A Preliminary Statistical Analysis on Risk Factors for Dementia and CIND from the Canadian Study of Health and Aging

By

Nan Zhao

A Research Paper presented to the
Department of Actuarial Science and Statistics
University of Waterloo
In partial fulfillment of the requirements
For the degree of Master of Mathematics in
Statistics and Computing

Waterloo, Ontario, Canada, Jan. 2008

Acknowledgement

To my supervisor Professor Wayne Oldford for his infinite patience, wise advice and everlasting strong support to me. To Professor Shai Ben-David for his valuable comments and suggestions.

I want to express my deepest gratitude to my family for their unstinting support, encouragement and love.

To all my friends in China and Canada and those who have shared with me every bittersweet moment in my life.

Abstract

Dementia and cognitive impairment no dementia (CIND) are considered as major health problems in aging population but have rarely been studied together. The Canadian Study of Health and Aging (CSHA) is a nationwide study of Canadian seniors by the University of Ottawa and Health Canada designed to study the prevalence, incidence and risk factors for dementia. Analysis has been done only for dementia on data from the first wave (1991-1996) of this study. In this paper, we present our analysis for both dementia and cognitive impairment no dementia on a subset of the data, namely those who were cognitively normal in 1991, and alive in 1996 and also those who alive in 2001. The adjusted odds ratios for the various risk factors calculated using logistic regression models and comparisons made between results from the CSHA-1 ~ CSHA-2 (1991-1996) and CSHA-2 ~ CSHA-3 (1996-2001) are included.

Table of Contents

Chapter 1

Introduction

1.1 Motivation and Organization of paper

Dementia and Cognitive Impairment No Dementia (CIND) are considered as major health problems in aging population but have rarely been studied together. In collaboration with Dr. Vladimir Hachinski, of the Lawson Health Research Institute, and Dr. Truls Ostbye, of Duke University, we luckily have access to the data from the Canadian Study on Health and Aging (CSHA). Our goals here are to understand the structure of the data at hand, and to generate hypotheses about some of the possible risk factors for dementia and CIND in an elderly population and try to examine if stroke poses a risk for dementia and CIND. Hypotheses generated in this project can be tested and explored on other data set, such as the data from the Framingham Heart Study, the clinical data supplied by Dr. David Spence's project and the data from Dr. Cechetto's animal models in order to produce significantly useful tools for stroke prevention. In terms of the methods, traditional statistical tools, namely logistic regression and odds ratios calculated from logistic regression models are used in detecting the possible risk factors and their effects to dementia and CIND.

In terms of the organization of this paper, the first chapter provides a description of the Canadian Study on Health and Aging and its data sets and the work we have done on data pre-processing is also included in this chapter. In addition, we briefly introduce logistic regression model, odds ratios and a few other relevant statistical concepts that we use in the analysis in this chapter as well. The statistical analysis of 13 factors on subsets of the data is included in the second chapter, which includes the models fitted on the data and results of test of significance and odds ratios. The third chapter provides the investigation on correlation between stroke and dementia, and also between stroke and CIND, i.e. whether stroke poses a risk for dementia, using the same statistical methods as above. Chapter 4 gives the conclusions and future extensions.

1.2 Introduction to the CSHA

1.2.1 Structure of the CSHA

Nowadays, people are living longer, but unfortunately diseases of aging are also becoming more common. Among such diseases, many different forms of dementia pose major threats. Dementia involves a progressive decline in person’s memory followed by other aspects of cognitive functioning due to brain dysfunction. These patients will eventually become incapable of caring for themselves and even unable to recognize close family members at the end.

The Canadian Study of Health and Aging (CSHA, www.csha.ca) is conducted as a multidisciplinary, nationwide, multi-center, population-based, longitudinal study of Canadian seniors which coordinated jointly by the University of Ottawa and Health Canada. In the CSHA official website shown above, its objectives were divided into four categories as the following:

1. “Core objectives addressed the prevalence, incidence and risk factors for dementia, and the impact of dementia on family caregivers”.
2. “Secondary objectives covered other health topics (such as disability, frailty or healthy aging) that could readily be addressed in the context of the study”.
3. “In addition, participating investigators were encouraged to add supplementary “add-on” studies of personal interest to them, and for which they could obtain separate funding”.
4. “Finally, it was anticipated that the CSHA results would generate “spin-off” studies that could be undertaken by the same team members under separate funding arrangements”.

The Canadian Study of Health and Aging has followed a nationally representative sample of 10,263 elderly Canadians over 10 years and has collected a wide range of information on their changing health status over that time. Participants were aged 65 and above, and include both community living and institutionalized elderly from urban and rural areas within Canada. The CSHA also involves a team of over 60 investigators (clinicians, epidemiologists, social scientists, psychologists and others) collaborated in 18 study centers across Canada. Data were collected at 5-yearly intervals: CSHA-1 in 1991, CSHA-2 in 1996, and CSHA-3 in 2001. Limited data were also collected in 1993.

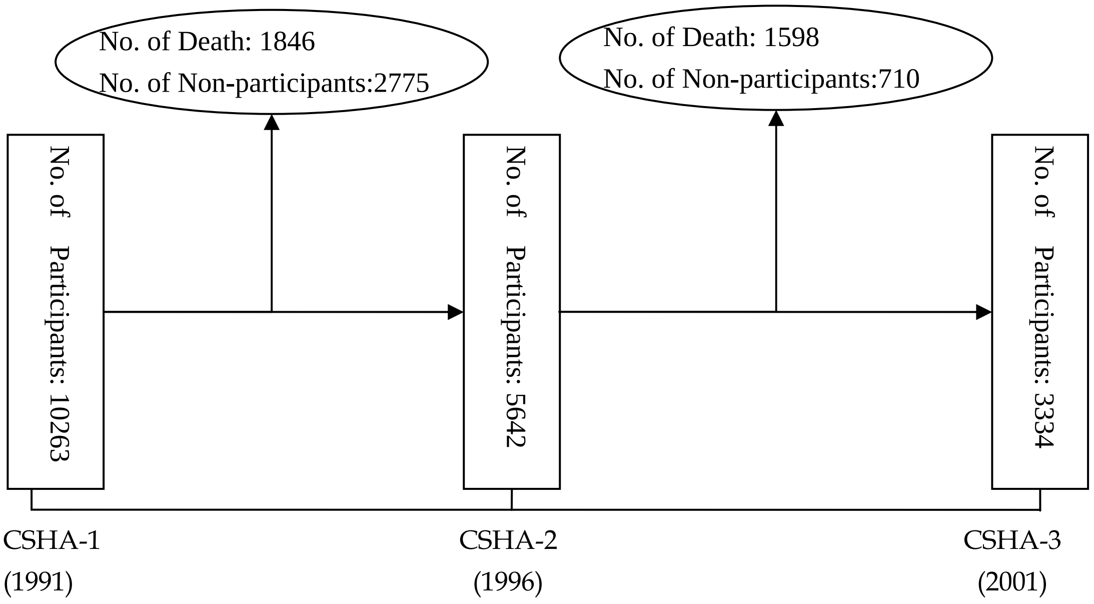
No. of Death: 1846
No. of Non-participants:2775
No. of Death: 1598
No. of Non-participants:710
No. of Participants: 10263
No. of Participants: 5642
No. of Participants: 3334
CSHA-1
(1991)
CSHA-2
(1996)
CSHA-3
(2001)

