Machine Learning of elderly people's willingness to delay retirement

--Based on CHARLS and HRS

Yetong Xu

I. EXECUTIVE SUMMARY

This paper examines whether the characteristics of elderly people can accurately influence their decisions to postpone retirement. With the continued aging of the global population, many countries are experiencing labor shortages and gaps in pension funds, leading to significant pressure on governments to provide resources for elderly care and healthcare. As a result, many countries have implemented policies to delay retirement, including China. However, to successfully implement these policies, we need to consider a question: **What factors influence people's willingness to delay retirement?**

Using three machine learning techniques and logistic regression, I attempted to answer this question by creating models that identify the characteristics that influence elderly people's willingness to delay retirement. These factors include gender, region, education level, family-related factors, health-related factors, social security-related factors, and COVID-19-related factors, etc. I tested these models on data from the 2018 China Health and Retirement Longitudinal Study and the 2020 Health and Retirement Survey for the US. Ultimately, I discovered that there are significant differences in the willingness to delay retirement between men and women in China. Meanwhile, in the US, there is no significant gender difference, and education, pension, and location are the main factors affecting people's willingness to delay retirement. Furthermore, I found that the COVID-19 pandemic did not significantly affect people's willingness to delay retirement. Last but not least, I think China can learn from the US to implement a flexible retirement system.

II. BACKGROUND

In recent years, the aging problem has become an urgent and significant concern for countries worldwide. In 2011, the IMF conducted a comparative study on pension payment pressures in developed and developing countries. The study found that as populations in both developed and developing countries age, the burden of pension payments would increase significantly. This could have an impact on the sustainable development of national pension systems, social stability, and economic development levels. As a result, the study recommended extending the retirement age. Now, the exacerbation of the aging problem has directly led to many countries globally adopting delayed retirement policies. For example, France, Spain, Russia, and the United States have already implemented or are currently implementing delayed retirement policies. Some of these countries have faced significant resistance. Recently, at least 1.1 million people protested on the streets of Paris and other French cities amid nationwide strikes against plans to raise the retirement age — but President Emmanuel Macron insisted he would press ahead with the proposed pension reforms(Corbet, 2023).

In response to the exacerbation of aging, China plans to implement a phased-in

delayed retirement policy in 2023 and achieve same-age retirement for men and women at 65 years old by 2055. According to relevant laws and regulations, the current retirement age in China is 60 years old for men, 55 years old for female cadres, and 50 years old for female workers. Certainly, it is without a doubt that China does not want to face intense opposition from the public, as was the case in France, when implementing its gradual delay retirement policy. Therefore, this study aims to analyze the factors that affect an individual's willingness to delay retirement from a personal perspective. It will then further analyze the obstacles and solutions to implementing the gradual delay retirement policy.

The first study (Wu et al., 2021) examines the factors affecting the willingness to delay retirement in China. The study is based on data from Charls 2015 and finds that financial concerns, work-related factors, and health status are the primary determinants of delayed retirement. The study (Tang, 2020) investigates the obstacles to delayed retirement in China. Her study uses data from the China Family Panel Studies and identifies health issues, social security, and job-related factors as the main barriers to delayed retirement.

The third study (RAND Corporation, 2010) focuses on delayed retirement among Americans and examines its potential benefits for social security and Medicare. The study suggests that delayed retirement could improve the financial stability of these programs. The fourth study (Levanon and Cheng, 2011) analyzes delayed retirement in the United States and its implications for businesses. The study is based on data from the Bureau of Labor Statistics and finds that older workers tend to delay retirement due to financial reasons and job satisfaction. The last study (Saure and Zoabi, 2011) investigates the role of occupations in determining retirement age across different countries, including Germany. The study uses data from the International Social Survey Programme and finds that occupations that require physical labor tend to have lower retirement ages.

Overall, these studies highlight the complex nature of delayed retirement and its determinants, including financial concerns, job-related factors, health status, and social security policies. However, there are some limitations of previous studies.

