Modeling Aging Trajectories and Urban-rural Disparities in Chronic Pain among Middle-aged and Older Chinese: Evidence from the Generalized Estimating Equations (GEE) and Generalized Linear Mixed Models (GLMMs)

Siyuan Chen

Student ID: 01908889

Promotor: Prof. Dr. Ronan Van Rossem

A dissertation submitted to Ghent University in partial fulfilment of the requirements for the degree of Master of Science in Statistical Data Analysis.

Academic year: 2023 - 2024

GHENT UNIVERSITY

The author and promoter give permission to consult this master dissertation and to copy it or parts of it for personal use. Every other use falls under the restrictions of the copyright, in particular concerning the obligation to mention explicitly the source when using results of this master dissertation.

Gent, August 23, 2024

The promotor,	The author,
Prof. Dr. Ronan Van Rossem	Siyuan Chen

Content

Contents i

Abstract ii

Table of Contents

Abstract

This study investigates the urban-rural and socioeconomic disparities in chronic pain prevalence and how these disparities evolve with aging, using nationally representative longitudinal data from China (2011-2020). The research applies two popular longitudinal analysis methods, Generalized Estimating Equations (GEE) and Generalized Linear Mixed Models (GLMMs), to examine the urban-rural disparities in pain prevalence and capture how they evolve over time. By examining pain prevalence across different socioeconomic groups and urban-rural populations, the study aims to provide insights into policies aimed at reducing pain inequalities, particularly in the context of China's rapidly aging population.

Our findings indicate that the rural population consistently reported the highest pain prevalence, followed by the semi-urban population, while the urban population exhibited the lowest risk of pain, demonstrating a health gradient in pain prevalence among urban-rural populations. In addition, we observe that pain prevalence increases with age, yet such pain-aging trajectories vary across the urban-rural populations, showing a converging trend in pain risk over the life course.

In addition, despite similar empirical conclusions from both modeling strategies, GEE and GLMM still have their own advantages and disadvantages, with GEE emphasizing more on marginal effects but lacking the possibility of exploring model complexity. Although GLMM can examine both fixed effects and random effects, its overly complex model leads to low computational efficiency and is sensitive to assumptions regarding random effects and error structures.

Chapter 1. Research Objectives and Outline

1.1 Introduction

1.1.1 Background of the Chronic Pain Context

Chronic pain (henceforth pain) is acknowledged as a barometer of health (Rubin and Zimmer 2015; Smith et al. 2018), closely linked to the quality of life for individuals (McNamee and Mendolia 2014) and imposing substantial economic costs on national healthcare systems (Gaskin and Richard 2012). It has received considerable attention in both biomedical and neuropsychological research, with a special focus on the proximate causes of pain at the individual level, such as emotional stress, traumatic experiences, inflammation, and degenerative diseases during the aging process (Cohen et al. 2021; Crofford 2015; Goubert and Trompetter 2017; Van Alboom et al. 2023). However, there remains a lack of understanding regarding why certain populations suffer more from pain while others experience lower prevalence and severity. To comprehend pain inequality fully at the population level, we must adopt a socio-structural lens, commonly referred to as the upstream cause of causes, as proposed by the sociology of pain (Zajacova et al. 2021b).

While research on social determinants of pain is still in its infancy (Peele and Schnittker 2022), there has recently been a notable increase in the literature, yielding some consistent conclusions, such as the higher risk of pain among women and lower socioeconomic classes (Goosby 2013; Grol-Prokopczyk 2017; Kennedy et al. 2014; Jay et al. 2019; Zajacova et al. 2020; Topping and Fletcher 2024). Nonetheless, research on broader socioeconomic characteristics and their intersections remains limited (Zajacova et al. 2021b), hindering a deeper understanding of how pain disparities are produced and maintained. For instance, although it is well documented that elevated pain is highly linked to the aging process (Rustøen et al. 2005), there is little empirical research on whether and how the pain-aging trajectories vary across socio-demographic groups. Besides, the scarce existing research is predominantly based on US and European data (Zimmer et al. 2020), leaving an absence in empirical research from non-Western societies that bear a disproportionately high pain burden (Blyth et al. 2019).

China provides one of the best contexts to address the aforementioned knowledge gap. The country has been experiencing a rapid population aging, with the proportion of individuals aged 65 and older increasing from 7% in 2000 to 12.6% in 2019, and this number is predicted to reach to 17.1% by 2030 (Zhan et al. 2021). Consequently, the healthcare burden associated

with pain is also projected to surge substantially due to the growing number of high-risk pain sufferers. Meanwhile, the longstanding household registration (*hukou*) system divides the population into urban and rural categories, resulting in profound inequalities in life chances and several health outcomes between the urban and rural populations (Dorélien and Xu 2020; Fu et al. 2018; Lu and Qin 2014; Wu 2019). However, it is still unclear whether these urban-rural inequalities are replicated in pain prevalence. Moreover, a comprehensive understanding of life-course related changes in pain disparities between urban and rural populations also remains absent. In addition, little is known about whether there are inequalities in chronic pain across socioeconomic status indicators such as education and income, and how such differences change with aging. Identifying vulnerable groups suffering from pain is crucial for providing vital insights for health policies aimed at reducing pain inequalities in the future, not only for China but also for countries experiencing concurrent population aging amidst deep-rooted institutional inequalities.

1.1.2 Research Questions

Hence, utilizing nationally representative longitudinal data from 2011 to 2020 from China, this study empirically contributes to the literature by answering the following questions:

Research Question 1:

Does the chronic pain prevalence vary across urban and rural populations?

Research Question 2:

Whether these disparities in chronic pain prevalence evolve across the life course?

1.2 The Necessity of Using Longitudinal Data Analysis Methods

In this research, a central aim is to uncover and understand how individual aging influences pain prevalence and whether it modifies the underlying socioeconomic gradient in pain prevalence. To capture the dynamic characteristics, it is essential that our data could reflect changes in pain at individual-level over time, necessitating the use of longitudinal follow-up survey data rather than cross-sectional or pooled cross-sectional data.

Although some studies have attempted to examine dynamic characteristics by combining multiple cross-sectional survey datasets to create a pooled dataset (Yang, 2008; Yang et al., 2004; Yang & Land, 2006, 2008), this method has limitations. Specifically, the practice of replacing samples in these cross-sectional surveys means that different respondents are surveyed at each time point, capturing only a snapshot of respondents' conditions at the time of the survey. Consequently, such data do not allow for the exploration of continuous processes. While this approach can partially account for the impact of survey year (period) by incorporating time indicators into analytical models, it fails to distinguish between 'between-individual effects' and 'within-individual effects.' As a result, it is inadequate for addressing research questions related to individual change and development over time.

Another flawed approach is to treat longitudinal data as pooled cross-sectional data with a time indicator distinguishing the survey wave, disregarding the fact that repeated measurements are nested within the same set of individuals or units. This oversight can lead to significant problems, as these correlated data violate the assumption of independent and identically distributed (i.i.d.) observations when applying conventional modeling techniques such as linear regression or generalized linear models. As a result, the standard errors of parameter estimates may be inaccurately estimated (even in the large sample), leading to incorrect statistical inferences and undermining the validity of the study's conclusions (Fitzmaurice et al. 2012; Hoffman 2015).

Therefore, in order to capture how pain prevalence changes with aging and to obtain consistent and efficient parameter estimates, we will use two common longitudinal data analysis frameworks: Generalized Estimating Equation (GEE) and Generalized Linear Mixed Models (GLMMs).

1.3 Outline: The Roadmap through This Thesis

In Chapter 1, we provide an overview of the background on chronic pain, present the research questions, and highlight the importance and necessities of employing longitudinal data analysis approaches to answer the research questions.

In Chapter 2, we offer a concise overview of two common longitudinal data analysis methods: Generalized Estimating Equations (GEE) and Generalized Linear Mixed Models (GLMMs). This chapter also introduces their basic model settings, assumptions, parameter estimation procedures, and relevant statistical inferences.

In Chapter 3, we present the overview of the datasets and measurements utilized in the empirical study, as well as the model specifications for the corresponding estimation methods. Additionally, we address potential challenges related to measurement inconsistency and mortality selection, and outline the strategies employed to mitigate these issues in the robustness checks.

In Chapter 4, we provide an exploratory analysis of the data, setting the foundation for subsequent statistical modeling. Then, we report the results obtained from the Generalized Estimating Equations (GEE) and Generalized Linear Mixed Models (GLMMs) and then compare these findings.

Finally, in Chapter 5, we discuss the results from the empirical analysis and evaluate the cons and pros of the two modeling approaches.

In the appendix, we include some descriptive statistical results that are not included in the main text, the results of the model robustness analysis, and the R code involved in the statistical modeling.

Chapter 2 Overview of the GEE and GLMMs

The repeated measures ANOVA is a popular technique for analyzing the measurement results of the same group of subjects under different conditions or time points, especially suitable for analyzing changes within the group and comparing differences between different conditions or time points. However, when our data is an unbalanced panel data, for example, the time intervals of repeated measurements are inconsistent, some respondents are missing a large number of time points, or there is a serious autocorrelation between the measured data, then using repeated measures ANOVA may lead to wrong conclusions (Park et al. 2009). Therefore, when the assumptions of repeated measures ANOVA are not met, researchers tend to apply the alternative approaches, for instance, generalized estimating equation (GEE) and generalized linear mixed models (GLMM).

Generalized estimating equation (GEE) and generalized linear mixed models (GLMM) are both commonly used statistical methods for dealing with data with correlation structure, when dealing with longitudinal data (repeated measurement data) or clustered data (such as multi-level structured data). In this chapter, we first quickly review their basic model settings, assumptions, parameter estimation, and relevant statistical inferences, and the specific analytic strategy and model setting are presented in Chapter 3.

2.1 Generalized Estimating Equation (GEE) for Correlated Data

Based on the Generalized Linear Models (GLM), which cannot handle the violation of the independent and identically distributed assumption, the Generalized Estimating Equation (GEE) is a method developed to relax the above assumptions by introducing a working correlation structure to deal with the correlation between observations. The term GEE was used because the model is derived from a generalization of the GLM estimating equation. The most commonly described GEE model was introduced by Liang and Zeger in 1986, this regression framework is designed to analyze correlated data in a population-averaged approach. Due to its flexibility, GEE is also widely applied for handling correlated longitudinal and clustered data (Ding 2024; Hardin and Hilbe 2002). GEE utilizes quasi-likelihood estimation methods to estimate parameters within generalized linear models, effectively accommodating the complexities inherent in repeated measurement data (Ding 2024; Liang and Zeger 1986; Zeger and Liang 1986).