Figure 1: Flow of Participants over 3 waves of CSHA

In 1991 - 1992, representative samples of men and women aged 65 years or older were drawn from 36 urban and rural areas over all 10 Canadian provinces. Within each of the above sampling areas, separate samples were drawn for community and for institutional people. Of the 10,263 participants, 9008 lived in the community while 1255 resided in institutions. However, not everybody selected in the sample participated all the way through the CSHA. Residents of the province of Newfoundland (395) were excluded from the prospective analysis because a legal interpretation of the province's advance directive legislation prohibited the use of proxy consent for persons unable to give full, informed consent to participate in research studies in 1996. In addition, 2380 participants were considered as being in the situation of either lost contact, not agree to participate or not selected for further study. In addition of 1846 death, there were 4680 participants who had to be omitted after the CSHA-1. In 1996 – 1997, the study continued on all subjects (5642) who could be contacted and who agreed to participate as the second wave of the study. The third wave of the study took place on all those who did not have dementia at CSHA-2 in 2001. 2308 subjects had left the study by that time.

1.2.2 Study Instruments and data files

Each wave of the study can be described as a diagnostic process, which contained the cognitive screening test and clinical assessment, combined with caregiver & risk factor questionnaires and other questionnaires regarding to additional information such as decedent information or institutionalization information which may vary from time to time. The general follow is shown in the following chart.

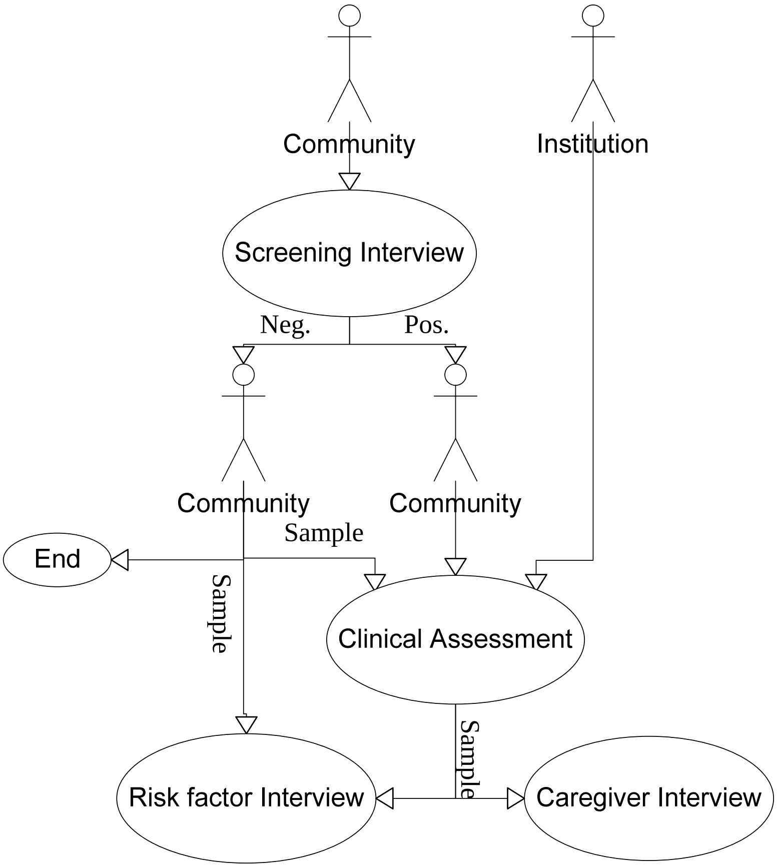

Figure 2: General follow of participants in CSHA

First of all, those in community were invited to the cognitive screening test for psychometrically testing of cognitive impairment. Participants were interviewed for 25 – 30 minutes about their health, presence of specific disorders and limitations in performing basic and instrumental activities of daily living based on an Activities of Daily Living (ADL) scale. All participants were tested for dementia by using the Modified Mini-Mental State (3MS) Examination. People who screened positive (i.e. 3MS < 78), a random sample of those who screened negative (i.e. 3MS $\geq$ 78) and all institutional subjects were sent for an extensive clinical evaluation. Within the clinical evaluation, there were 3 main components. A nurse first re-examed the 3MS score and collected information on the participant's medical and family history. Then every participant needed to take a standardized physical and neurological exam conducted by a formal physician. After that, all subjects with a 3MS score of 50 or above continued to neuropsychological testing. Finally, preliminary diagnoses were made by formal physician based on the results from the standardized physical and neurological exam, and followed by case conference where a consensus diagnosis was produced according to Diagnostic and Statistical Manual of Mental Disorders. The clinical exam permitted confirmation of the classification of cognitive impairment, the diagnosis of dementia, and the differential diagnosis of the type of dementia. Further more, participants who found to be cognitively normal based on either the screening test or clinical exam results were asked to complete the risk factor questionnaire by themselves. Proxies for around a half of these participants were asked to re-answered the risk factor questions in order to either ensure the accuracy of provided information about the subjects or use proxy as a representative for the participant.