Firstly, most of the previous studies used data before 2020 or even before 2018. With the passage of time and the development of the era, especially the impact of the global COVID-19 pandemic since 2020, people's willingness to delay retirement may have undergone significant changes. Secondly, most existing studies use a single model and prior assumptions for empirical analysis, such as Probit and Logistic regression models. If the relationship between the independent variable and the dependent variable does not conform to the regression form, it will not be extracted by the model, which can lead to the loss of important variables and result in one-sided empirical results, which cannot achieve a comprehensive, systematic and integrated examination.

III. Data

3.1 Data Description

The data used in this study mainly comes from survey data, specifically the 2018 China Health and Retirement Longitudinal Study (CHARLS, 2018) data and the 2020 Health and Retirement Study (HRS, 2020) data from the United States. The CHARLS survey collected micro-level data from Chinese households and individuals aged 45 and above to analyze the health of the elderly population and aging issues in China. The national baseline survey of CHARLS covered 150 county-level units, 450 village-level units, and approximately 17,000 people from 10,000 households. Meanwhile, The Health and Retirement Study (HRS) is a longitudinal panel study that surveys a representative sample of approximately 20,000 people in America. Through its unique and in-depth interviews, the HRS provides an invaluable and growing body of multidisciplinary data that researchers can use to address important questions about the challenges and opportunities of aging.

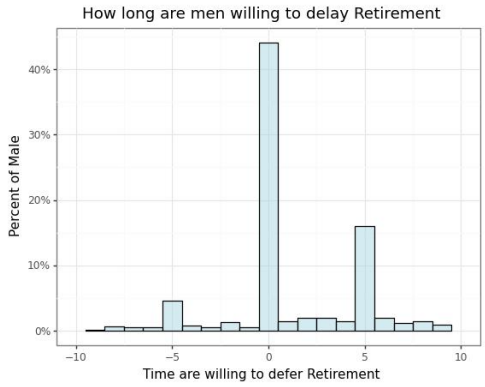
Of course, the data selected for this study has its limitation. I believe that the COVID-19 pandemic since 2020 has had a significant impact on everyone worldwide. In this situation, I prefer to use CHARLS data after 2020. Unfortunately, the research team at Peking University only updated a new round of data starting in the second half of 2020 and did not make this version of data publicly available even now. Therefore, this study can only choose to use the 2018 CHARLS data to investigate the delayed retirement intentions of Chinese elderly people. So, how can we examine the impact of the COVID-19 pandemic on elderly people's intentions to delay retirement? This study plans to use the 2020 HRS data to explore whether the COVID-19 pandemic has changed the intentions of American elderly people to delay retirement. We will observe whether the delayed retirement intentions of elderly people have changed due to economic downturns and other factors caused by the pandemic.

3.2 Descriptive Summary

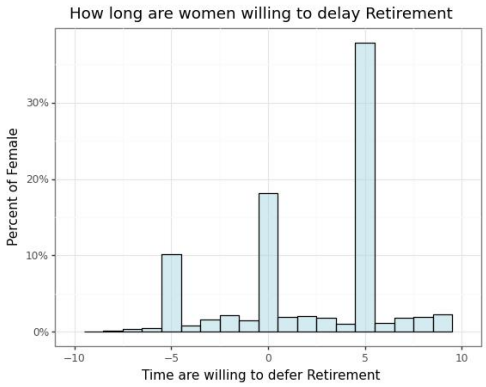
Table 1 Willingness to Delay Retirement

	Male		Female		Total	
	Number	**%**	**Number**	**%**	**Number**	**%**
Charls 2018	1519	53.00%	1525	76.71%	3044	62.71%
HRS 2020	688	73.50%	826	69.35%	1514	71.15%