The GEE can be written as:

$$g(\mu_{it}) = E(Y_{it} \mid X_{it}) = \boldsymbol{X}_{it}^{T} \boldsymbol{\beta}$$

with $\mathrm{Var}(Y_{it}) = \mathrm{v}(\mu_{it})\phi$

$g(\cdot)$ is the setting of link function; $\mathrm{v}(\cdot)$ presents the relation between μ_{it} and variance; ϕ is unknown parameter of scale. Using the estimated covariance matrix, we can update the GEE estimate to improve efficiency (Ding 2024). There are three common choices of the working covariance structure matrix, called "independent", "exchangeable", and "unstructured".

Independent working covariance matrix:

$$R_{it} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}$$

Exchangeable working covariance matrix:

$$R_{it} = \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{pmatrix}$$

Unstructured working covariance matrix:

$$R_{it} = \begin{pmatrix} 1 & a & \cdots & j \\ a & 1 & \cdots & k \\ \vdots & \vdots & \ddots & \vdots \\ j & k & \cdots & 1 \end{pmatrix}$$

In longitudinal data analysis, the elements on the main diagonal of the covariance matrix are all equal to 1, as they represent the correlation between the t-th observation of the i-th individual and itself. In the Independent working covariance matrix, the off-diagonal elements are all 0, indicating that each repeated measurement within an individual is independent. The Exchangeable working covariance matrix allows for correlation between repeated measurements, with a constant correlation coefficient ρ. The Unstructured working covariance matrix further relaxes these assumptions by allowing for varying correlation coefficients.

Although different choices of the working covariance matrix do not affect the consistency of the estimation (Ding 2024), to get a more efficient $\hat{\beta}$, we still need to consider the strategy of choosing the working covariance matrix. While the independent working covariance matrix is sufficient for many applications despite potential efficiency losses, a carefully selected working covariance matrix can lead to efficiency gains compared to the simpler independent covariance matrix. Therefore, in practice, one usually starts with an independence or exchangeability structure. If these structures do not capture the correlation in the data well, one can try other structures such as an unstructured working covariance matrix.

In addition, as for the parameter estimation, GEE uses the method of "quasi-likelihood estimation", which does not require the full specification of the data distribution, but only requires the specification of the mean and variance structure of the data. The parameters of GEE are obtained by solving an iterative estimating equation based on the assumed correlation structure and an extension of the generalized linear model (GLM). Moreover, the Wald test is a commonly used method in the GEE model to test whether a single or multiple regression coefficients are significant.

Regarding model fit evaluation, QIC (Quasi-likelihood under the Independence model Criterion) is provided to indicate the goodness of fit of the model. QIC is a model selection criterion, similar to the role of AIC (Akaike Information Criterion) in traditional models, with a smaller value indicating the better model, and can be used to compare the analysis results of correlation matrices of different operations, thus is widely used to compare different GEE models to select the most optimal one.

Although GEE does not require data distribution and independent outcomes, there are several key issues that need to pay attention to ensure the validity of the model and the reliability of the results. First, though data points within the same individual can be correlated, data between different individuals are assumed to be independent. Second, we need to use the right link function to correctly reflect the relationship between the response variable and the linear predictors. Third, even if the working correlation matrix assumption is not completely correct, the parameter estimation of GEE is still robust, but the closer the working correlation matrix assumption is to the actual situation, the more efficient the estimation results. Therefore, it is still necessary to compare and select the most appropriate working covariance matrix.

2.2 Generalized Linear Mixed Models (GLMMs) for Correlated Data

When the observations are not independent and the i.i.d. assumption is violated, another common method to solve the problem caused by intra-group correlation is linear mixed models. If the link function extends to non-linear format, for instance, considering the outcome is a categorical measure, we extend the linear one to a logit setting, thus get the generalized linear mixed models (GLMMs). Different from the GEE approach focusing on the "population-averaged" effect, GLMMs dealing with correlated or clustered measures by incorporating the fixed effect and random effect, thus named the mixed effect model (Fitzmaurice et al. 2012).

In GLMM, fixed effects are used to describe the average effect of the population level, while random effects are used to capture the variation caused by the hierarchical structure of the data or differences between groups. GLMM not only models the explainable part of the data (fixed effects), but also captures the unexplainable part (random effects), thus more accurately reflecting the complexity of the data (Hox et al. 2017; Rabe-Hesketh and Skrondal 2008).

In this study, we use the multilevel modelling (MLM) framework (also called hierarchical linear model, HLM) to deal with the longitudinal data, which is a series of models designed to analyze mixed effects using nested or longitudinal data, integrating predictors from different levels to explain a common dependent variable. This model incorporates measurements across various levels and accounts for the contributions of information at different hierarchies (Bryk and Raudenbush 1992). In its mathematical expression, the intercept or slope at a lower level can be explained by predictors at a higher level. This approach not only accurately addresses the computation of model parameters in multilevel data analysis but also enables the simultaneous examination of micro and macro variables, as well as cross-level interaction effects. Additionally, it can mitigate estimation bias arising from correlated error terms across different levels, allowing for the estimation and analysis of both fixed and random effects.

In this section, we quickly review the framework of the multilevel modelling using the linear model as an example. In the empirical part of the paper, the multilevel modelling and its specific parameter settings are described in detail in Chapter 3.

First, we start from the null model or the unconditional means model. The equation is as following:

Level-1 / within-individual level:

$$Y_{ti} = \beta_{0i} + u_{ti}$$

Level-2 / between-individual level:

$$\beta_{0i} = \gamma_{00} + \delta_{0i}$$

Substitution of the second equation into the first one we get:

$$Y_{ti} = \gamma_{00} + \delta_{0i} + u_{ti}$$

Where $u_{ti} \sim N\ (0,\ \sigma_u^2)$, $v_{0i} \sim N\ (0,\ \sigma_\delta^2)$, and Cov($u_{ti}$, δ_{0i}) = 0.

Using the unconditional mean model, we can estimate the proportion of total variation attributable to each level (e.g., within and between individuals). This helps identify sources of variation and guides subsequent model construction. Specifically, we decompose the outcome variable's variance into within- and between-individual components and calculate the proportion of between-individual variance using the intraclass correlation (ICC), ranging from 0 to 1, with the following equation:

$$\text{ICC} = \frac{\sigma_b^2}{\sigma_w^2 + \sigma_b^2} = \frac{\sigma_\delta^2}{\sigma_u^2 + \sigma_\delta^2}$$

Where σ_{w}^2 represents the within-group (within-individual) variance, σ_b^2 is the between-group (between individual) variance. The value $\sigma_{\text{w}}^2 + \sigma_b^2$ represents the total variance. The larger the ICC, the more similar the observations in the same group are, the smaller the individual differences within the group, and the greater the proportion of inter-group differences in the total variation, indicating that the group (or random effect) considered in the model has a greater impact on the data. Therefore, in this case, ignoring the hierarchy or group effect may lead to inaccurate models or large errors.

Next, we could add the predictors and covariates into the MLMs and thus allow the random-coefficients regression models. For instance, we could consider the existing of random intercept and random slopes at the same time. The MLM can be written as:

Level-1 / within-individual level:

$$Y_{ti} = \beta_{0i} + \beta_{1i} X_{ti} + \varepsilon_{ti}$$

Level-2 / between-individual level:

$$\beta_{0i} = \gamma_{00} + \gamma_{01} W_i + \tau_{0i}$$

$$\beta_{1i} = \gamma_{10} + \gamma_{11} W_i + \tau_{1i}$$

Substitution of the second equation into the first one we get:

$$Y_{ti} = \gamma_{00} + \gamma_{01} W_i + \tau_{0i} + \gamma_{10} X_{ti} + \gamma_{11} W_i X_{ti} + \tau_{1i} X_{ti} + \varepsilon_{ti}$$

Where $\varepsilon_{ti} \sim N\ (0,\ \sigma_{\varepsilon}^2)$, $\begin{pmatrix} \tau_{0i} \\ \tau_{1i} \end{pmatrix} \sim \mathrm{N}\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{00}^2 & \sigma_{01}^2 \\ \sigma_{10}^2 & \sigma_{11}^2 \end{pmatrix}\right]$

By setting the value of τ_{0i} to non-zero, we can obtain a random intercept model, and by setting the value of τ_{1i} to non-zero, we can obtain a random slope model.

In multilevel modeling (MLM), parameter estimation involves estimating fixed and random effects. Given the complexity of MLM's hierarchical structure, estimation is more intricate than in general linear models, typically using Maximum Likelihood Estimation (MLE) or Restricted Maximum Likelihood Estimation (REML). Estimation is often performed through iterative algorithms, like the Expectation-Maximization (EM) or Newton-Raphson algorithms, which maximize the likelihood function by adjusting the parameters iteratively.

Model evaluation and comparison can also be achieved through Log-likelihood, which is reported as -2LL (-2 * loglikelihood), with a smaller value meaning better model fitting. In addition, when comparing models, it is not appropriate to use the likelihood ratio test (LR test) for non-nested models. In this case, alternative approaches can be used for various information criteria developed based on -2LL including AIC (Akaike information criteria), and BIC (Bayesian information criteria), with smaller values indicating better model fitting.

Chapter 3 Data and Methods

3.1 Dataset and analytic sample

We used data from the China Health and Retirement Longitudinal Study (CHARLS)[1], a nationally representative survey of the Chinese population aged 45 years and older, conducted by Peking University. CHARLS adopted a stratified multistage probability sampling design to obtain the sample, which covered 450 communities from 150 county-level units in 28 provinces, and interviewed 17708 participants from 10257 households, in the baseline wave (Zhao et al. 2014). After the baseline wave in 2011, CHARLS conducted four follow-up surveys in 2013, 2015, 2018, and 2020.

