying that those who have not seen a doctor recently may be healthier overall. A similar trend is noted in the female cohort: subjects who have not seen their dentist in the past 6 months tend to have a higher FEV1/FVC predicted value, with the exception of women who have never visited a dentist. Again, assessment centre location (a proxy variable for location of residence) is significantly associated with the outcome in the female cohort, this time with Sudbury residents experiencing an increased FEV1/FVC percent of predicted value. These multivariate models suggest that individuals with an asthma diagnosis are significantly associated with a greater FEV1/FVC predicted value (approximately a 0.15% increase for males, and a 0.38% increase for females, p =0.09 in males and p<0.001 in females). This relationship requires deeper investigation because this finding is defies current biological understanding of FEV1/FVC percent predicted.

8.4 Diagnostics Interpretation

Model diagnostics are a good way to evaluate the fit of a model by checking the assumptions of multivariate linear regression. Because the data was cleaned prior to analysis, looking for outliers would be redundant, hence the goal of the diagnostic plots for the six model will be identification of non-constant variance in the residuals and non-normal residuals. The quantile-quantile (Q-Q) plot will address the assumption of normality: if the residuals deviate signficantly from the y=x line, then the normality assumption does not hold. In the residuals versus fitted values plot for a model, there should be a random scattering of point above and below the line $\epsilon = 0$, and almost all of the points should be within $2\hat{\sigma}$ of 0. The Q-Q plots for the six models generally look like the model assumptions are reasonable - the residuals tend to fall along the normal line, though in the female FVC final model plot, there seems to be slight deviation from the normal in the upper quantiles, which is also reflected somewhat in the female FEV1/FVC percent predited Q-Q plot. Perhaps this occurrence is worth investigating further, but this slight deviation does not negate the assumption of normality. In the residuals versus fitted values plots, the constant variance assumptions seem to hold true for all of the outcomes.

8.5 Summary of Model Results

It is evident that in the six models, SES has an effect on all three spirometric indices (FEV1, FVC, and percentage of predicted value for FEV1/FVC), after adjusting for related physiological (age, height, weight and body mass index) and health variables (asthma diagnoses and current smoking status), and the relationships of these variables to the outcome differ between males and females. The conventional measures of SES, income, education level and working status, proved to be significantly associated with lung function; a higher education, higher income and/or full-time employment corresponded to improved lung function, as measured by the described indicators. Different measures were emphasized between males and females: male lung function was more responsive to educational and working variables, whereas improved female lung function was associated with higher income and marital status. FEV1/FVC percent predicted models demonstrated that the SES variables that have an impact on lung function act in a similar way in males and females, with the same strength of effect. In addition to the conventional SES measures described above, marital status, number supported by income, health care visits and region of residence are significantly associated with changes in average lung function. The exploratory nature of these analyses means that a deeper investigation into these relationships may provide greater insight into the true socioeconomic forces at work. As well, the differences in the socioeconomic indices between genders should be examined further.

9 Further Investigation and Analysis

9.1 Extension of Multivariate Analysis

As demonstrated in the above interpretation, there are several routes that can be taken following the multivariate analysis of SES and lung function. The continuation of this particular study could involve closely examining the differences in SES between males and females as it relates to lung function, perhaps giving futher consideration to the impact of assessment centre (location of residence) and additional area-level variables. Census information for each region could be related to each individual and considered as another level of SES, particularly relevant to the female group who demonstrated a significant relationship with this variable. Additional subanalyses could be conducted within the male and female samples to look at the mechanisms of SES within particular subgroups, such as overweight and obese individuals, people with doctor-diagnosed asthma, as well as the other ethnicities in the study, though there would be concern about a lack of power for certain ethnicities in this cohort. Looking at the same OHS cohort and different lung function or physical measures outcomes, such as forced expiratory flow when the lungs are at 50% capacity ($FEF_5 0$), or other physical measures like blood pressure, with respect to SES would be interesting as well. It would be beneficial to have a deeper look at the FEV1/FVC percent predicted values as well as the absolute ratios, however, the distribution of the outcome values does not necessarily fit with previous research, which implies that more a more complex analysis would be required to obtain accurate esimates for SES assocation.