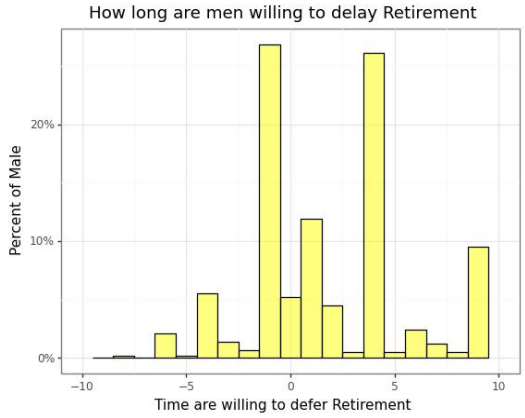
As we can see from Table 1, both in China and the United States, over 50% of men and women have the intention to delay retirement. However, in China, the percentage of women who have the intention to delay retirement is 76.71%, much higher than the 53% of men.



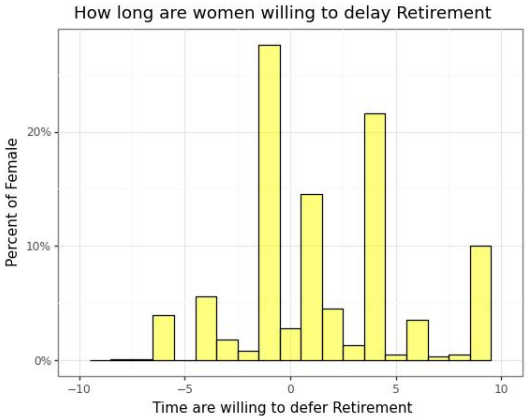
(a) Male in Charls



(b) Female in Charls



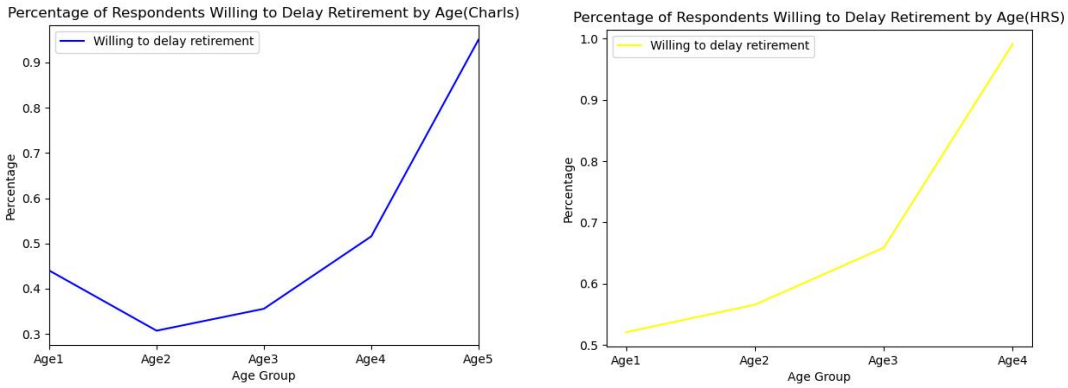
(c) Male in HRS



(d) Female in HRS

Figure 1 Distribution of Delayed Retirement Duration

From Figure 1 (a) and (b), we can see that more than 40% of elderly Chinese men do not have the intention to delay retirement. Among those who are willing to delay retirement, the highest percentage of Chinese elderly men are willing to delay it for 5 years. On the other hand, more than 30% of Chinese elderly women are willing to delay retirement for 5 years, and interestingly, more women than men are willing to retire early. From Figure 1 (c) and (d), we can observe that the distribution of delayed retirement duration is very similar between elderly American men and women.



(a) Charls **(b) HRS**

Figure 2 Delayed Retirement Intention Distribution of different Age group

Figure 2 shows the changes in people's willingness to delay retirement with age. From the figure, we can see that as age increases, the number of middle-aged and elderly people in the United States who are willing to delay retirement gradually increases. In China, the proportion of middle-aged and elderly people who are willing to delay retirement first experiences a brief decline between the ages of 45 and 50, and then gradually increases with age.