We excluded individuals (N = 1229, percentage = 6.9%) with missing information on outcome, urban-rural classification, demographic and educational indicators used for this study at the baseline wave. After incorporating data from four subsequent waves, we obtained an unbalanced panel dataset comprising 16479 participants.

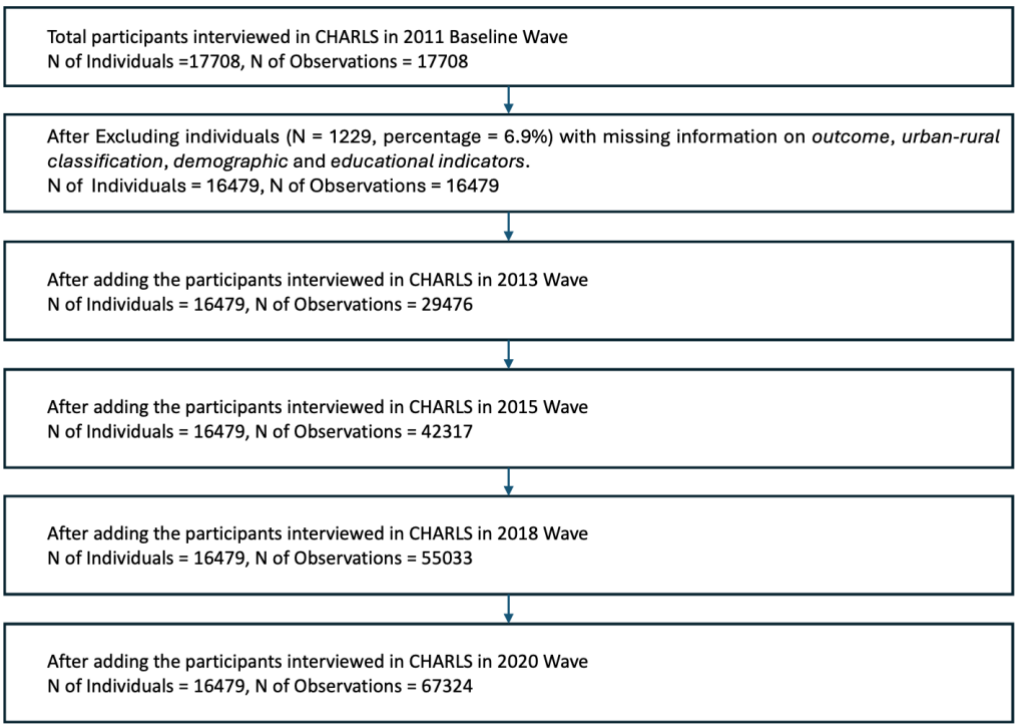

Figure 1. Data cleaning flow for the final analytic samples

[1] The CHARLS is an open source data set. User can submit an application to download it from the official website. The website link is: https://charls.pku.edu.cn/en/

3.2 Measures

Pain. In 2011, 2015, 2018, and 2020 waves of CHARLS, all respondents were asked "Are you often troubled with any body pains?" In the 2013 wave, the question was modified to "Yesterday, did you feel any pain?" Respondents who answered "No" were classified as living without pain in that survey wave (Pain = 0); otherwise, they were recorded as suffering from pain in the corresponding wave (Pain = 1).

Urban-rural classification. The urban-rural classification was determined by a combination of *hukou* status and place of residence at the baseline wave. Among the rural *hukou* population, respondents living in rural regions were classified as rural population, and those dwelling in urban areas were grouped to semi-urban population. Given the small number of urban *hukou* holders living in rural areas (N = 285, percentage = 1.7%) and our primary focus on the impact of *hukou* as an institutional factor on pain, no differentiation was made concerning urban or rural residence among the urban *hukou* holders. Therefore, all respondents with an urban *hukou* were classified as urban population, regardless of whether they lived in urban or rural areas. Hence, we obtained a key predictor with three categories: rural population, semi-urban population, and urban population.

Aging process. The years after the baseline wave is another key predictor in our study. We created a measure of continuous variable called aging process, which was obtained by subtracting the baseline from the corresponding survey year.

Covariates. Considering that rural-to-urban migration, pain prevalence and severity are influenced by sociodemographic characteristics and health conditions (Hao and Tang 2018; Lu and Qin 2014; Rubin and Zimmer 2015; Zajacova et al. 2021b), we adjusted for potential confounders using the baseline wave data. Specifically, for demographic characteristics, we included gender (female = 0; male = 1), age groups in 2011 (45-49 = 0; 50-54 = 1; 55-59 = 2; 60-64 = 3; 65–69 = 4; 70+ = 5), and marital status, which was categorized into married/cohabiting, and single (separated, divorced, widowed, and never married) as the reference group. We controlled for education level (illiterate = 0; less than elementary school = 1; up to elementary school = 2; middle school = 3; high School or beyond = 4). In addition, we also adjusted for health conditions including arthritis, hypertension, diabetes, and dyslipidemia. These conditions were assessed in CHARLS by the following question using a dichotomous measure: "Have you been diagnosed with [the condition] by a doctor?"

Table 1. List of the used variables and their codes in our study

Variable	Code
Outcome	
Pain prevalence	Not suffering from pain = 0; Suffering from pain = 1.
Key predictors	
Urban-rural classification	Rural population; Semi-urban population; Urban population.
Aging process	Year of survey wave - 2011
Covariates	
Gender	Female = 0; Male = 1.
Age groups in baseline wave	45-49 = 0; 50-54 = 1; 55-59 = 2; 60-64 = 3; 65–69 = 4; 70+ = 5
Marital status in baseline wave	Single (separated, divorced, widowed, and never married) = 0; Married/cohabiting = 1.

Education level	illiterate = 0;
	less than elementary school = 1;
	up to elementary school = 2;
	middle school = 3;
	high School or beyond = 4.

Health conditions in baseline wave	
Arthritis	Not having arthritis = 0; Having arthritis = 1; Missing = 2.
Hypertension	Not having hypertension = 0; Having hypertension = 1; Missing = 2.
Diabetes	Not having diabetes = 0; Having diabetes = 1; Missing = 2.
Dyslipidemia	Not having dyslipidemia = 0; Having dyslipidemia = 1; Missing = 2.

3.3 Analytic strategy

3.3.1 Descriptive Analysis

First, descriptive statistics for sample characteristics, stratified by urban-rural classification, were computed. Chi-square tests were employed to assess the statistical significance of covariate differences among urban-rural populations and to evaluate outcome variations across urban-rural groups for every survey wave. Then, in order to preliminarily assess how the pain prevalence varies with the aging process, this study generated line graphs depicting the average pain prevalence over the aging process across different socio-demographic groups.

Since the descriptive analysis above only captures population-averaged trends, to further explore the heterogeneity in these trajectories—assuming that pain-aging trajectories may vary across individuals—we randomly selected 30 participants and plotted their individual pain trajectories across the five survey rounds.

3.3.2 Multivariate Analysis

Then we adjusted for the covariates in the regression analysis to assess the urban-rural and socio-economic disparities in pain prevalence. Due to the longitudinal nature of the data that repeated outcome measurements of the same set of subjects over time and could result in potential correlations between repeated measures nested in the same subject, the generalized estimating equation (GEE) and the generalized linear mixed models (GLMMs) were conducted to account for within-subject correlations.

3.3.2.1 Model setting for the GEE logit model

An unstructured covariance matrix was applied in all GEE models. A logit link function was used to deal with binary outcome, and all results were presented as odds ratios with 95% confidence intervals. In addition, this study focused on the average effect of the aging on the outcome rather than the differences within subjects, thus the grand mean centered value of the aging process was applied. We estimated the GEE logit models using *geepack* package in R. The full model including the interaction term can be expressed as:

$$\text{logit (odds)} = log(\frac{\pi_{ti}}{1-\pi_{ti}}) = \beta_0 + \beta_1 Years_{ti} + \beta_2 X_i + \beta_3 Years_{ti} X_i + \sum_{n=4}^{n} \beta_n C_i$$

(π_{ti} refers to the probability of pain = 1, at time point t for individual i.; $Years_{ti}$ refers to aging process; X_i is Urban-rural classification, and C_i indicates the covariates.)

3.3.2.2 *Model setting for the GLMMs*

We then utilized a multilevel format for the GLMMs to assess the multivariate associations between pain prevalence and the predictors, as well as depict the pain-aging trajectory and its variation across different urban-rural and socio-economic groups using the five-wave unbalanced panel data. We also use a logit format as the link function to fit the probability that an individual was suffering from pain in the specific wave (Ward and Ahlquist 2018). We estimated the GEE logit models using *lme4* package in R.

Model 0*: check the necessity for random intercept.*

In this research, we first start from the unconditional means model (null model, or empty model), to estimate how much of the total variation comes from the variations at the individual level (level-2 or within the individual level), as a basis for subsequent model constructions. The expression for the unconditional means model is as follows:

Level 1:

$$\text{logit (odds)} = log(\frac{\pi_{ti}}{1-\pi_{ti}}) = \beta_{0i}$$

Level 2:

$$\beta_{0i} = \gamma_{00} + \delta_{0i}$$

Thus, the combined format:

$$\text{logit (odds)} = log(\frac{\pi_{ti}}{1-\pi_{ti}}) = \gamma_{00} + \delta_{0i}$$

(π_{ti} refers to the probability of pain = 1, at time point t for individual i.; β_{0i} is the random intercept; γ_{00} is the fixed intercept; and δ_{0i} is the deviation of the individual-specific intercept from the fixed intercept, that is the within-individual residual)

The interclass correlation case:

$$\text{ICC} = \frac{\tau_{00}}{\tau_{00} + \frac{\tau^2}{3}}$$

(τ_{00} is the variance of random intercept, and for the binomial distribution model, the residual variance is usually assumed to be $\frac{\tau^2}{3}$).

***Model 1**: Adding our predictors and covariates to answer question 1 (Does the chronic pain prevalence vary across urban and rural populations?)*

Based on the empty model, we then add urban-rural classification and covariates into the random intercept part. For model simplicity, we assume that our predictors and covariates are all time-invariant variables.