In addition to above processes, laboratory tests were undertaken in the follow-up study. For example, CT scans were brought into the study in

CSHA-2, and blood samples were collected in CSHA-3 for genetic and other analysis. Given the interrelationship between dementia/cognition and vascular disease, detailed information on cardiovascular risk factors and vascular outcomes were also collected. Finally, abstracted information from death certificates and provincial health care utilization records were collected for the study period.

Regarding to the data set holding most of the information we described above, there are 31 data files in total we have on hand. The following table describes the correspondence between instruments performed and data files.

Table 1: Correspondence between processes and data files

		CSHA-1	CSHA-2	CSHA-3
Screening Test		SCREEN1	SCREEN2	SCREEN3
Clinical Assessment	Nurse exam	NURSE1	NURSE2	
	Physician test	DOCTOR1	DOCTOR2	DOCTOR3
	Neuropsychological exam	NEURO1	NEURO2	NEURO3
Caregiver/Informant Questionnaire		CARE1	CARE2	CARE3
Risk Factor Questionnaire.		PROXY1 & SELF1		
Decedent			DECED2NEW	DECED3
Institutionalization Questionnaire			INSTIT2	INSTIT3
Link file		LINK3		
Informant Component				INFORM3
Non-participation codes		STATUS1		
Maintaining contact study		MCS_DIST		

Additional datasets of derived variables:

Imputed-Date of Institutionalization:	INSTDATA
Imputed_Age at Death:	DEATHAGE
Cognitive status at death (CSHA-2):	DCAM2FIX
Cognitive status at death (CSHA-3):	DCAM3FIX
Correction to the phase 2 WAIS-R Comprehension Test:	WAISCOM2
Case/Control, Incidence/Prevalence as used for paper:	RF_CC_IP
Estimated Date of Onset:	ONSETDAT

1.3 Pre-processing on data

As described in previous session, numerous information have been collected on the participants regarding to their health, history, daily life, hobbies, jobs etc. all most everything related to or may have effect on their health status over 10 years were recorded. After going through every data file we have on hand, in excess of 6000 demographic, psychological, social, clinical and risk factor variables are available in the data set. It is apparent that all 6000 variable can't be analyzed at once. They need to be logically grouped into different categories such that each category will contain factors that may have similar effect on the prediction of Dementia and CIND. Using such grouping strategy, we are hoping to eliminate most of the factor variables from the true potential risk factors within the same group.

Base on nature of the variables we have, there are five categories that we think is appropriate to distribute variable factors into:

- Sociodemographic factors:

  Participant's age, sex, institutional/commoditized status, years of education, rural/urban status, region, race and language are included.

- General medical factors:

  This category includes subject's medical status and history, such as cholesterol, blood pressure, ApoE4 (i.e. ApoE4 is indicating the presence of apolipoprotein E4 which is proposed to be a genetic risk factor for dementia and other diseases), history of diabetes, history of cerebral disease, history of cerebro-vascular events such as strokes, history of cardiovascular events, current medication etc.

- Physical factors:

  In this category, we consider variables regarding to subject’s behavior and ability of daily living. For example, smoking, alcohol consumption, social support, mood, family etc.

- Psychological factors:

  Psychological factors are the variables regarding short and long term memory, ability of abstract thinking and executive functioning and so on. CT scans and 3MS scores are also considered in this category.

- Environmental factor:

  Environmental factors are variables describing the outside factor that the subject is/was interacting with. Things like jobs, hobbies, living standard etc are considered.

There may be overlapping of variables across different groups which is not much of a problem here, because grouping variables is not the way to eliminate factors. For the purpose of our analysis, we choose to analyze the socio-demographic factors and other factors which are proposed by other study as possible risk factors such as arthritis, wine consumption, coffee consumption, regular physical activity and ApoE4.

Since we are at the preliminary stage of the analysis, we want to start with a sample set that is simple but would lead the investigation to a meaningful stage. In this case, we choose to start with the 921 participants who were clearly diagnosed as cognitively normal in CSHA-1 (1991). 300, 152 and 86 of them were cognitively normal, CIND and demented respectively in CSHA-2 (1996). By the time of the third wave, 107, 85 and 52 of them were cognitively normal, CIND and demented respectively.

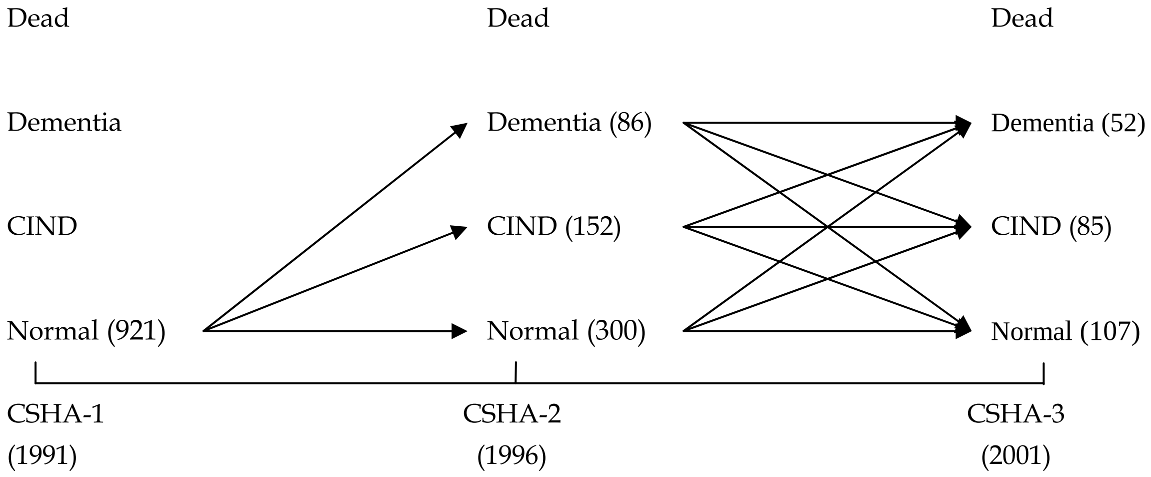

Figure 3: Number of test samples at each status.

The reason to choose subjects that were clearly diagnosed as cognitively normal instead of subject whose screening result was negative is that we try to avoid any possible false negative cases. However, there is one limitation about this strategy, which is the size of the data set will be relatively small.