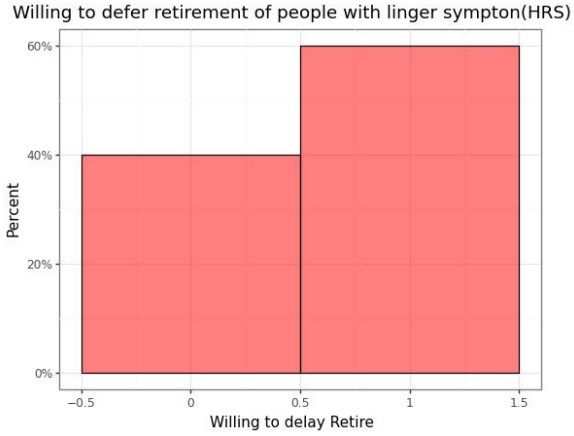


Figure 3 Distribution of delayed retirement intention of peopele who get linger symptons caused by COVID-19

We can see from Figure 3 that among people who experience linger symptoms caused by Covid-19, 60% of them are willing to delay retirement. This proportion is much lower compared to the overall willingness to delay retirement of over 70% in Table 1. We can infer that Covid-19 has a potential impact on people's work and retirement timing choices.

IV. Methodology

My project conducted inference analyses. I aimed to create a model that could identify specific characteristics associated with delayed retirement intentions. For all models in this paper, the outcome variable or target array was delayed retirement intention. In China, according to the latest statutory retirement age, individuals with an expected retirement age greater than 60 for men and 55 for women are considered to have a willingness to delay retirement. In the United States, individuals with an expected retirement age greater than 66 are considered to have a willingness to delay retirement. The feature matrix of individuals' characteristics, both in Charls and HRS, include sex, age, marital status, location, education level, annual salary, type of employer, whether they have grandchildren to take care of, whether they need to provide financial help to children, whether they have a pension, and whether they have medical insurance. Additionally, in the HRS, this study includes variables related to Covid-19, such as lingering symptoms, contract Covid-19 or not, household member contract Covid-19 or not, and work affected, etc.

To investigate the characteristics associated with people's delayed retirement intention, I will use some non-parametric techniques such as neural networks, random forests, and XG-Boost to identify which variables are important. Then I will further use a parametric technique, the logistic regression model, to study the mechanism of the characteristics influencing the willingness to delay retirement.

Neural networks can handle linear and nonlinear data relationships, especially in nonlinear data fitting. The neural network model used in this paper has powerful nonlinear mapping capabilities. Since many independent variables in our data are nonlinearly related to the dependent variable, and the strength and form of the relationships are unknown, neural networks can solve this problem well.

The random forest model builds a forest in a random way, which contains many decision trees that have high prediction accuracy and weakly or even non-correlated. Random forests are actually an ensemble learning method and it can handle high-dimensional data. During the training process, random forests can identify variables that play an important role in classification and provide measures of the importance of these variables.

XG-Boost is a type of boosting algorithm, which integrates many weak classifiers into a strong classifier. The basic idea of XG-Boost algorithm is to continuously generate trees and grow trees by continuously splitting feature variables. Each time a tree is generated, a new function is re-learned to fit the residual of the previous prediction, so as to continuously improve the learning quality and approximate the actual value.

Logistic regression model is a multivariate regression analysis method mainly used to study the relationship between a binary dependent variable and its influencing factor variables, that is, to study whether an event occurs and what is the probability of its occurrence when the influencing factor variables take different values.

All the methods above have been used widely in extent literature. For example, Seligman and Tuljapurka (2017) used neural networks and random forest to investigate how machine learning may add to the understanding of social determinants of health based on the Health and Retirement Study. Schiltz et al. (2017)

also used random forest to identify specific combinations of chronic conditions, functional limitations, and geriatric syndromes associated with direct medical costs and inpatient utilization. Ogunleye and Wang (2019) used the XGBoost model for Chronic Kidney Disease Diagnosis. Additionally, Zwerling et al. (1996) used a logistic regression model to examine risk factors for occupational injury among older workers.