Level 1:

$$\text{logit (odds)} = log(\frac{\pi_{ti}}{1-\pi_{ti}}) = \beta_{0i}$$

Level 2:

$$\beta_{0i} = \gamma_{00} + \gamma_{01} X_i + \sum_{n=2}^{n} \gamma_{0n} C_i + \delta_{0i}$$

Thus, the combined format:

$$\text{logit (odds)} = log\left(\frac{\pi_{ti}}{1-\pi_{ti}}\right) = \gamma_{00} + \gamma_{01} X_i + \sum_{n=2}^{n} \gamma_{0n} C_i + \delta_{0i}$$

(X_i is the key predictors: Urban-rural classification; C_i refers to the covariates: Gender, Age group in the baseline wave, marital status in the baseline wave, educational level, and health conditions in the baseline wave).

***Model 2**: add aging process to Model 1*

Based on the above model, we further add the time indicator variable *aging process* in level-1 to examine the fixed effect of aging process and its random effect, and get the growth model:

Level 1:

$$\text{logit (odds)} = log(\frac{\pi_{ti}}{1-\pi_{ti}}) = \beta_{0i} + \beta_{1i}\ \text{Years}_{ti}$$

Level 2:

$$\beta_{0i} = \gamma_{00} + \gamma_{01} X_i + \sum_{n=2}^{n} \gamma_{0n} C_i + \delta_{0i}$$

$$\beta_{1i} = \gamma_{10} + \delta_{1i}$$

Thus, the combined format:

$$\text{logit (odds)} = log\left(\frac{\pi_{ti}}{1-\pi_{ti}}\right) = \gamma_{00} + \gamma_{01} X_i + \sum_{n=2}^{n} \gamma_{0n} C_i + \delta_{0i} + \gamma_{10}\text{Years}_{ti} + \delta_{1i}\ \text{Years}_{ti}$$

(The β_{1i} refers to the random slope for the pain-aging trajectory, meaning that the effect of aging on pain prevalence could vary across individuals; and δ_{1i} is the deviation of the individual-specific slope from the fixed slope γ_{10}; $Years_{ti}$ refers to aging process).

By setting δ_{1i} equal to zero or not, we could have two different type of unconditional growth models, with δ_{1i} is a model with random slope, otherwise without random slope. Then we use

likelihood ratio test to compare the above model and decide whether we need to add random slope in our MLM analysis or not.

***Model 3**: Adding the interaction term to answer question 2 (Whether these disparities in chronic pain prevalence evolve across the life course?)*

Finally, we add the interaction term in the above model:

Level 1:

$$\text{logit (odds)} = log(\frac{\pi_{ti}}{1-\pi_{ti}}) = \beta_{0i} + \beta_{1i}\,\text{Years}_{ti}$$

Level 2:

$$\beta_{0i} = \gamma_{00} + \gamma_{01}\,X_i + \sum_{n=2}^{n}\gamma_{0n}\,C_i + \delta_{0i}$$

$$\beta_{1i} = \gamma_{10} + \gamma_{11}\,X_i + \delta_{1i}$$

Thus, the combined format:

$$\text{logit (odds)} = log\left(\frac{\pi_{ti}}{1-\pi_{ti}}\right) = \gamma_{00} + \gamma_{01}\,X_i + \sum_{n=2}^{n}\gamma_{0n}\,C_i + \delta_{0i} + \gamma_{10}\text{Years}_{ti} + \gamma_{11}\text{Years}_{ti}\,X_i + \delta_{1i}\,\text{Years}_{ti}$$

(X_i is the key predictors: Urban-rural classification; C_i refers to the covariates: Gender, Age group in the baseline wave, marital status in the baseline wave, educational level, and health conditions in the baseline wave).

3.3.3 Robustness Check

Due to inconsistencies in outcome measurement during the 2013 wave, we excluded this wave and analyzed the above models for a sensitivity check. Additionally, our sample consists of middle-aged and elderly individuals, and some unhealthy elderly participants may pass away during the survey between 2011 and 2020. Since mortality risk varies across individuals based on urban-rural classification and pain levels, restricting the analysis sample to surviving elderly adults may introduce collider bias. This could, in turn, lead to a biased estimate of the relationship between urban-rural classification and pain prevalence.

To address the potential mortality selection bias, we also conducted a robustness check by analyzing a sub-sample of participants aged 45 to 69 at baseline wave. This age group was below the average life expectancy of 77.4 years old in 2019 for Chinese (World Health Organization 2021) and thus less influenced by mortality risk.

Chapter 4 The Results of Analysis

4.1 Descriptive Analysis

Table 2 presents the descriptive statistics for the full analytic sample, stratified by urban-rural classification. There were significant differences in pain prevalence among urban-rural populations across all survey waves. The rural population consistently reported the highest pain prevalence, followed by the semi-urban population, while the urban population exhibited the lowest risk of pain, demonstrating a health gradient in pain prevalence among urban-rural populations. In addition, an extensive escalation of pain prevalence was observed between 2011 and 2020, rising from approximately 33% in the baseline wave to nearly 58% in the final wave, on average. However, the magnitude of this increase varied across urban-rural groups, with the urban population experiencing the largest increase (approximately 29%), followed by the semi-urban population (around 26%), and the rural population showing the smallest increase (about 23%). This pattern suggests that while age-related pain increased across all three populations, the increase rate varied by urban-rural groups, and the disparities in pain prevalence appeared to wane across the life course. This also reminds us that in subsequent modeling, we should take into account that the aging effect on pain prevalence may be differentiated by urban and rural groups. Therefore, the interaction terms need to be included in the multivariate models.

Moreover, χ^2 test results presented in Table 2 indicate significant associations between urban and rural categories and covariates. The data reveal that individuals in urban areas tend to have higher levels of education (χ^2 = 2614.032, $p < 0.001$), and a lower prevalence of arthritis (χ^2 = 138.155, $p < 0.001$). However, the urban population also exhibits a higher risk of cardiovascular disease. Figure 2 illustrates the prevalence of chronic pain across various demographic groups at different stages of aging. The data reveal that rural populations consistently exhibit the highest prevalence rates, while urban populations demonstrate the lowest. From a demographic standpoint, women are more likely to experience chronic pain than men, and individuals who are single face a heightened risk. The likelihood of chronic pain also increases with age. Regarding socioeconomic factors, there is a pronounced educational gradient; individuals with higher levels of education are less likely to suffer from chronic pain. Additionally, an individual's health status is a significant determinant of chronic pain risk. In particular, individuals with arthritis face a substantially higher risk of chronic pain compared

to those without arthritis. The above results show that it is necessary to consider the above confounders in subsequent modeling to obtain unbiased estimates.

Before going to the MLM, we randomly selected 25 individuals and plotted their pain-aging trajectories. Figure 3 shows that the trajectories of individuals differed greatly during the five observation periods. Although the pain trajectory of most individuals showed an upward trend, some individuals showed different trajectories, such as remaining unchanged or showing a V-shaped increase. Therefore, it is necessary to consider the random slope model in subsequent analyses to explore the heterogeneity of aging effects.

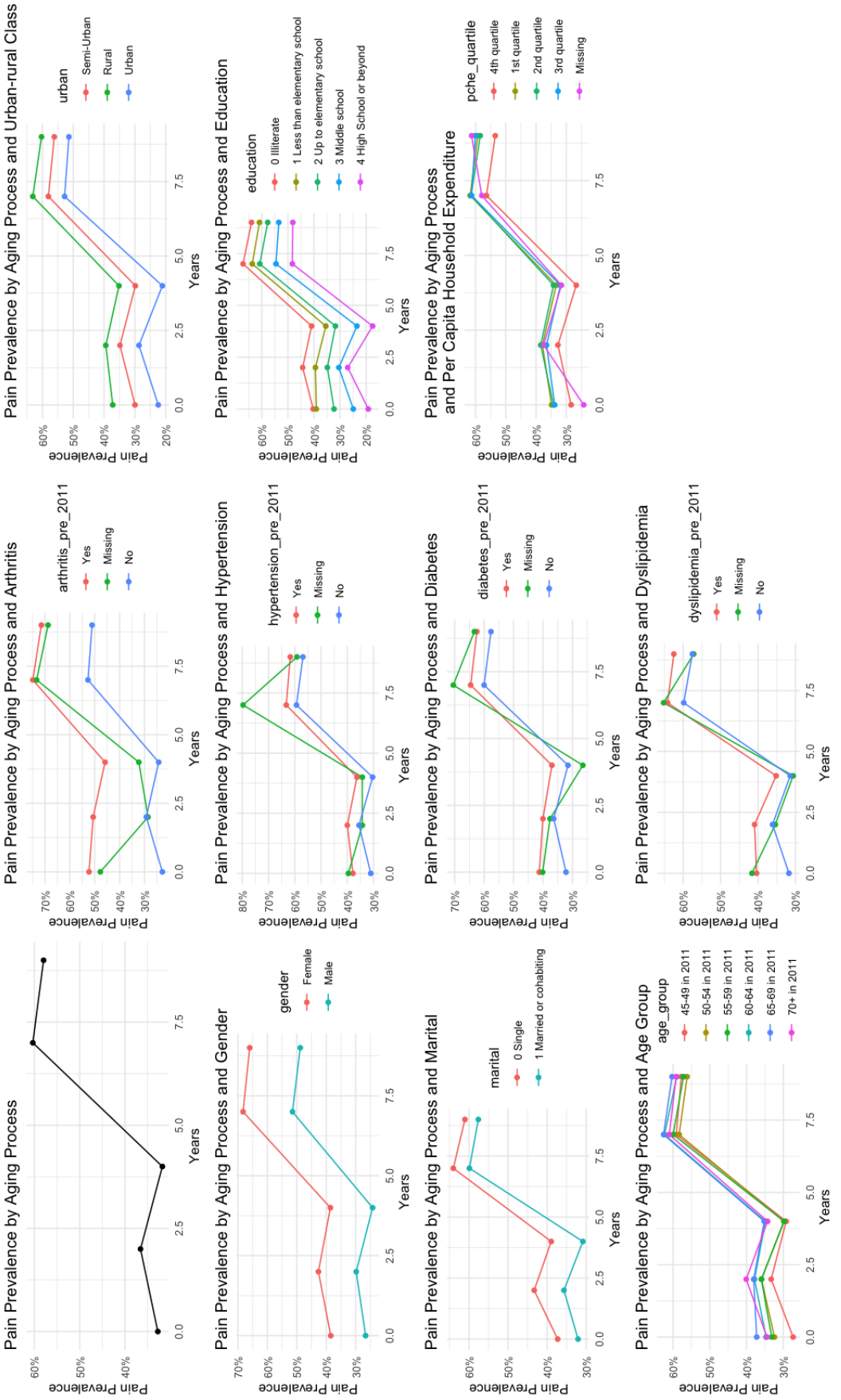

Fig. 2. Pain-aging trajectories across different populations.