To take the univariate analysis further with the OHS Pilot data, it would be interesting to obtain Statistics Canada ecological descriptive statistics from community and health profiles to implement an in-depth comparison of frequencies and prevalences with the OHS Pilot study results. This process would get at the heart of statistical analysis by evaluating how well a sample of the population represents the entire population, and would be useful for describing generalizability of the OHS Pilot results.

Because socioeconomic position and its relationship to health outcomes is a focus of current health research, creating an aggregate variable for SES would provide insight into the multidimensional nature of SES, but would also be helpful for future OHS analyses.

As mentioned previously in the report, a study of IPAQ results and influence of missing values on the interpretation of OHS Pilot data should be studied further, particularly in a sensitivity analysis. As well, it is necessary to investigate the FEV1/FVC percent predicted models further to determine why asthma has counterintuitive associations with the outcome.

9.2 Multiple Correspondence Analysis: Quantification of SES : The next step

After conducting a literature review on different socioeconomic variables used in Canadian population health research, the derived variables used in different cohorts didn't seem to translate well into analysis of SES with the OHS Pilot data. The material deprivation and social deprivation indices, used by Statistics Canadxa and Canadian Institute of Health Information, addressed SES at the individual level and regional level and contained information comparable to that gathered in the OHS Pilot study, but the income was represented continuously in these studies while the rest of their SES variables are binary. After identifying relevant SES variables from the OHS Pilot, all of the potentially important predictors of SES are categorical. Consequently, these previously validated SES forms cannot be used in the OHS analysis, nor can principal components analysis - the method of derivation of this SES structure - be used to derive similar factors for the OHS Pilot analysis.

Principal component analysis (PCA) is a method that can reduce the dimension of the indicator space by creating new variables - "components" - that are a linear combination of the original variables. The first linear combination explains the greatest amount of variation among the original variables, and the second combination, constructed to be orthogonal (uncorrelated) with the first component, explains the next greatest amount of variation in the data. Each component is orthogonal with the preceding component, and the number of components selected should strike a balance between information explained and minimizing the number of new variables. The variance explained by the identified variables is known as the total inertia. Mathematically, this is equivalent to the total Pearson chi-square for the two-way contingency table divided by the number of observations total. After standardizing the variables, PCA can be performed using eigenvalue/ singular-value decomposition of a data covariance matrix. Used as an exploratory analysis, the final results can be interpreted as the summary scores of the linear combinations that make up the principal component, but the factor loadings, the weight of the original variable in a particular principal component, are interesting as well. For a multidimensional concept like SES that's not clearly quantitatively defined, this decomposition method is particularly useful as it can take several identified SES predictors and construct a score based on all measured components. Additionally, researchers can identify which factors contribute the most to different indices of SES. Due to the categorical nature of the SES variables in the OHS Pilot study, PCA cannot be performed to create SES "scores" as in the Statistics Canada study. The OHS Pilot study obtained more in-depth information than described by the Statistics Canada SES indices, and conducting an OHS-specific correspondence analysis may contribute to a better understanding of predictors of SES in Ontario.

The multi-faceted nature of SES and the ordinal nature of the related variables in the OHS Pilot study demand a technique like Multiple Correspondence Analysis (MCA). Correspondence analysis is a descriptive method for contingency table analysis that provides a measure of correspondence between the rows and columns. A simple correspondence analysis (only two categorical variables) would produce a cross tabulation table of relative frequencies which represent the Euclidean distances between individual rows and/or columns in a low-dimensional space. The objective of MCA involves describing a lower-dimensional space that will retain almost all of the information about the differences between the rows. [19]

A mathmematical primer on MCA : Assume that there are k categorical ordinal indicators, with each indicator I_k having J_k categories, with each J_k binary variable corresponding to each category of the indicator variable. MCA uses the following notation:

- $\mathbf{X_{n,J}}$: the matrix of n observations on the K indicators decomposed into J_k variables, where $J = \sum_{k=1}^{K} J_k$ is the total number of categories. This is the correspondence matrix.
- n_j : the absolute frequency of category J; the column mass of $\mathbf{X}$.
- n : the sum of the elements of matrix $\mathbf{X}$; e.g., $n \pm K$.
- $f_j = \dfrac{n_j}{n}$: the relative frequency of category j.
- $f_j^i = \dfrac{X_{i,j}}{X_{i.}}$, where $\mathbf{X_{i.}}$ is the row sum; the row mass. The set $f_J^i = \{f_j^i, j = 1, J\}$ is called the *profile* of observation i.