V. ANALYSIS

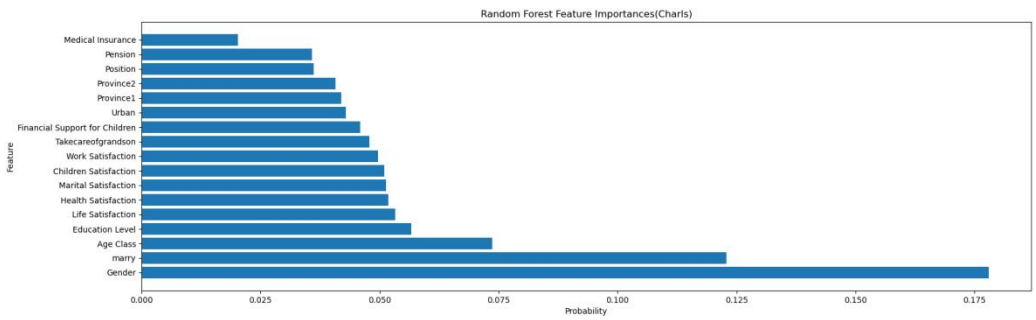
5.1 Fit in three Machine Learning Models

After cleaning the data and performing a series of preprocessing steps, I ran the Charls and HRS data separately through Random Forest, Neural Networks, and XG-Boost. I set the training size to 80%, which means that 20% of the samples were set as the test group. The results of the models are shown in the following table. For both Charls and HRS, we can see that Random Forest has the highest prediction accuracy, at 71.82% and 76.30%, respectively. At the same time, we can see that in Charls, Random Forest has the highest recall, while in HRS, Random Forest's recall is only slightly lower than XG-Boost's recall. In comparison, we can basically conclude that Random Forest performs the best in this study.

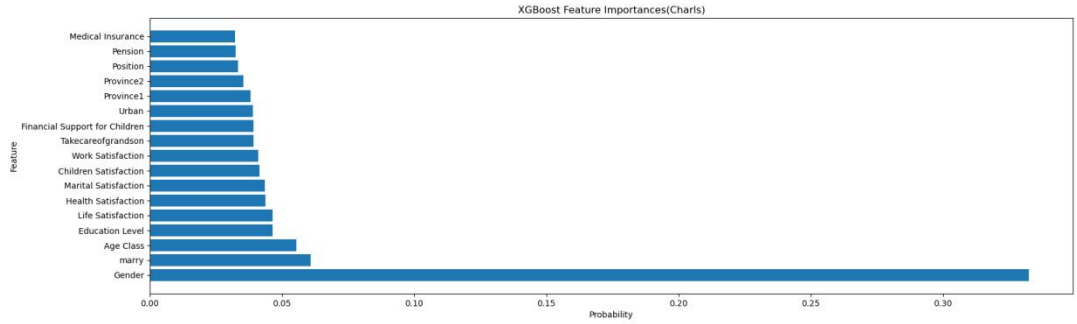
Then, I listed the most important factors for delayed retirement intentions found by Random Forest and XGBoost in the Charls and HRS, respectively, as shown in Figure 4 and Figure 5. After comparing the important variables selected by Random Forest and XGBoost, I found that in Charls, they both selected gender, age, education, marital satisfaction, and urban-rural status, etc. In HRS, they both selected age, education, number of grandchildren, number of children, and whether someone had died due to Covid-19, etc.

Table 2 Models' Accuracy and Recall

	Random Forest		Neural Networks		XG-Boost	
	Accuracy	**Recall**	**Accuracy**	**Recall**	**Accuracy**	**Recall**
Charls 2018	71.82%	51.72%	66.36%	41.38%	64.55%	39.39%
HRS 2020	76.30%	80.33%	68.89%	80.19%	73.33%	82.41%