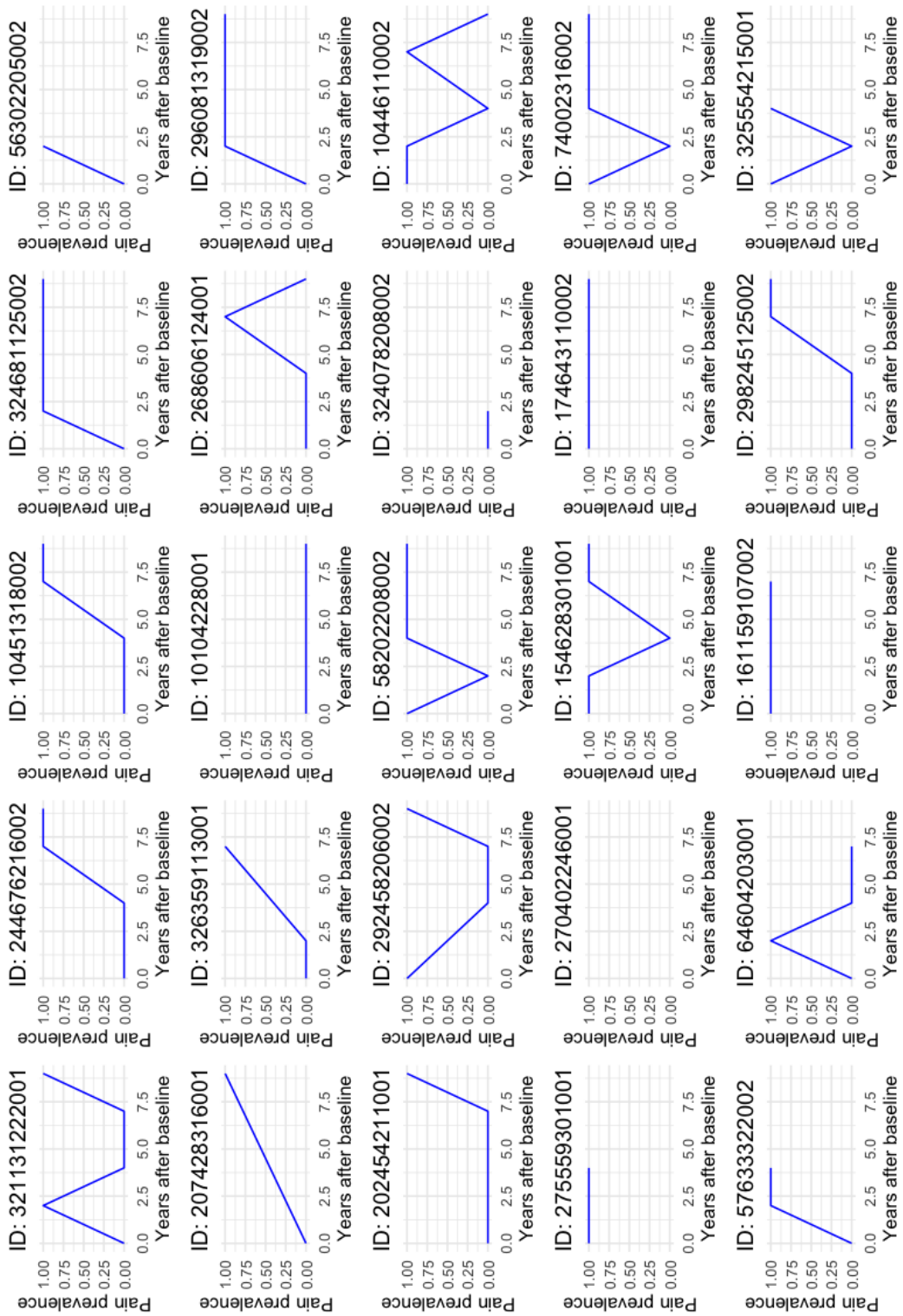

Fig. 3. Pain-aging trajectories for randomly chosen 25 individuals

Table 2. Descriptive statistics for sample characteristics, stratified by urban-rural classification

	Urban-rural classification				
	Rural	Semi-urban	Urban	Total	χ^2
Number of individuals at baseline	9801 (59.476)	3381 (20,517)	3297 (20,007)	16479 (100.000)	
Covariates at baseline					
Gender					15.921 ***
Female	5048 (51.505)	1830 (54.126)	1624 (49.257)	8502 (51.593)	
Male	4753 (48.495)	1551 (45.874)	1673 (50.743)	7977 (48.407)	
Marital Status					3.578
Single	1311 (13.376)	417 (12.334)	409 (12.405)	2137 (12.968)	
Married or cohabiting	8490 (86.624)	2964 (87.666)	2888 (87.595)	14342 (87.032)	
Age Group					49.505 ***
45 - 49	1926 (19.651)	782 (23.129)	660 (20.018)	3368 (20.438)	
50 - 54	1471 (15.009)	549 (16.238)	465 (14.104)	2485 (15.080)	
55 -59	2042 (20.835)	703 (20.793)	664 (20.140)	3409 (20.687)	
60-64	1676 (17.100)	563 (16.652)	555 (16.833)	2794 (16.955)	
65-69	1092 (11.142)	300 (8.873)	351 (10.646)	1743 (10.577)	
70+	1594 (16.264)	484 (14.315)	602 (18.259)	2680 (16.263)	
Education					2614.032 ***
Illiterate	3422 (34.915)	860 (25.436)	326 (9.888)	4608 (27.963)	
Less than elementary school	1951 (19.906)	709 (20.970)	332 (10.070)	2992 (18.156)	
Up to elementary school	2154 (21.977)	824 (24.371)	587 (17.804)	3565 (21.634)	
Middle school	1725 (17.600)	733 (21.680)	948 (28.753)	3406 (20.669)	
High school or beyond	549 (5.601)	255 (7.542)	1104 (33.485)	1908 (11.578)	
Arthritis					138.155 ***
Yes	3594 (36.670)	1072 (31.707)	853 (25.902)	5520 (33.497)	
No	6177 (63.024)	2300 (68.027)	2438 (73.946)	10915 (66.236)	
Missing	30 (0.306)	9 (0.266)	5 (0.152)	44 (0.267)	
Hypertension					126.249 ***
Yes	2191 (22.355)	776 (22.952)	1045 (31.695)	4012 (24.346)	
No	7543 (76.962)	2585 (76.457)	2243 (68.032)	12371 (75.071)	
Missing	67 (0.684)	20 (0.592)	9 (0.273)	96 (0.583)	
Diabetes					371.746 ***
Yes	397 (4.051)	194 (5.738)	325 (9.857)	916 (5.559)	
No	9299 (94.878)	3155 (93.316)	2952 (89.536)	15406 (93.489)	
Missing	105 (1.071)	32 (0.946)	20 (0.607)	157 (0.953)	
Dyslipidemia					163.042 ***
Yes	639 (6.520)	267 (7.897)	571 (17.319)	1477 (8.963)	
No	8929 (91.103)	3042 (89.973)	2690 (81.589)	14661 (88.968)	
Missing	233 (2.377)	72 (2.130)	36 (1.092)	341 (2.069)	
Pain prevalence					
Pain in 2011 (N = 16479)					259.202 ***
Pain = No	6157 (62.820)	2368 (70.038)	2588 (77.586)	11083 (67.255)	
Pain = Yes	3644 (37.180)	1013 (29.962)	739 (22.414)	5396 (32.745)	
Pain in 2013 (N = 12997)					95.014 ***
Pain = No	4882 (60.571)	1685 (65.159)	1677 (71.331)	8244 (63.430)	
Pain = Yes	3178 (39.429)	901 (34.841)	674 (28.669)	4753 (36.570)	
Pain in 2015 (N = 12841)					160.813 ***
Pain = No	5275 (64.843)	1794 (70.160)	1696 (78.920)	8765 (68.258)	
Pain = Yes	2860 (35.157)	763 (29.840)	453 (21.080)	4076 (31.742)	
Pain in 2018 (N = 12716)					82.257 ***
Pain = No	2955 (36.887)	1074 (41.937)	1012 (47.201)	5041 (39.643)	
Pain = Yes	5056 (63.113)	1487 (58.063)	1132 (52.799)	7675 (60.357)	
Pain in 2020 (N = 12291)					57.762 ***
Pain = No	3085 (39.709)	1088 (43.853)	992 (48.604)	5165 (42.023)	
Pain = Yes	4684 (60.291)	1393 (56.147)	1049 (51.396)	7126 (57.977)	

Note. † $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. The Chi-square tests were computed to test the associations between pain prevalence and urban-rural classification, and associations between covariates and urban-rural classification. Percentages in parentheses indicate the column ones.

4.2 The Results from the GEE and GLMMs

4.2.1 Multivariate Analysis from the GEE Logit Models

Table 3 presents the estimates from the GEE models on the associations between pain prevalence and urban-rural populations for the full analytic sample. Consistent with the earlier bivariate patterns, after adjusting for covariates, Model 4 indicates that pain prevalence varied significantly by urban-rural groups. Compared to semi-urban population, rural population exhibited a nearly 21% higher risk of suffering from pain (OR = 1.208, CI = 1.143-1.278), showing that the that urban dwelling is associated with a lower risk of suffering from pain among the rural *hukou* populations. Moreover, urban population was more likely to be free from pain than semi-urban population, with a nearly 15% lower odds of developing it (OR = 0.847; CI = 0.786-0.913).

Although we also observed the elevated pain prevalence during aging process, the magnitude of these rising pain-aging trajectories varied across urban-rural populations. For semi-urban population, the odds of experiencing pain increased by about 17% for each year of aging (OR = 1.166; CI = 1.154-1.178). With each additional year of aging, the likelihood of experiencing pain decreased by 1.5% in the rural population compared to the semi-urban population (OR = 0.985; CI = 0.973-0.997), indicating a narrowing disparity in pain prevalence between the two groups as age increases. Conversely, the urban population showed a more pronounced increase in pain prevalence with age compared to the semi-urban population, with a 2.3% higher rate of increase (OR = 1.023; CI = 1.007-1.039).