1.4 Methods

1.4.1 Logistic regression model

Logistic regression, which is part of a category of statistical models called generalized linear models, is chosen since it aims at predicting a discrete class response variable based on a set of explanatory variables or features. Moreover, logistic regression allows variables to be continuous, discrete, dichotomous, or a mix of any of these, which is a particular important characteristic of the data set.

In logistic regression model, we have a response variable y that indexes an object's class. In our case here, we only have 2 classes, $y=1$ represents that the subject has dementia or CIND and $y=0$ represents that the subject is cognitively normal. We also have d explanatory variables,

$$X=(x_1,\ldots,x_d)^T,$$

describing each object or case, also in our case here, we have $\max(d)=13$.

In logistic regression, we model the probabilities of belonging to the various classes given explanatory variable information. A binary-valued random Y is representing the distribution of the possible values (0 and 1) of y. Y is Bernoulli random variable, which is jest a special case of a binomial random variable with one trial. Thus, we will be modeling the conditional probability

$$p(X) \equiv p(Y=1|X).$$

Specifically, a logistic regression model has the form

$$\log\left(\frac{p(X)}{1-p(X)}\right)=\eta(X).$$

Often $\eta(X)$ is a linear predictor. For example, $\eta(X)=\beta_0+\beta_1 x_1+\ldots+\beta_d x_d$.

Now we can rewrite the logistic model as

$$p(X)=p(Y=1|X)=\frac{\exp[\eta(X)]}{1+\exp[\eta(X)]}.$$

In more general form, we can write the logistic model as

$$p(Y=y|X)=\left(\frac{\exp[\eta(X)]}{1+\exp[\eta(X)]}\right)^y\left(\frac{1}{1+\exp[\eta(X)]}\right)^{1-y} \qquad (y=0,1)$$

For each case i in the data set, we know their class $Y_i=y_i$ and their explanatory variables X_i. As we described above, we are modeling

$$p(X_i)=p(Y_i=y_i|X_i)=\frac{\exp[\eta(X_i)]}{1+\exp[\eta(X_i)]}.$$

The unknown parameters in $\eta(X)$, e.g., $\beta_0,\beta_1,\ldots\beta_d$ are estimated via maximum likelihood with the assumption of Y_i are independent. The maximization may be carried out by least square algorithm which is not included here.

1.4.2 Residual deviance

The residual deviance is a measure of fit of a logistic regression model or a generalized linear model that is similar to the residual sum of squares in linear models. The deviance for the fitted model is defined to be

$$D(y;\hat{\beta}) = 2[l_{\max} - l(y;\hat{\beta})],$$

where $l(y;\hat{\beta})$ denote the log likelihood for a fitted model with maximum likelihood estimates $\hat{\beta}$, and $l_{\max}$ denote the maximum possible log likelihood. McCullagh and Nelder suggested that the deviance is most useful for comparing nested models. Suppose model A is nested within model B and let m denote the change in number of fitted parameters between two models, the change in deviance

$$D(y;\hat{\beta}_A) - D(y;\hat{\beta}_B)$$

is a likelihood ratio statistics for testing whether the m extra parameters are zero. In another word, we can use this to test the significance of the m extra terms to the model.

1.4.3 Odds ratio from logistic regression model

The odds of the outcome being present among individuals with *x* present (i.e. *x*=1) is defined as

$$\frac{p(Y \mid x=1)}{1-P(Y \mid x=1)}.$$

Then the odds ratio is defined as

$$Odds\ ratio = \frac{p(Y \mid x=1)/1-P(Y \mid x=1)}{p(Y \mid x=0)/1-P(Y \mid x=0)} = \frac{p(Y \mid x=1)[1-P(Y \mid x=0)]}{p(Y \mid x=0)[1-P(Y \mid x=1)]}$$

The odds ratio is often used to assess the risk of a particular outcome if a certain factor is present, and it can be derived from a logistic regression model. If we recall from the previous page that in logistic regression model, the dependent variable is a *logit*, which is the log of the odds. Odds ratio can be derived from this relationship as the following:

$$
\begin{aligned}
&\log(Odds(x)) = \log\left(\frac{p(Y \mid x=1)}{1-p(Y \mid x=1)}\right) = \eta(x) \\
&\rightarrow \log(Odds(x)) = \beta_0 + \beta_x \\
&\rightarrow Odds(x) = \exp(\beta_0 + \beta_x) \\
&\rightarrow Odds\ ratio(x) = \frac{\exp(\beta_0 + \beta_x)}{\exp(\beta_0)} \\
&\rightarrow Odds\ ratio(x) = \exp(\beta_x)
\end{aligned}
$$

This is correct when the independent variable has been coded as 0 or 1 to indicate its presence. A confidence interval estimate for the odds ratio is obtained by calculating the endpoints of a confidence interval for β_x and then exponentiating the two values. In general, the endpoints are given by

$$\exp(\hat{\beta}_x \pm z_{1-\alpha/2} \times \hat{se}(\hat{\beta}_x)) .$$

When x is a continuous variable, the odds ratio can be estimated as

$$Odds\ ratio(x) = \exp(a\beta_x)$$

where a is the difference between 25% and 75% quantiles of the continuous variable. And the 95% confidence interval of the odds ratio is given by

$$\exp(a\hat{\beta}_x \pm z_{1-\alpha/2} \times a \times \hat{se}(\hat{\beta}_x)) .$$

Chapter 2

Results of analysis on risk factors for Dementia and CIND

2.1 Results for Dementia

2.1.1 From CSHA-1 to CSHA-2

We first look at those people who were cognitively normal in CSHA-1(1991) and alive but also cognitively normal or demented in CSHA-2(1996). As we described in Chapter 1, we are analyzing 13 factors which are age, sex, institutional/commoditized status, years of education, rural/urban status, region, race, language, arthritis, wine consumption, coffee consumption, regular physical activity and ApoE4. Due to missing values in the dataset, we have 263 subjects (Dementia: 42, Normal: 221) left with complete information on 13 factors.