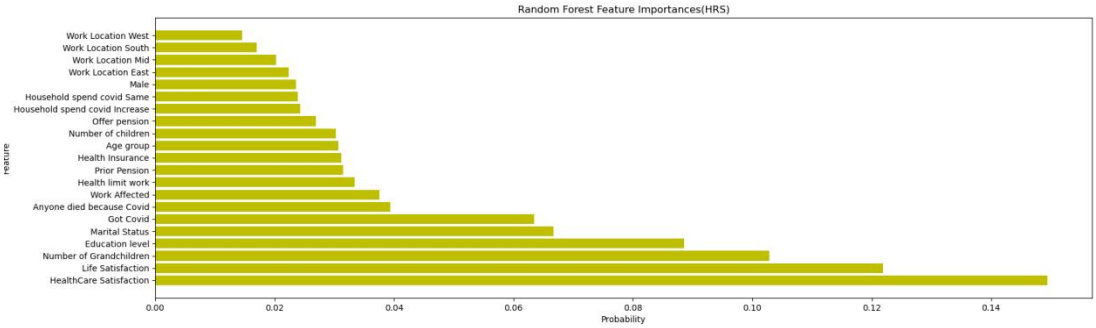


(a) Random Forest

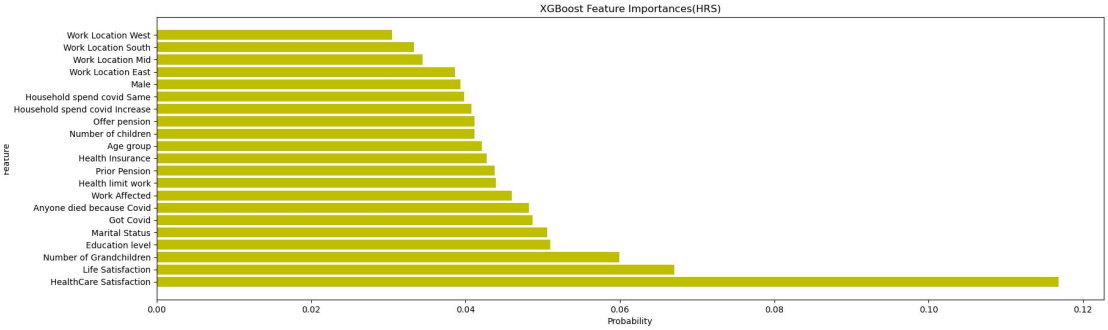


(b) XG-Boost

Figure 4 Important Features of Charls



(a) Random Forest



(b) XG-Boost

Figure 4 Important Features of HRS

From this, I have found that in the process of classification modeling, neural networks can fit the nonlinear relationship between variables well, but cannot provide a measure of variable importance. Random Forest can provide a measure of variable importance, but due to the limitations of the model, it is prone to overfitting. XG-Boost solves the problem of overfitting well, and can also be compared with the important variables of Random Forest. However, from the perspective of the classification model, we can only determine that these variables have an important

relationship with "delayed retirement intentions", and the specific quantitative relationship and significance need to be determined through a regression model.

5.2 Fit in Logistic Model

Based on previous research, we found that in China, gender is an important factor affecting people's delayed retirement intentions, while the impact of gender on Americans' delayed retirement intentions is not as significant. Therefore, we conducted separate logistic regressions by gender on the Charls data and a logistic regression on the full sample for the HRS data. The regression results are shown in Tables 3-5 below.

Table 3 Logistics Results of Male (Charls)

	Coef	p-value	sig	sig2	sig3
Marital Status	-3.08	2.63e-07	*	**	***
Age	0.47	7.00e-03	*	**	***
Province1	0.73	4.21e-02	*	**	
Province2	0.90	1.16e-02	*	**	

Table 4 Logistics Results of Female (Charls)

	Coef	p-value	sig	sig2	sig3
Marital	0.92	0.04	*	**	
Education	0.52	0.02	*	**	
Marital Satisfaction	0.55	0.08	*		
Financial Support for Children	1.85	0.02	*	**	
Urban	-0.55	0.04	*	**	
Province2	0.67	0.04	*	**	
Work Position	-0.78	0.003	*	**	***
Medical Insurance	-1.02	0.05	*	**	