The above pattern is graphically presented in Fig. 4. Although the rural population exhibits the highest pain prevalence, the increase in pain with age is the smallest. In contrast, the urban population has a lower initial prevalence, but the risk escalates more rapidly with age. The disparities in pain prevalence among the above three groups become smaller across life course, rather than diverging.

Table 3. Generalized estimating equations for pain prevalence (Full sample)

	Model 1 (OR, 95% CI)	Model 2 (OR, 95% CI)	Model 3 (OR, 95% CI)	Model 4 (OR, 95% CI)
Intercept	0.867 (0.790-0.952) **	0.940 (0.843-1.048)	2.922 (2.501-3.414) ***	2.930 (2.507-3.424) ***
Gender (ref = Female)				
Male	0.539 (0.516-0.563) ***	0.578 (0.551-0.606) ***	0.610 (0.582-0.639) ***	0.609 (0.581-0.638) ***
Marital Status (ref = Single)				
Married or cohabiting	0.899 (0.838-0.965) **	0.908 (0.846-0.974) **	0.922 (0.859-0.989) *	0.921 (0.859-0.989) *
Age Group in 2011 (ref = 45 - 49)				
50 - 54	1.083 (1.006-1.165) *	1.065 (0.989-1.146) †	1.015 (0.943-1.092)	1.014 (0.943-1.091)
55 -59	1.129 (1.056-1.208) ***	1.040 (0.971-1.115)	0.963 (0.899-1.032)	0.963 (0.898-1.032)
60-64	1.252 (1.167-1.344) ***	1.117 (1.037-1.202) **	0.988 (0.919-1.063)	0.988 (0.918-1.063)
65-69	1.310 (1.207-1.422) ***	1.182 (1.086-1.286) ***	1.023 (0.940-1.114)	1.022 (0.939-1.113)
70+	1.266 (1.173-1.366) ***	1.101 (1.014-1.196) *	1.006 (0.926-1.093)	1.007 (0.926-1.094)
Education (ref = Illiterate)				
Less than elementary school		1.024 (0.958-1.095)	0.995 (0.931-1.063)	0.996 (0.933-1.065)
Up to elementary school		0.910 (0.852-0.972) **	0.914 (0.856-0.976) **	0.916 (0.858-0.978) **
Middle school		0.750 (0.699-0.806) ***	0.772 (0.719-0.829) ***	0.774 (0.720-0.831) ***
High school or beyond		0.657 (0.599-0.721) ***	0.687 (0.626-0.753) ***	0.686 (0.625-0.753) ***
Arthritis (ref = Yes)				
No			0.393 (0.376-0.412) ***	0.394 (0.376-0.412) ***
Missing			0.722 (0.508-1.027) †	0.721 (0.507-1.025) †
Hypertension (ref = Yes)				
No			0.890 (0.844-0.938) ***	0.890 (0.844-0.938) ***
Missing			0.888 (0.660-1.195)	0.887 (0.660-1.193)
Diabetes (ref = Yes)				
No			0.829 (0.751-0.915) ***	0.827 (0.749-0.913) ***
Missing			0.878 (0.674-1.145)	0.877 (0.673-1.144)
Dyslipidemia (ref = Yes)				
No			0.779 (0.721-0.843) ***	0.778 (0.720-0.841) ***
Missing			0.734 (0.609-0.885) **	0.733 (0.608-0.883) **
Urban-rural classification (ref = Semi-urban)				
Rural	1.263 (1.195-1.335) ***	1.240 (1.173-1.312) ***	1.208 (1.143-1.278) ***	1.208 (1.143-1.278) ***
Urban	0.747 (0.697-0.801) ***	0.857 (0.795-0.923) ***	0.850 (0.789-0.916) ***	0.847 (0.786-0.913) ***
Aging process	1.150 (1.144-1.155) ***	1.150 (1.145-1.156) ***	1.159 (1.154-1.165) ***	1.166 (1.154-1.178) ***
Aging process * Urban-rural classification (ref = Semi-urban)				
Aging process * Rural				0.985 (0.973-0.997) *
Aging process * Urban				1.023 (1.007-1.039) **
Observations	67324	67324	67324	67324
Number of individuals	16479	16479	16479	16479
QIC	86557	86298	83320	83292

Note. † p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001.

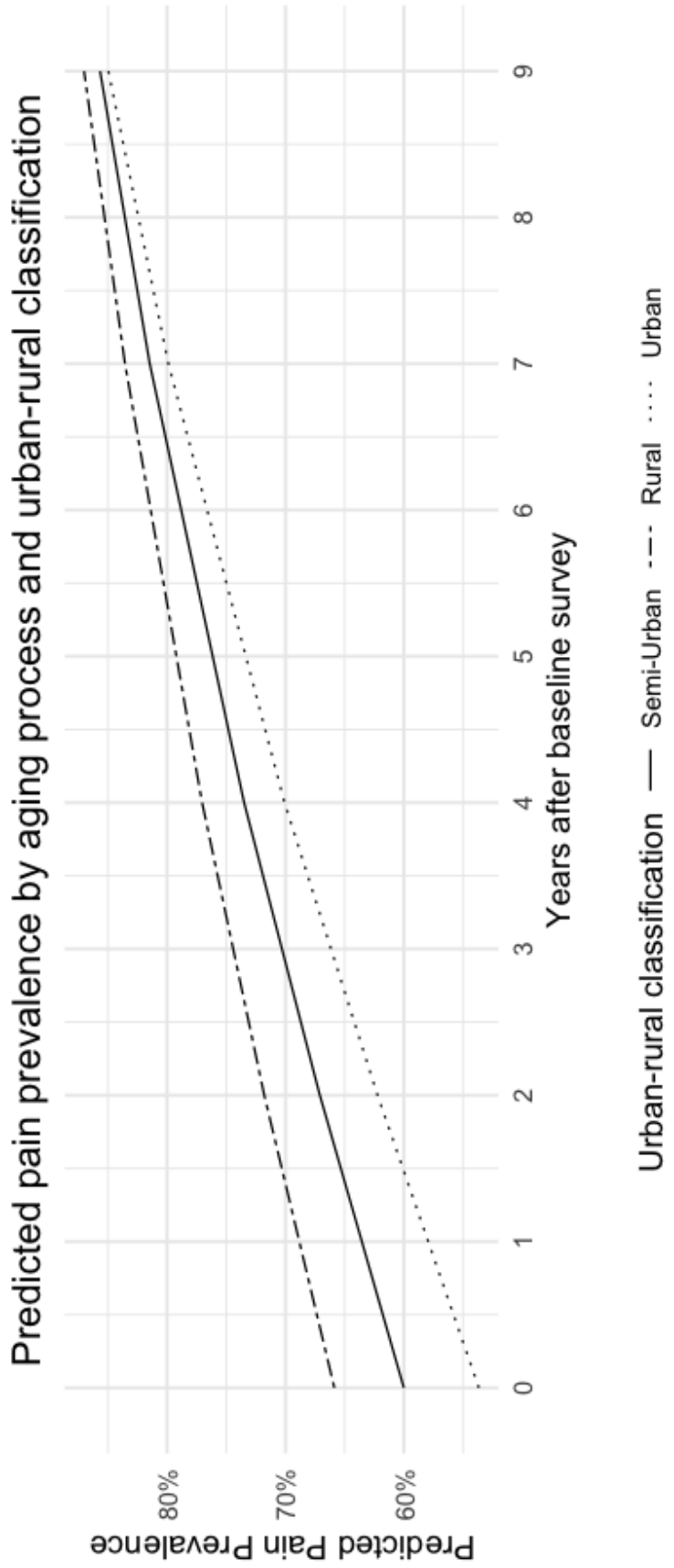

Fig. 4. Predicted pain-aging trajectories by urban-rural classification

4.2.2 Multivariate Analysis from Generalized Linear Mixed Models

Based on the results from the null model, the ICC is around 0.314, indicating that that differences between individuals (Level-2 units) explain 31.4% of the total variation. In other words, 31.4% of the variation is due to differences between individuals, not to variation within individuals. This shows that individual characteristics play an important role in predicting pain prevalence. Hence, it is reasonable and necessary to use a mixed effects model to consider the random effects between individuals.

Table 4. shows the multivariate results from the multilevel logit models for chronic pain prevalence. The results are similar to those of the GEE model. The presence of a significant urban-rural gradient in the prevalence of chronic pain was consistently observed across studies. However, it is important to highlight that in the analysis of interaction terms, differences were only identified in the trajectories of chronic pain between urban and rural populations. No significant differences were observed in pain-aging trajectories between rural and semi-urban populations, a result that is inconsistent with the conclusion of the GEE model.