MCA applies the PCA methodology to the matrix $\mathbf{X}$ - to the set of the J binary variables in the $\mathbf{R}^n$ space, but with the χ^2-metric on row and column profiles, instead of the usual Euclidean metric. The differences between MCA and PCA are evident in two main properties:

1. Marginalization bias. MCA overweights the primary indicators with fewer categories. In the instance of a binomial indicator, the marginal categories will receive a higher weight because the covariance is the same for both categories.
2. Duality. MCA can be applied on the matrix $\mathbf{X}$ to the row profiles (observations) or the column profiles (categories). In terms of SES, the composite SES scores of the OHS Pilot data is the average of the standardized factorial weights of the K potential SES predictors. Equivalently, the weight of a given poverty category is the mean of the composite SES standardized scores for the corresponding SES category.[20]

Therefore, it seems that there is potential for MCA for the study of SES in the OHS Pilot data. This area will be researched further, within the analysis of spirometric measures and SES, but with other outcomes as well.

10 Summary

Gaining practical experience doing biostatistical work with the Ontario Health Study this summer has helped me develop my statistical analysis skills and sharpen my own independent research abilities. Concepts like data cleaning aren't discussed frequently in a lecture setting, but this practice is essential for any statistical analysis, and though it seemed overwhelming and tedious at first, over time I became more efficient and quick to pick up data quality issues. The most beneficial aspect of my experience this summer was the independence needed to conduct these analyses and structure my own analytic plan. While everyone at the OHS was quick to help me with questions, to construct a suitable analytic plan for the SES analysis with spirometry required me to draw on skills from my entire undergraduate and graduate academic career. In all, it became clear that in epidemiological studies, statistical analyses and model building is an "art". However, mathematical rigour and a solid statistical foundation are required for the results to truly make a scientific impact. I enjoyed and learned a great deal from working with professionals, researchers and clinicians with several different backgrounds, and it seems that being well-versed in many disciplines may be a strong asset in epidemiology, so that statistical needs are met with clinical desires.

11 Appendix I

Table 16: FEV1 in white males by education level

Level	Mean	SD
Elementary	3.16	0.59
High School	3.47	0.48
Technical Certificate	3.52	0.50
College	3.60	0.47
University Certificate	3.55	0.44
Bachelor's Degree	3.65	0.49
Graduate Degree	3.67	0.49

Table 17: FEV1 in white females by education level

Level	Mean	SD
Elementary	2.29	0.59
High School	2.54	0.48
Technical Certificate	2.50	0.50
College	2.67	0.47
University Certificate	2.58	0.44
Bachelor's Degree	2.73	0.49
Graduate Degree	2.76	0.49

Table 18: Descriptive statistics for FVC in white females by Education levels

Level	Mean	SD	
Elementary	4.08	0.83	63.00
High School	4.48	0.75	389.00
Technical Certificate	4.53	0.80	284.00
College	4.63	0.77	399.00
University Certificate	4.62	0.79	93.00
Bachelor's Degree	4.69	0.77	649.00
Graduate Degree	4.76	0.82	406.00

Table 19: FEV1/FVC % predicted in white males by Education levels

Level	Mean	SD
Elementary	75.18	1.65
High School	75.95	1.59
Technical Certificate	75.92	1.65
College	76.44	1.74
University Certificate	75.87	1.73
Bachelor's Degree	76.29	1.84
Graduate Degree	76.17	1.76