From the comparison between Table 3 and Table 4, we can see that women's delayed retirement intentions in China are influenced more by family factors. Women with higher marital satisfaction and those who provide financial support to their children are more likely to delay their retirement. This may be related to the fact that the previous generation of Chinese women placed more emphasis on family values and had more traditional views. We also found that women with higher education

levels are more likely to delay retirement, and rural women are more willing to delay their retirement compared to urban women. This is reasonable because among many people with non-agricultural jobs in rural areas are self-employed, so the retirement boundary is more blurred.

Table 5 Logistics Results (HRS)

	Coef	p-value	sig	sig2	sig3
Education	0.07	2.00e-02	*	**	
Age	0.57	2.85e-10	*	**	***
Pension	-0.58	5.25e-03	*	**	***
Work Location Mid	-1.47	2.92e-02	*	**	
Work Location South	-1.26	5.25e-02	*		
Work Location West	-1.59	1.87e-02	*	**	

Table 5 shows that in the United States, the factors that have the greatest impact on people's willingness to delay retirement are educational level, age, pension, and location of work. As educational level increases and age increases, people are more likely to delay retirement. However, as people have more pension, their willingness to delay retirement decreases.

Additionally, we can see that people in the Mid, Southern, and Western parts of the United States are less willing to delay retirement compared to those in the Eastern part of the country. Among them, people in the Western part of the United States are the least willing to delay retirement. This may be related to factors such as cost of living, cultural background, and job opportunities in different regions. However, the more detailed and accurate reasons behind these patterns will require further exploration in the future.

VI. CONCLUSION

There is no doubt that this study has some limitations. While the Charls and HRS surveys are comprehensive, they may not capture all relevant factors that could influence delayed retirement intentions. Additionally, the data may be subject to selection bias, as the surveys only include individuals who agree to participate in the study. Also, while these two surveys are large, the analysis may be limited by the sample size, especially when looking at specific subgroups of the population. What's more, the age cutoffs used to define delayed retirement intention may not be applicable to all individuals or populations, and could potentially skew the results.

6.1 Ethical Considerations

One ethical consideration is informed consent. Informed consent means that participants should be fully informed about the purpose of this study, the procedures involved, and the potential risks and benefits of participation, and they should be

given the opportunity to voluntarily consent to participate. In this study, the data used are already collected from existing surveys, and the consent procedures for these surveys are not disclosed. However, we can assume that the survey organizations have obtained informed consent from the participants before collecting their personal information.

Another ethical consideration is the protection of participants' confidentiality and privacy. If the data collected contains sensitive information, such as personal financial details, health status, or family relationships, there is a risk that this information could be leaked or hacked, leading to harm to the participants. This could result in emotional distress, stigmatization, and discrimination, particularly for vulnerable groups such as older adults or those with chronic illnesses. In this study, the data used are de-identified, which means that the personal information of the participants has been removed or masked, so that they cannot be identified.

Additionally, it is important to consider the potential impact of the research on the participants and on society as a whole. The results of this study could have implications for policy and practice related to retirement planning and workforce management. It is important to consider the potential consequences of these implications and to ensure that the research is conducted in a fair and unbiased manner.

6.2 Policy Implications and Future Considerations

In China, there are significant differences in the willingness to delay retirement between men and women. Women are more susceptible to family-related factors such as marital satisfaction and financial support for children. Meanwhile, the impact of education on women is also higher, as highly educated women are more willing to delay retirement. In contrast, in the US, we find that there is not a significant gender difference, and education, pension, and location are the main factors affecting people's willingness to delay retirement.

Furthermore, this study basically negates the impact of the COVID-19 pandemic on people's willingness to delay retirement. I think this is reasonable because although the COVID-19 pandemic has swept the world for three years, most regions have resumed normal life rhythms.

However, it is still worth noting that there are differences in national conditions and culture between China and the United States. I hope that in the future, I can further explore whether the COVID-19 pandemic has an impact on the retirement intentions of the Chinese people.