Table 4. Generalized Linear Mixed Models for pain prevalence (Full sample)

	Model 1 (OR, 95% CI)	Model 2 (OR, 95% CI)	Model 3 (OR, 95% CI)
Fixed effect			
Intercept	3.220 (2.713-3.820) ***	3.864 (3.203-4.466) ***	3.762 (3.115-4.544) ***
Gender (ref = Female)			
Male	0.554 (0.525-0.584) ***	0.535 (0.505-0.567) ***	0.544 (0.513-0.577) ***
Marital Status (ref = Single)			
Married or cohabiting	0.941 (0.869-1.018)	0.928 (0.852-1.011) †	0.923 (0.847-1.007) †
Age Group in 2011 (ref = 45 - 49)			
50 - 54	1.057 (0.974-1.147)	1.055 (0.964-1.154)	1.026 (0.937-1.123)
55 -59	0.994 (0.921-1.073)	0.994 (0.913-1.081)	0.968 (0.889-1.053)
60-64	1.026 (0.945-1.114)	1.032 (0.943-1.130)	0.995 (0.909-1.090)
65-69	1.046 (0.951-1.149)	1.071 (0.965-1.188)	1.044 (0.941-1.159)
70+	0.941 (0.857-1.033)	1.063 (0.959-1.177)	1.038 (0.937-1.150)
Education (ref = Illiterate)			
Less than elementary school	1.037 (0.962-1.118)	1.001 (0.923-1.087)	1.006 (0.927-1.092)
Up to elementary school	0.932 (0.866-1.003) †	0.895 (0.826-0.971) **	0.892 (0.823-0.968) **
Middle school	0.767(0.708-0.831) ***	0.719 (0.658-0.785) ***	0.708 (0.648-0.774) ***
High school or beyond	0.668 (0.603-0.740) ***	0.612 (0.547-0.685) ***	0.593 (0.529-0.664) ***
Arthritis (ref = Yes)			
No	0.345 (0.328-0.364) **	0.304 (0.287-0.322) ***	0.308 (0.291-0.327) ***
Missing	0.779 (0.483-1.256)	0.656 (0.389-1.105)	0.746 (0.442-1.259)
Hypertension (ref = Yes)			
No	0.915 (0.862-0.971) ***	0.868 (0.813-0.927) ***	0.880 (0.824-0.940) ***
Missing	1.047 (0.744-1.474)	0.888 (0.660-1.195)	0.873 (0.599-1.272)
Diabetes (ref = Yes)			
No	0.828 (0.741-0.925) ***	0.806 (0.714-0.910) ***	0.823 (0.728-0.929) **
Missing	0.903 (0.673-1.212)	0.863 (0.626-1.192)	0.860 (0.622-1.188)
Dyslipidemia (ref = Yes)			
No	0.689 (0.741-0.925) ***	0.745 (0.675-0.821) ***	0.729 (0.661-0.805) ***
Missing	0.689 (0.558-0.850) ***	0.689 (0.548-0.868) **	0.698 (0.554-0.879) **
Urban-rural classification (ref = Semi-urban)			
Rural	1.285 (1.207-1.367) ***	1.284 (1.200-1.375) ***	1.320 (1.233-1.414) ***
Urban	0.781 (0.719-0.849) ***	0.807 (0.737-0.884) ***	0.855 (0.779-0.937) ***
Aging process		1.207 (1.200-1.215) ***	1.210 (1.194-1.226) ***
Aging process * Urban-rural classification (ref = Semi-urban)			
Aging process * Rural			0.990 (0.975-1.005)
Aging process * Urban			1.033 (1.013-1.054) **
Random effect			
Intercept	1.050	1.384	1.398
Slope	/	0.001	0.003
Observations	67324	67324	67324
Number of individuals	16479	16479	16479
AIC	83887	79567	79552
BIC	84096	79804	79808

Note. † p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001.

Chapter 5 Conclusion and Discussion

5.1 The Conclusion and Discussion for the Empirical Results

Based on nationally representative longitudinal data, our study contributes new knowledge on urban-rural disparities in pain prevalence and how they evolve across life course among middle-aged and older Chinese for the first time. The findings reveal a pronounced urban-rural gradient in pain prevalence, with the highest prevalence observed among individuals with rural *hukou* dwelling in rural areas, followed by rural-to-urban migrants with rural *hukou*, and finally, the urban *hukou* population, who experienced the lowest pain prevalence. Our results indicate that both urban residence and urban *hukou* are associated with lower pain prevalence. This finding provides evidence supporting that urban geographic and institutional characteristics act as health advantage resources shaping pain inequalities, consistent with prior research (Dorélien and Xu 2020; Song and Smith 2019). Notably, the observed heightened vulnerability to pain risk among individuals with rural *hukou*, regardless of their place of residence, compared to those with urban *hukou*, demonstrating that urban *hukou* is closer to advantages in health outcomes, and has a greater impact than geographic characteristics in shaping health conditions. This pattern highlights the role of the *hukou* as an unequal and fundamental institutional arrangement, exerts an entrenched influence on producing and maintaining health inequalities. In this sense, urban-rural health inequality shaped by institutional factors deserves more of research attention when research focuses on examining the urban-rural division in China.

In addition, the study indicates that aging is associated with a significantly elevated risk of developing pain, which is in line with previous research (Zajacova et al. 2021a). Notably, our results reveal substantial urban-rural disparities in the pain-aging trajectory. In rural *hukou* groups, although pain prevalence is higher among rural residents compared to rural-to-urban migrants, the disparity in pain prevalence diminishes over the life course. While the urban *hukou* population exhibits a lower initial pain prevalence, they experience the most pronounced increase in prevalence over time. As individuals age and their physical function declines, the differences in pain prevalence between urban and rural populations tend to converge. This finding supports the age-as-leveler hypothesis, as opposed to the cumulative (dis)advantage hypothesis, and contrasts with the conclusions of some prior studies. For instance, research conducted in the United States has found that the socioeconomic status (SES) gap in pain

prevalence widens with age (Zajacova et al. 2021a). We argue that the contrasting pattern observed in our study may be attributable to two key reasons. First, compared to the US, China provides lower levels of social security for middle-aged and elderly populations, including less generous pension benefits, which may result in insufficient support for pain prevention and treatment. Second, our sample includes individuals aged 45 years and older, all of whom were born before 1966. These cohorts endured substantial hardships in their early years, marked by material scarcity, malnutrition, high-intensity physical labor, and limited access to timely medical care. The experiences and lasting negative impact of these early-life conditions may have a more pronounced effect on health outcomes in later life (Fan and Qian 2015; Song et al. 2009; Zhang et al. 2017). Future research needs to consider and explore the influences of early-life conditions on chronic pain in later life. In addition, based on the results from the sensitivity analysis, we contend that the narrowing health gap is unlikely to be attributable to mortality selection. Instead, it may be influenced by other biological factors associated with the aging process. Future research should aim to incorporate a broader range of physiological indicators to more comprehensively test this hypothesis.

This research has several limitations that should be acknowledged. First, the measurement of pain relies on retrospective self-reported data, which is inherently subjective to recall bias and reporting heterogeneity. Additionally, the measurement does not distinguish between chronic and acute pain, nor does it specify the duration of pain, potentially compromising its validity. Second, the use of an overall pain measure, rather than focusing on specific types of pain, limits the ability to explore heterogeneity in pain disparities. The pain prevalence and its sites can vary significantly between urban and rural populations due to differing working- and life-style patterns. For example, lower back and joint pain, often associated with prolonged physical labor, may be more prevalent in rural populations, whereas conditions such as sciatica, linked to sedentary behaviors, may be more common in urban groups. Furthermore, age-related pain patterns are not uniform across all pain sites. Certain types of pain are more closely associated with physiological decline regardless of socioeconomic status, while others may be more influenced by non-physiological factors. Future research should aim to refine measurement approaches and focus on more specific indicators to assess pain inequalities in middle-aged and older Chinese.

5.2 A Conclusion and Discussion for the GEE and GLMMs modelling

As more longitudinal data become available to researchers, the challenge is how to effectively extract and utilize the information related to change and development hidden within these data. This study applied the GEE logit model and GLMM logit model to model the urban-rural disparities in pain prevalence and its dynamic characteristics within a Chinese context. Despite the general consistency in the conclusions of the empirical studies, substantial differences remain between the two models in practical application.

When comparing the GEE Logit model and the GLMM (using a multilevel framework), each model has distinct characteristics. The GEE Logit model is designed to handle correlated data by modeling the correlation between observations through a working correlation matrix. Its parameters directly reflect marginal effects, making it suitable for interpretation at the population level. The strength of the GEE model lies in its robustness to misspecifications of the error structure as we discussed in Chapter 2 and its relative computational efficiency. However, it may be less suitable and less predictive than GLMM in dealing with complex hierarchical data structures.

GLMM, on the other hand, incorporates random effects to handle the hierarchical structure within the data, allowing for better modeling of variability between individuals. Its flexibility enables it to manage complex nested data and excel in individual-level predictions. However, GLMM is computationally more complex and more sensitive to assumptions regarding random effects and error structures, which can present challenges, especially with large datasets or intricate model structures.

In all, GEE models and GLMM models each have their own application scenarios and assumptions. GEE is more suitable for use when focusing on overall marginal effects and uncertainty about the correlation structure, while GLMM is more suitable for scenarios that need to capture individual differences and accurately model random effects. The choice of model should be based on the actual needs of the research problem, the data structure, and the degree of dependence on assumptions.

Reference

Blyth, F. M., Briggs, A. M., Schneider, C. H., Hoy, D. G., and March, L. M. (2019), "The global burden of musculoskeletal pain—where to from here?," *American Journal of Public Health*, 109, 35–40. https://doi.org/10.2105/AJPH.2018.304747.

Bryk, A. S., and Raudenbush, S. W. (1992), *Hierarchical linear models: applications and data analysis methods.*, Sage Publications, Inc.

Cohen, S. P., Vase, L., and Hooten, W. M. (2021), "Chronic pain: an update on burden, best practices, and new advances," *The Lancet*, 397, 2082–2097. https://doi.org/10.1016/S0140-6736(21)00393-7.

Crofford, L. J. (2015), "Chronic pain: where the body meets the brain," *Transactions of the American Clinical and Climatological Association*, American Clinical and Climatological Association, 126, 167.

Ding, P. (2024), "Linear Model and Extensions," *arXiv preprint arXiv:2401.00649*.

Dorélien, A., and Xu, H. (2020), "Estimating rural–urban disparities in self-rated health in China: Impact of choice of urban definition," *Demographic Research*, 43, 1429–1460. https://doi.org/10.4054/DemRes.2020.43.49.

Fan, W., and Qian, Y. (2015), "Long-term health and socioeconomic consequences of early-life exposure to the 1959–1961 Chinese Famine," *Social Science Research*, 49, 53–69. https://doi.org/10.1016/j.ssresearch.2014.07.007.

Fitzmaurice, G. M., Laird, N. M., and Ware, J. H. (2012), *Applied longitudinal analysis*, John Wiley & Sons.

Fu, Q., Wu, C., Liu, H., Shi, Z., and Gu, J. (2018), "Live like mosquitoes: *Hukou*, rural–urban disparity, and depression," *Chinese Journal of Sociology*, 4, 56–78. https://doi.org/10.1177/2057150X17748313.

Gaskin, D. J., and Richard, P. (2012), "The Economic Costs of Pain in the United States," *The Journal of Pain*, 13, 715–724. https://doi.org/10.1016/j.jpain.2012.03.009.