All 13 factors are firstly sent to the stepAIC process to choose the best logistic regression model according to its Akaike Information Criterion (AIC) value. Test of significance and odds ratios will be calculated for those variables contained in this model. For this dataset, the chosen variables are age, years of education, coffee consumption and ApoE4. Most of the socio-demographic factors are eliminated by stepAIC process. In order to verify the influence of sex, arthritis, wine consumption and regular physical activity on risk of dementia since they are proposed by other researchers, they will also be included in further analysis.

Table 2 shows the significance of all 8 variables described above. The smaller the value is, the more significant the variable is. We can see that age, years of education, coffee consumption and ApoE4 are highly significant whereas other variables don't demonstrate their significance in the model.

Table 2: Significance of selected risk factors, CSHA-2

Variables	Test of Significance*
Age	5.673871e-07
Sex	0.6629166
Years of Education	0.01611759
Arthritis	0.887537
Wine	0.2435375
Coffee	0.044697
Regular physical activity	0.4347878
ApoE4 †	0.009902521

* Logistic regression model used: factor(SUMMNEW) ~ CSHA1AGE + EDUC + factor(SEX)+ factor(ARTH) + factor(WINE) + factor(COFF) + factor(EXER) + factor(E4). The rest of the models are nested within above model.

† ApoE4, apolipoprotein allele

Table 3 shows the odds ratios of each of the 8 variables. We can see that coffee consumption shows statistically significant association with a reduced risk of dementia whereas ApoE4 shows statistically significant association with an increase risk of dementia. Moreover, regular physical activity, wine consumption and arthritis present the benefit of reducing risk of dementia and being female increases the chance of having dementia. However the effects from these four factors are not found to be statistically significant here. Age and years of education continue to show statistical significant influence on risk of dementia here as they do in Table 2.

Table 3: Risk of Dementia associated with selected variables, CSHA-2

Variables	Odds Ratio§	95% confidence interval	
Age (by year)	3.44	1.98	5.96
Sex (Male:Female)	0.69	0.36,	1.31
Years of Education (by year)	0.54	0.34	0.87
Arthritis (No:Yes)	1.01	0.54,	1.82
Wine (Yes:No)	0.31	0.07	1.43
Coffee (No:Yes)	2.22	1.08	4.56
Regular physical activity (No:Yes)	1.36	0.67	2.75
ApoE4 (Yes: No)	3.82	1.64	8.87

§ Adjusted for age, sex, and education

Table 4 shows us a clear view of the effects from each subgroup of age and years of education. Both groups of age show the trend that higher age group has a high probability of being demented. Both groups of years of education show that lower education increase the chance of being demented.

Table 4: Risk of Dementia associated with age group and years of education, CSHA-2

Variables	Odds Ratio§	95% confidence interval	
Age (≥85:75-84) §	2.84	1.39	5.81
Age (65-74:75-84) §	0.20	0.07	0.53
Years of education (≥13 : 9-12)*	0.57	0.23	1.37
Years of education (0-8 : 9-12)*	1.98	1.02	3.84

§ Odds ratios were adjusted for sex and years of education.

* Odds ratios were adjusted for sex and age.

Some of the hypotheses generated from above analysis are matching with the results from other studies. In the prospective analysis of CSHA data done by Joan Lindsay, increasing age, low educational level and ApoE4 are also found to be significantly associated with increased risks of Alzheimer's disease (one particular type of dementia) and coffee consumption was significantly associated with reduced risk. Moreover, they identified that the benefit of regular physical activity and wine consumption were statistically significant which our analysis did not find. One major difference between two analyses is the starting point. We start with people who were clearly diagnosed as cognitively normal in CSHA-1 whereas they also include people who were screened negative but didn't have the clinical assessment. Within those who were screened negative and did get the clinical assessment, we find that around 1/5 of them are actually demented or CIND. In this case, including those who were only screened negative would give a larger data set but also would have the risk of false negative.

2.1.2 From CSHA-1 to CSHA-3

After the analysis on the period from CSHA-1 to CSHA-2, we look at those people who were cognitively normal in CSHA-1(1991) and alive as well as cognitively normal or demented in CSHA-3(2001). The same 13 factors were considered as well. Due to the missing values in the dataset, we have 123 subjects (Dementia: 36, Normal: 87) left with information of all 13 factors present.

Same as before, the stepAIC process is used to choose the best logistic regression model. Only age and years of education are chosen at this time. Most of the factors are eliminated. In order to verify the influence of sex, arthritis, coffee consumption, ApoE4, wine consumption and regular physical activity on risk of dementia and make comparison with their effect from CSHA-2, they will also be included in further analysis.

Table 5 shows the significance of all 8 variables described above. We can see that age and years of education are still highly significant whereas coffee consumption and ApoE4 have lost its significance and other variables remain insignificance to the data.

Table 5: Significance of selected risk factors, CSHA-3

Variables	Test of Significance *
Age (by year)	0.02946907
Sex	0.5776802
Years of Education (by year)	0.0001030493
Arthritis	0.2920819
Wine	0.2523947
Coffee	0.9203443
Regular physical activity	0.5716076
ApoE4†	0.9824153

*Logistic regression model used: factor(SUMMNEW) ~ CSHA1AGE + factor(ARTH) + factor(WINE) + factor(COFF) + factor(EXER) + EDUC + factor(SEX) + factor(E4). The rest of the models are nested within above model.

† ApoE4, apolipoprotein allele

Table 6 contains the odds ratios and its 95% confident interval of each of the 8 variables. We can see that age and years of education continually shows statistically significant association with risk of dementia. ApoE4 shows an association with an increase risk of dementia but the effect is not found to be significant. Moreover, regular physical activity, coffee consumption, wine consumption and arthritis have the same effect as before on the risk of dementia and being female increases the chance of having dementia. However the effects from these five factors are not shown statistically significant here, even though coffee consumption was significantly associated with a decrease risk of dementia previously. Age and years of education continue to show statistically significant and strong influence on risk of dementia.