Table 20: Descriptive statistics for FVC in white females by Education Levels

Level	Mean	SD
Elementary	2.84	0.69
High School	3.18	0.58
Technical Certificate	3.18	0.63
College	3.34	0.60
University Certificate	3.30	0.59
Bachelor's Degree	3.46	0.64
Graduate Degree	3.48	0.63

Table 21: Descriptive statistics for FEV1/FVC % predicted in white females by Education Levels

Level	Mean	SD
None		
Elementary	77.12	1.65
High School	78.01	1.59
Technical Certificate	77.94	1.65
College	78.65	1.74
University Certificate	77.92	1.73
Bachelor's Degree	78.73	1.84
Graduate Degree	78.72	1.76
Don't Know		
Prefer not to answer		

Table 22: FEV1 in white males by income category

Level	Mean	SD
Low	3.35	0.67
Medium	3.56	0.62
High	3.77	0.62

Table 23: FEV1 in white females by income category

Level	Mean	SD
Low	2.45	0.48
Medium	2.66	0.48
High	2.84	0.46

Table 24: FVC in white males by income category

Level	Mean	SD
Low	4.35	0.82
Medium	4.59	0.77
High	4.84	0.77

Table 25: FVC in white females by income category

Level	Mean	SD
Low	3.09	0.57
Medium	3.34	0.61
High	3.59	0.61

Table 26: FEV1/FVC% predicted in white males by income category

Level	Mean	SD
Low	75.32	1.60
Medium	76.07	1.75
High	76.82	1.58

Table 27: FEV1/FVC % predicted in white females by income category

Level	Mean	SD
Low	77.52	1.54
Medium	78.53	1.80
High	79.13	1.54

Table 28: FEV1 in white males by working status

Level	Mean	SD
Full time	3.77	0.60
Part time	3.38	0.59
Unable to work	3.49	0.83
Looking after family	3.82	0.63
Retired	3.35	0.61
Unemployed	3.57	0.61
Unpaid work	3.14	0.49

Table 29: FEV1 in white females by working status

Level	Mean	SD
Full time	2.77	0.47
Part time	2.71	0.47
Unable to work	2.49	0.45
Looking after family	2.80	0.54
Retired	2.39	0.44
Unemployed	2.80	0.43
Unpaid work	2.57	0.48

Table 30: FVC in white males by working status

Level	Mean	SD
Full time	4.82	0.76
Part time	4.43	0.73
Unable to work	4.53	1.02
Looking after family	4.91	0.94
Retired	4.35	0.75
Unemployed	4.61	0.77
Unpaid work	4.10	0.50

Table 31: FVC in white females by working status

Level	Mean	SD
Full time	3.49	0.61
Part time	3.42	0.58
Unable to work	3.10	0.55
Looking after family	3.51	0.63
Retired	3.03	0.55
Unemployed	3.50	0.60
Unpaid work	3.30	0.72

Table 32: FEV1/FVC % predicted in white males by working status

Level	Mean	SD
Full time	77.12	1.61
Part time	75.43	1.35
Unable to work	76.68	1.48
Looking after family	77.31	1.57
Retired	74.78	0.86
Unemployed	76.87	1.22
Unpaid work	74.93	1.34

Table 33: FEV1/FVC % predicted in white females by working status

Level	Mean	SD
Full time	79.26	1.52
Part time	78.60	1.70
Unable to work	78.23	1.36
Looking after family	79.60	1.82
Retired	76.78	0.77
Unemployed	79.35	1.54
Unpaid work	77.64	1.38

References

[1] Williams, Tracey. 2011. Ontario Health Study: physicians to play integral role in study implementation and evolution. Ontario Medical Review feature, May 2011. Retrieved from the Ontario Health Study network July 13, 2011.

[2] Framingham Heart Study. 2011. *Research Milestones.* From the Framingham Heart Study website. Accessed July 21, 2011 at http://framinghamheartstudy.org/about/milestones.html.

[3] Cancer Care Ontario. October 2008. Proposal to Establish the Ontario Population Cohort Study. Ontario Population Cohort Study Proposal for Ethics Review. (Ontario Health Study).