Based on this study’s results, I think China can learn from the US and implement a flexible retirement system. The flexible retirement system stipulates the legal age for early retirement and normal retirement. People who retire before the early retirement age can receive partial pension benefits each month, and those who retire at the normal retirement age can receive full pension benefits. Those who retire later than the retirement age can not only receive full pension benefits but also receive certain bonus payments. This gradual and flexible retirement system gives individuals

greater choice and can motivate some people to continue working, reducing government spending on elderly care and easing the strain on medical resources, and promoting social and economic development to some extent.

Also in the future, I want to incorporate data from other countries or surveys to gain a more comprehensive understanding of the factors that influence delayed retirement intentions across different populations and contexts. Moreover, I want to use more comprehensive measures of delayed retirement intention, including more factors to gain a more nuanced understanding of the reasons behind delayed retirement intentions. Last but not least, I think I can also try to do longitudinal studies, which can help to better understand the trajectory of delayed retirement intentions over time, and how they are influenced by different factors.

BIBLIOGRAPHY

Corbet, S. (2023, January 19). France: Over 1 million march against raising retirement age. AP News. https://apnews.com/article/france-retirement-age-limit-protests-866eb86aea5cf0d39894b96d2888c26f

Wu, Y., Liu, Y., & Liu, X. (2021). Delayed retirement willingness and its influencing factors: Evidence from the China Health and Retirement Longitudinal Study. Jilin University Journal Social Sciences Edition, 61(3), 1-11. https://doi.org/10.15939/j.jujsse.2021.03.sh1

Tang, S. (2020). The resistance of delayed retirement [Doctoral dissertation, Capital University of Economics and Business]. CNKI. https://doi.org/10.27338/d.cnki.gsjmu.2020.000611

Anonymous. (2010). Delayed retirement among Americans may bolster future of Social Security and Medicare, study finds [Press release]. NewsRx Health.

Levanon, G., & Cheng, B. (2011). US workers delaying retirement: Who and why and implications for businesses. Business Economics, 46(4), 223-232.

Saure, P. U., & Zoabi, H. (2011). Retirement age across countries: The role of occupations. SSRN Electronic Journal. http://apps.eui.eu/Personal/JMF/PhilipSaure/Retirement_SNB.pdf

Zhao, Y., & Strauss, J. (2020). China Health and Retirement Longitudinal Study [Data set]. National School of Development, Peking University. https://charlsdata.com/pages/Data/2018-charls-wave4/zh-cn.html

Weir, D. R. (2022). Health and Retirement Study [Data set]. The National Institute of Aging. https://hrsdata.isr.umich.edu/data-products/public-survey-data?_ga=2.160246261.910769648.1677873413-900753830.1677520506

Seligman, B., Tuljapurkar, S., & Rehkopf, D. (2018). Machine learning approaches to the social determinants of health in the health and retirement study. SSM - Population Health, 4, 95-99. https://doi.org/10.1016/j.ssmph.2017.11.008.

Schiltz, N. K., Warner, D. F., Sun, J., Bakaki, P. M., Dor, A., Given, C. W., Stange, K. C., & Koroukian, S. M. (2017). Identifying Specific Combinations of Multimorbidity that Contribute to Health Care Resource Utilization: An Analytic

Approach. Medical care, 55(3), 276–284. https://doi.org/10.1097/MLR.0000000000000660

Ogunleye, A., & Wang, Q. -G. (2020). XGBoost Model for Chronic Kidney Disease Diagnosis. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 17(6), 2131-2140. doi: 10.1109/TCBB.2019.2911071.

Zwerling, C., Sprince, N. L., Wallace, R. B., Davis, C. S., Whitten, P. S., & Heeringa, S. G. (1996). Risk factors for occupational injuries among older workers: An analysis of the health and retirement study. American Journal of Public Health, 86, 1306-1309. https://doi.org/10.2105/AJPH.86.9.1306