Table 6: Risk of Dementia associated with selected variables, CSHA-3

Variables	Odds Ratio §	95% confidence interval	
Age (by year)	2.66	1.32	5.35
Sex (Male:Female)	0.69	0.28	2.03
Years of Education (by year)	0.38	0.21	0.68
Arthritis (No:Yes)	1.05	0.48,	1.79
Wine (Yes:No)	0.86	0.35	2.14
Coffee (No:Yes)	1.94	0.68	5.58
Regular physical activity (No:Yes)	1.51	0.61	3.73
ApoE4 (Yes:No)	1.42	0.48	4.17

§ Adjusted for age, sex, and education

In Table 7, we can observe that both groups of years of education show that lower education increases the chance of being demented and both groups of age show the trend that higher age group has a high probability of being demented as before. However, the group of age more than 85 is not significant anymore. The reason is that there are only 4 elderly subjects made to CSHA-3 with age more than 95 in 2001, so that the percentage of people being in the ≥85 age group is very small. In fact, among 4 of these subjects, 2 of them are demented and 2 of them are cognitively normal. Therefore, the ability of identifying the influence of such age group on dementia is limited.

Table 7: Risk of Dementia associated with age group and years of education, CSHA-3

Variables	Odds Ratio§	95% confidence interval	
Age (≥85:75-84) §	2.55	0.30	21.48
Age (65-74:75-84) §	0.42	0.18	0.98
Years of education (≥13 : 0-8)*	0.18	0.06	0.62
Years of education (9-12 : 0-8)*	0.32	0.13	0.78

§ Odds ratios were adjusted for sex and years of education.

* Odds ratios were adjusted for sex and age.

Among 13 selected variables, only age and years of education have shown continued strong and significant effect on risk of dementia through both of the two phases of the study. Coffee consumption and ApoE4 showed their significant relationship with the risk of dementia for the period of CSHA-2, but lost their significance by the time of CSHA-3. The rest of the variable show different impact on risk of dementia, but the relationship is not found to be statistically significant.

2.2 Results for Cognitive impairment no dementia (CIND)

2.2.1 From CSHA-1 to CSHA-2

For the analysis of CIND, we look at those people who were cognitively normal in CSHA-1(1991) and alive and also CIND in CSHA-2(1996). In this case, we model the probability of having CIND instead of being demented now. The same analysis strategy and tools are used here as well. The same set of 13 factors is selected and to be analyzed. Due to the missing values in the dataset, we have 306 subjects (CIND: 85, Normal: 221) left with information of all 13 factors present.

The stepAIC process is used again. For this dataset, the chosen variables are age, years of education and sex. Most of the socio-demographic factors are eliminated. In order to verify the influence of arthritis, wine consumption, coffee consumption, ApoE4 and regular physical activity on risk of CIND and to compare with the results for dementia, they will also be included in further analysis.

Table 8 shows the significance of all 8 variables. We observe that age, years of education are the only two factors which are highly significant whereas other variables don’t demonstrate their significance to the data.

Table 8: Significance of selected risk factors, CSHA-2

Variables	Test of Significance*
Age (by year)	0.04550026
Sex	0.1435019
Years of Education (by year)	0.001065580
Arthritis	0.9244194
Wine	0.9203443
Coffee	0.943628
Regular physical activity	0.5071225
ApoE4 †	0.4548712

* Logistic regression model used: factor(SUMMNEW) ~ CSHA1AGE + EDUC + factor(SEX)+ factor(ARTH) + factor(WINE) + factor(COFF) + factor(EXER) + factor(E4). The rest of the models are nested within above model.

† ApoE4, apolipoprotein allele

Table 9 contains the odds ratio and the 95% confidant interval of the odds ratio for each variable. The same situation as in Table 8 appears again here. Age and years of education continue to show statistical significant influence on risk of CIND. Coffee consumption, regular physical activity, wine consumption and arthritis show a statistically insignificant association with a reduced risk of CIND. ApoE4 and being female show an association with an increase risk of CIND, but not statistically significant.

Table 9: Risk of CIND associated with selected variables, CSHA-2

Variables	Odds Ratio §	95% confidence interval	
Age (by year)	1.41	1.02	1.96
Sex (Male:Female)	0.68	0.42	1.10
Years of Education (by year)	0.60	0.43	0.82
Arthritis (No:Yes)	1.02	0.64	1.63
Wine (Yes:No)	0.88	0.44	1.76
Coffee (No:Yes)	1.02	0.59	1.77
Regular physical activity (No:Yes)	1.30	0.82	2.06
ApoE4 (Yes:No)	1.70	0.87	3.33

§ Adjusted for age, sex, and education

In Table 10, we can have a clear view of the effects on risk of CIND from each subgroup of age and years of education. Both groups of age indicate a trend that higher age group has a higher risk of CIND. Both groups of years of education show that lower education increase the chance of being cognitive impairment but not demented. However, only the trend in years of education seems to be statistically significant.

Table 10: Risk of CIND associated with age group and years of education, CSHA-2

Variables	Odds Ratio§	95% confidence interval	
Age (≥85:75-84) §	1.26	0.54	2.94
Age (65-74:75-84) §	0.75	0.42	1.33
Years of education (≥13 : 0-8)*	0.49	0.25	0.97
Years of education (9-12 : 0-8)*	0.35	0.19	0.65

§ Odds ratios were adjusted for sex and years of education.

* Odds ratios were adjusted for sex and age.

2.2.2 From CSHA-1 to CSHA-3

After analysis on the first two waves of CSHA, we look at those people survived through the third wave of the study, namely who were cognitively normal in CSHA-1(1991) and alive and also cognitively normal or CIND in CSHA-3(2001). The same 13 factors were considered as well. Due to the missing values in the dataset, we have 160 subjects (CIND: 73, Normal: 87) left with complete information.

The stepAIC process chooses the model using age, years of education and sex as the best logistic regression model this time, which is the same model as before. For the same reason as before, sex, arthritis, coffee consumption, ApoE4, wine consumption and regular physical activity are included in further analysis.

Table 11 shows the significance of all 8 variables. We can say that age is highly significant and years of education is on the edge of being significant whereas other variables are not statistically significant in the model.

Table 11: Significance of selected risk factors, CSHA-3

Variables	Test of Significance *
Age (by year)	0.02505617
Sex	0.05743312
Year of Education (by year)	0.05004352
Arthritis	0.9203443
Wine	0.3427817
Coffee	0.3928832
Regular physical activity	0.1522062
ApoE4†	0.9775893

*Logistic regression model used: factor(SUMMNEW) ~ CSHA1AGE + factor(ARTH) + factor(WINE) + factor(COFF) + factor(EXER) + EDUC + factor(SEX) + factor(E4). The rest of the models are nested within above model.