[4] Colley, Rachel C., Didier Garriguet, Ian Janssen, Cora L. Craig, Janine Clarke, & Mark S. Tremblay. Canadian Health Measures Survey: Physical activity of youth and adults. *Statistics Canada*, Catalogue no. 82-003-X. Health Reports, Volume 22, no.1.

[5] Bauman, Adrain, Bull, Fion, Chey, Tien, et al. 2009. The Internation Prevalence Study on Physical Activity: results from 20 countries. *Internation Journal of Behavioural Nutrition and Physical Activity*, 6 (21): 1381-95.

[6] World Health Organization. 2011. *Global Physical Activity Questionnaire and Analysis Guide.* Accessed online on July 5, 2011 at http://www.who.int/chp/steps/GPAQ/en/index.html

[7] R.A. and R.M. Hauser. Socioeconomic status (SES) and health at midlife; a comparison of educational attainment with occupation-based indicators. Ann Epidemiol. 2001; 11: 75-84

[8] Shavers, Vicki. 2007. Measurement of Socioeconomic Status in Health Disparities Research. *Journal of the National Medical Association*, 99(9): 1013-1023

[9] Pampalon, Robert, Denis Hamel, and Philippe Gamache. 2009. A comparison of individual and area-based socio-economic data for monitoring social inequalities in health - Methodological Insights. *Statistics Canada*, Catalogue no. 82-003-XPE. Health Reports, Volume 20, no.3.

[10] Canadian Institute for Health Information, Reducing Gaps in Health: A Focus on Socio-Economic Status in Urban Canada (Ottawa, Ont.: CIHI, 2008).

[11] Canadian Institute for Health Information, How Healthy Are Rural Canadians? An Assessment of Their Health Status and Health Determinants (Ottawa, Ont.: CIHI, 2006).

[12] Pomerleau, Joceline, et al. Health behaviours and socio-economic status in Ontario, Canada. *European Journal of Epidemiology.* 1997; 13: 613-622.

[13] Public Health Agency of Canada. 2008. The Chief Public Health OfficerâĂŹs Report on the State of Public Health in Canada. *Social and Economic Factors that Influence Our Health and Contribute to Health Inequalities* [Chapter 4].

[14] Ranjit, Nalini, Diez-Roux, Ana V., Shea, Steven, et al. 2007. Socioeconomic Position, Race/Ethnicity, and Inflammation in the Multi-Ethnic Study of Atherosclerosis. Circulation, 116: 2383-2390.

[15] Singh-Manoux, Archana, et al. 2006. Subjective social status: its determinants and its association with measures of ill-health in the Whitehall II study. *Social Science & Medicine*; 56: 1321-1333.

[16] Hankinson, John L., Kawut, Steven M., Shahar, Eyal., et al. 2010. Performance of American Thoracic Society-Recommended Spirometry Reference Values in a Multiethnic Sample of Adults.: The Multi-Ethnic Study of Atherosclerosis (MESA) Lung Study. Chest, 137(1): 138-145.

[17] Miller, M.R., Hankinson, J., Brusasco, V., et al. 2005. Standardisation of spirometry. *Eur Respir J*, 26(2): 319-338.

[18] Statistics Canada. 2007. Ontario (Code35) (table). 2006 Community Profiles. 2006 Census. *Statistics Canada Catalogue* no. 92-591-XWE. Ottawa. Released March 13, 2007. Accessed August 5, 2011 at
http://www12.statcan.ca/census-recensement/2006/dp-pd/prof/92-591/index.cfm?Lang=E

[19] Panagiotakos, Demosthenes B. and Christos Pitsavos. 2004. Interpretation of Epidemiological Data Using Multiple Correspondence Analysis and Log-linear Models. Journal of Data Science 2: 75-86.

[20] Nenadic, Oleg, and Michael Greenacre. 2007. *Computation of Multiple Correspondence Analysis, with code in R* [electronic]. Accessed online on August 2, 2011 from http://www.econ.upf.edu/docs/papers/downloads/887.pdf

[21] Gelman, Andrew,& Jennifer Hill. 2009. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press: New York.