† ApoE4, apolipoprotein allele

Table 12 shows the odds ratios and its 95% confident interval of each of the 8 variables. Age and years of education continually shows statistically significant association with risk of CIND which agrees with the result in Table 11. ApoE4 is not found to be significant for an association with an increase risk of CIND at this stage. Moreover, regular physical activity, coffee consumption, wine consumption and arthritis have the same effect as previous on the risk of CIND and being male increases the chance of having CIND now. However the effects from these five factors are not found statistically significant here. Age and years of education continue to show statistical significant influence on risk of CIND.

Table 12: Risk of CIND associated with selected variables, CSHA-3

Variables	Odds Ratio §	95% confidence interval	
Age (by year)	1.89	1.09	3.30
Sex (Male:Female)	1.84	0.93	3.62
Years of Education (by year)	0.66	0.46	0.95
Arthritis (No:Yes)	1.07	0.55	2.07
Wine (Yes:No)	0.56	0.22	1.42
Coffee (No:Yes)	1.46	0.72	2.96
Regular physical activity (No:Yes)	1.70	0.84	3.47
ApoE4(Yes:No)	1.08	0.46	2.54

§ Adjusted for age, sex, and education

In Table 13, we can observe that both groups of years of education show that lower education increases the chance of getting CIND and both groups of age show the trend that higher age group has a higher risk of CIND. However, none of age group and years of education seems to be significant anymore.

Table 13: Risk of CIND associated with age group and years of education, CSHA-3

Variables	Odds Ratio§	95% confidence interval	
Age (≥85:75-84) §	2.07	0.35	12.36
Age (65-74:75-84) §	0.53	0.27	1.05
Years of education (≥13 : 9-12)*	0.60	0.26	1.39
Years of education (0-8 : 9-12)*	1.37	0.65	2.89

§ Odds ratios were adjusted for sex and years of education.

* Odds ratios were adjusted for sex and age.

In conclusion, younger age and more years of education seem to be protective to CIND always. Regular physical activity, coffee consumption, wine consumption and arthritis keep having the same effect on the risk of CIND regardless of the time. However, the risk of CIND shifts from female to male at the end of the second study. The ApoE4 shows negative impact on protecting CIND. However the findings on ApoE4, sex, regular physical activity, coffee consumption, wine consumption and arthritis do not pass the statistical significance test.

To compare the risk factors' influence on dementia and CIND, we find that age and years of education are common risk factors for both, and have the same effects to the risk of both disease. Coffee consumption has a significant association with reduced risk of dementia in the early phase of the study, but it doesn't appear to be significantly beneficial for CIND. Similarly, ApoE4 is significantly associated with an increase risk of dementia but it is not the case for CIND. The rest of the risk factors are all associated with dementia and CIND in the same way, and none of these relationships appear to be statistically significant.

Chapter 3

Result of analysis on Stroke as a risk factor for Dementia and CIND

3.1 Results for Dementia

3.1.1 From CSHA-1 to CSHA-2

In this analysis, since we are focusing on relationship between stroke and dementia or CIND, we only model the probability of dementia or CIND using age, sex, years of education and history of stroke as predictors in the logistic regression model.

We first look at those people who were cognitively normal in CSHA-1(1991) and alive and also cognitively normal or demented in CSHA-2(1996). Without any cases with incomplete information, we have 364 subjects left to work with.

Table 14 shows the result of test of significance of all of the 4 variables described above. Age and years of education are highly significant whereas sex and history of stroke do not demonstrate their significance to in the model.

Table 14: Significance of selected risk factors, CSHA-2

Variables	Test of Significance *
Age (by year)	2.241873e-12
Sex	0.3681203
Years of Education (by year)	0.001698318
Stroke	0.806496

* Logistic regression model used: factor(SUMMNEW) ~ CSHA1AGE + EDUC + factor(SEX)+ factor(HXSTROKE). The rest of the models are nested within above model.

Table 15 includes the odds ratios for each of the 4 variables. Age and years of education shows statistically significant association with the risk of dementia. History of stroke is associated with an increase risk of dementia and being female is also increasing the risk of dementia. But they are not statistically significant to the data.

Table 15: Risk of Dementia associated with selected variables, CSHA-2

Variables	Odds Ratio §	95% confidence interval	
Age (by year)	3.47	2.35	5.13
Sex (Male:Female)	0.67	0.37	1.21
Years of Education (by year)	0.56	0.39	0.82
Stroke (Yes:No)	1.08	0.22	5.24

§ Adjusted for age, sex, and education

3.1.2 From CSHA-1 to CSHA-3

After analysis on the first two waves of the study, we look at those people who were cognitively normal in CSHA-1(1991) and alive and also cognitively normal or demented in CSHA-3(2001). Due to the missing values in the dataset, we have 145 subjects left with information of all 4 factors present.

Table 16 and Table 17 include the significance and the odds ratios of all 4 variables. We are having the same situation as in CSHA-2. Age and years of education are significant as usual. And history of stroke indicates a negative impact on protection of dementia but not found significant.

Table 16: Significance of selected risk factors, CSHA-3

Variables	Test of Significance
Age (by year)	0.000726577
Sex	0.3149028
Years of Education (by year)	0.000858635
Stroke	0.2131354

*Logistic regression model used: factor(SUMMNEW) ~ CSHA1AGE + factor(SEX) + EDUC + factor(HXSTROKE). The rest of the models are nested within above model.

Table 17: Risk of Dementia associated with selected variables, CSHA-3

Variables	Odds Ratio §	95% confidence interval	
Age (by year)	2.54	1.43	4.51
Sex (Male:Female)	0.63	0.26	1.52
Year of Education (by year)	0.42	0.25	0.71
Stroke (Yes:No)	6.64	0.37	118.01

§ Adjusted for age, sex, and education

3.2 Results for CIND

3.2.1 From CSHA-1 to CSHA-2

The same analysis as the analysis for dementia is done on people who were cognitively normal in CSHA-1(1991) and alive and also cognitively normal or CIND in CSHA-2(1996). There are 426 cases included in the analysis.

Table 18 shows the significance of all 4 variables described above. We notice that age, years of education are highly significant but sex and history of stroke are not significant enough to be considered.

Table 18: Significance of selected risk factors, CSHA-2

Variables	Test of Significance *
Age (by year)	0.0007463988
Sex	0.2505921
Years of Education (by year)	0.0005704541
Stroke	0.2087607

* Logistic regression model used: factor(SUMMNEW) ~ CSHA1AGE + EDUC + factor(SEX)+ factor(HXSTROKE). The rest of the models are nested within above model.

Table 19 includes the odds ratios of each of the 4 variables. Age and years of education once again shows statistically significant association with the risk of CIND. History of stroke is associated with an increase risk of CIND and being female is also increasing the risk of CIND, but both effects are not found statistically significant at all.

Table 19: Risk of CIND associated with selected variables, CSHA-2

Variables	Odds Ratio §	95% confidence interval	
Age (by year)	1.56	1.17	2.09
Sex (Male:Female)	0.71	0.46	1.10
Years of Education (by year)	0.60	0.45	0.78
Stroke (Yes:No)	1.72	0.72	4.12

§ Adjusted for age, sex, and education

3.2.2 From CSHA-1 to CSHA-3

Finally, we perform the analysis on those people who were cognitively normal in CSHA-1(1991) and alive and also cognitively normal or CIND in

CSHA-3(2001). Due to the missing values in the dataset, we have 177 subjects left with information of all 4 factors present.

Table 20 and Table 21 include the significance and the odds ratios of all 4 variables. We are having the same situation as in CSHA-2. Age and years of education are significant as usual. And history of stroke indicates an insignificant negative impact on protection of CIND. In addition, the risk of CIND shifts from female to male in this second study, but is not found significant.

Table 20: Significance of selected risk factors, CSHA-3

Variables	Test of Significance
Age (by year)	0.02324637
Sex	0.07926053
Years of Education (by year)	0.04443382
Stroke	0.1042035

*Logistic regression model used: factor(SUMMNEW) ~ CSHA1AGE + factor(SEX) + EDUC + factor(HXSTROKE). The rest of the models are nested within above model.

Table 21: Risk of Dementia associated with selected variables, CSHA-3

Variables	Odds Ratio §	95% confidence interval	
Age (by year)	1.73	1.04	2.89
Sex (Male:Female)	1.82	0.96	3.43
Year of Education (by year)	0.72	0.53	0.99
Stroke(Yes:No)	5.30	0.55	50.76

§ Adjusted for age, sex, and education

Combining the results from two phases of the study, stroke does show an association with an increase risk of dementia and CIND, and the association is getting stronger over time which can be seen through the odds ratios. However, the confident intervals of the odds ratios do not confirm the association is significant. At the same time, we confirm the strong association of age and years of education with the risk of dementia and CIND once again.

Chapter 4

Conclusion & Extension

In conclusion, through our analysis, we find that age and years of education have a strong significant effect on the risk of dementia and CIND. Younger age and more years of education seem to be protective for Dementia and CIND. Coffee consumption may reduce the risk of dementia whereas the presence of ApoE4 genes may increase the risk of dementia at CSHA-2. The influence of other factors on the risk of dementia and CIND as described in previous 2 chapters isn't found to be significant. History of stroke does show an association with an increase risk of dementia and CIND, but not statistically significant enough to be considered. All finding in our analysis may be considered as hypothesis and to be further investigated through other methods or on other dataset. Because of our way of choosing subsets of data in order to avoid any false negative cases, our sample sets have relative small size. Therefore, the distribution of values for some variables may be largely uneven. In this case, some of the relationships between factors and disease may only indicate themselves but could not form any statistical significance.

Regarding to the extension of this project, there are a lot more things can be done with the data from CSHA. On one hand, a lot of other factors within the dataset have not been analyzed yet, and any hypothesis generated from this dataset can be evaluated on other studies. On the other hand, due to the rich texture of the data from CSHA, such as image data, genetic data, numerical data, text etc., new algorithm can be developed particular for this kind of dataset and the dataset would also be good for testing any existing analytical methods as well. A long term goal for the mathematical research would be to show that the data mining methods of important interest to dementia prevention research could be proved to be statistically valid when generating a hypothesis from a single data set. Our analysis here is only trying to fulfill the immediate term goals which are to understand the structure of the data at hand, and to generate hypotheses about dementia and CIND prevention based on that understanding. This analysis should be just a beginning of a series of analysis. Further research is definitely needed.

References

[1] Joan Lindsay, Danielle Laurin, Rene Verreault, Rejean Hebert, Barbara Helliwell, Gerry B. Hill, Ian McDowell, December 2001, *Risk Factors for Alzheimer's Disease: A Prospective Analysis from the Canadian Study of Health and Aging*, American Journal of Epidemiology

[2] P. McCullagh, and J. A. Nelder, *Generalized Linear Models,* Chapman and Hall, 1983

[3] Stephen E. Fienberg, *The analysis of Cross-Classified Categorical Data*, The Massachusetts Institute of Technology, 1980

[4] Teng EL, Chui HC., *The Modified Mini-Mental State (3MS) examination,* J Clin Psychiatry 1987;48(8):314–8.

[5] The Canadian Study of Health and Aging Working Group., *The incidence of dementia in Canada.* The Canadian Study of Health and Aging Working Group, Neurology 2000; 55(1):66 –73.

[6] The Canadian Study of Health and Aging Working Group. The *Canadian Study of Health and Aging: Study methods and prevalence of dementia.* CMAJ 1994; 150(6): 899-913

[7] Tuokko H, Frerichs R, Graham J, Rockwood K, Kristjansson B, Fisk J, et al. *Five-year follow-up of cognitive impairment with no dementia.*Arch Neurol 2003; 60(4):577– 82.

[8] Wayne Oldford, *Identifying Risk Profiles: A Data Mining and Visualization Approach,* University of Waterloo, 2007

[9] William J. Welch, *Computational Exploration of Data,* University of Waterloo, 2001

[10] Ya-Ping Jin, Silvia Di Legge, Truls Ostbye, John W. Feightner, Vladimir Hachinski, *The reciprocal risks of stroke and cognitive impairment in an elderly population*, Alzheimer's & Dementia 2 (2006) 171-178