Goosby, B. J. (2013), "Early Life Course Pathways of Adult Depression and Chronic Pain," *Journal of Health and Social Behavior*, 54, 75–91. https://doi.org/10.1177/0022146512475089.

Goubert, L., and Trompetter, H. (2017), "Towards a science and practice of resilience in the face of pain," *European Journal of Pain*, 21, 1301–1315. https://doi.org/10.1002/ejp.1062.

Grol-Prokopczyk, H. (2017), "Sociodemographic disparities in chronic pain, based on 12-year longitudinal data," *Pain*, 158, 313–322. https://doi.org/10.1097/j.pain.0000000000000762.

Hao, P., and Tang, S. (2018), "Migration destinations in the urban hierarchy in China: Evidence from Jiangsu," *Population, Space and Place*, Wiley Online Library, 24, e2083. https://doi.org/10.1002/psp.2083.

Hardin, J. W., and Hilbe, J. M. (2002), *Generalized estimating equations*, chapman and hall/CRC.

Hoffman, L. (2015), *Longitudinal analysis: Modeling within-person fluctuation and change*, Routledge.

Hox, J., Moerbeek, M., and Van de Schoot, R. (2017), *Multilevel analysis: Techniques and applications*, Routledge.

Jay, M. A., Bendayan, R., Cooper, R., and Muthuri, S. G. (2019), "Lifetime socioeconomic circumstances and chronic pain in later adulthood: findings from a British birth cohort study," *BMJ open*, British Medical Journal Publishing Group, 9, e024250. https://doi.org/10.1136/bmjopen-2018-024250.

Kennedy, J., Roll, J. M., Schraudner, T., Murphy, S., and McPherson, S. (2014), "Prevalence of persistent pain in the US adult population: new data from the 2010 national health interview survey," *The Journal of Pain*, Elsevier, 15, 979–984. https://doi.org/10.1016/j.jpain.2014.05.009.

Liang, K.-Y., and Zeger, S. L. (1986), "Longitudinal data analysis using generalized linear models," *Biometrika*, 73, 13–22. https://doi.org/10.1093/biomet/73.1.13.

Lu, Y., and Qin, L. (2014), "Healthy migrant and salmon bias hypotheses: A study of health and internal migration in China," *Social Science & Medicine*, 102, 41–48. https://doi.org/10.1016/j.socscimed.2013.11.040.

McNamee, P., and Mendolia, S. (2014), "The effect of chronic pain on life satisfaction: Evidence from Australian data," *Social Science & Medicine*, 121, 65–73. https://doi.org/10.1016/j.socscimed.2014.09.019.

Park, E., Cho, M., and Ki, C.-S. (2009), "Correct use of repeated measures analysis of variance," *The Korean journal of laboratory medicine*, The Korean Society for Legal Medicine, 29, 1–9. https://doi.org/10.3343/kjlm.2009.29.1.1.

Peele, M., and Schnittker, J. (2022), "The nexus of physical and psychological pain: consequences for mortality and implications for medical sociology," *Journal of*

Health and Social Behavior, 63, 210–231. https://doi.org/10.1177/00221465211064533.

Rabe-Hesketh, S., and Skrondal, A. (2008), *Multilevel and longitudinal modeling using Stata*, STATA press.

Rubin, S., and Zimmer, Z. (2015), "Pain and self-assessed health: Does the association vary by age?," *Social Science & Medicine*, 130, 259–267. https://doi.org/10.1016/j.socscimed.2015.02.024.

Rustøen, T., Wahl, A. K., Hanestad, B. R., Lerdal, A., Paul, S., and Miaskowski, C. (2005), "Age and the Experience of Chronic Pain: Differences in Health and Quality of Life Among Younger, Middle-Aged, and Older Adults," *The Clinical Journal of Pain*, 21, 513–523. https://doi.org/10.1097/01.ajp.0000146217.31780.ef.

Smith, D., Wilkie, R., Croft, P., Parmar, S., and McBeth, J. (2018), "Pain and mortality: mechanisms for a relationship," *Pain*, 159, 1112–1118. https://doi.org/10.1097/j.pain.0000000000001193.

Song, Q., and Smith, J. P. (2019), "Hukou system, mechanisms, and health stratification across the life course in rural and urban China," *Health & Place*, 58, 102150. https://doi.org/10.1016/j.healthplace.2019.102150.

Song, S., Wang, W., and Hu, P. (2009), "Famine, death, and madness: Schizophrenia in early adulthood after prenatal exposure to the Chinese Great Leap Forward Famine," *Social Science & Medicine*, 68, 1315–1321. https://doi.org/10.1016/j.socscimed.2009.01.027.

Topping, M., and Fletcher, J. (2024), "Educational attainment, family background and the emergence of pain gradients in adulthood," *Social Science & Medicine*, Elsevier, 116692. https://doi.org/10.1016/j.socscimed.2024.116692.

Van Alboom, M., Baert, F., Bernardes, S. F., Bracke, P., and Goubert, L. (2023), "Public chronic pain stigma and the role of pain type and patient Gender: An experimental vignette Study," *The Journal of Pain*, 24, 1798–1812. https://doi.org/10.1016/j.jpain.2023.05.007.

Ward, M. D., and Ahlquist, J. S. (2018), *Maximum likelihood for social science: Strategies for analysis*, Cambridge University Press.

World Health Organization (2021), *World Health Statistics 2021*.

Wu, X. (2019), "Inequality and social stratification in postsocialist China," *Annual Review of Sociology*, 45, 363–382. https://doi.org/10.1146/annurev-soc-073018-022516.

Zajacova, A., Grol-Prokopczyk, H., and Zimmer, Z. (2021a), "Pain trends among American adults, 2002–2018: patterns, disparities, and correlates," *Demography*, Duke University Press, 58, 711–738. https://doi.org/10.1215/00703370-8977691.

Zajacova, A., Grol-Prokopczyk, H., and Zimmer, Z. (2021b), "Sociology of Chronic Pain," *Journal of Health and Social Behavior*, 62, 302–317. https://doi.org/10.1177/00221465211025962.

Zajacova, A., Rogers, R. G., Grodsky, E., and Grol-Prokopczyk, H. (2020), "The relationship between education and pain among adults aged 30–49 in the United States," *The Journal of Pain*, Elsevier, 21, 1270–1280. https://doi.org/10.1016/j.jpain.2020.03.005.

Zeger, S. L., and Liang, K.-Y. (1986), "Longitudinal data analysis for discrete and continuous outcomes," *Biometrics*, JSTOR, 121–130. https://doi.org/10.2307/2531248.

Zhan, P., Ma, X., and Li, S. (2021), "Migration, population aging, and income inequality in China," *Journal of Asian Economics*, 76, 101351. https://doi.org/10.1016/j.asieco.2021.101351.

Zhang, Z., Song, S., and Wu, X. (2017), "Exodus from hunger: The long-term health consequences of the 1959–1961 Chinese famine," *Biodemography and Social Biology*, Taylor & Francis, 63, 148–166. https://doi.org/doi.org/10.1080/19485565.2017.1311203.

Zhao, Y., Hu, Y., Smith, J. P., Strauss, J., and Yang, G. (2014), "Cohort profile: the China health and retirement longitudinal study (CHARLS)," *International Journal of Epidemiology*, 43, 61–68. https://doi.org/10.1093/ije/dys203.

Zimmer, Z., Zajacova, A., and Grol-Prokopczyk, H. (2020), "Trends in pain prevalence among adults aged 50 and older across Europe, 2004 to 2015," *Journal of Aging and Health*, 32, 1419–1432. https://doi.org/10.1177/0898264320931665.

Appendix

A.1 Tables for robustness check

Table A.1.1. Generalized estimating equations for pain prevalence (Exclude 2013 wave)

	Model 1	Model 2	Model 3	Model 4
Intercept	0.921 (0.835-1.015) †	0.993 (0.886-1.114)	3.251 (2.756-3.836) ***	3.263 (2.764-3.851) ***
Aging process	1.151 (1.146-1.157) ***	1.150 (1.145-1.158) ***	1.162 (1.156-1.168) ***	1.170 (1.157-1.184) ***
Urban-rural classification (ref = Semi-urban)				
Rural	1.275 (1.203-1.351) ***	1.240 (1.173-1.328) ***	1.220 (1.150-1.294) ***	1.219 (1.149-1.293) ***
Urban	0.748 (0.695-0.805) ***	0.857 (0.795-0.932) ***	0.856 (0.791-0.926) ***	0.852 (0.787-0.923) ***
Gender (ref = Female)				
Male	0.532 (0.508-0.557) ***	0.578 (0.551-0.598) ***	0.601 (0.572-0.631) ***	0.600 (0.572-0.631) ***
Marital Status (ref = Single)				
Married or cohabiting	0.909 (0.844-0.979) *	0.908 (0.846-0.990) *	0.932 (0.864-1.004) †	0.931 (0.864-1.004) †
Age Group in 2011 (ref = 45 - 49)				
50 - 54	1.070 (0.992-1.115) †	1.065 (0.989-1.139)	1.002 (0.928-1.081)	1.002 (0.928-1.081)
55 -59	1.118 (1.043-1.199) **	1.040 (0.971-1.105)	0.950 (0.884-1.022)	0.950 (0.884-1.021)
60-64	1.254 (1.165-1.351) ***	1.117 (1.037-1.204) **	0.983 (0.910-1.061)	0.982 (0.909-1.060)
65-69	1.320 (1.211-1.439) ***	1.182 (1.086-1.298) ***	1.023 (0.935-1.120)	1.022 (0.934-1.118)
70+	1.236 (1.140-1.340) ***	1.101 (1.014-1.172)	0.977 (0.895-1.067)	0.977 (0.895-1.067)
Education (ref = Illiterate)				
Less than elementary school		1.024 (0.958-1.114)	1.009 (0.940-1.082)	1.011 (0.942-1.084)
Up to elementary school		0.910 (0.852-0.991) *	0.929 (0.867-0.995) *	0.931 (0.869-0.998) *
Middle school		0.750 (0.699-0.807) ***	0.771 (0.716-0.832) ***	0.773 (0.717-0.834) ***
High school or beyond		0.657 (0.599-0.713) ***	0.677 (0.614-0.746) ***	0.676 (0.613-0.745) ***
Arthritis (ref = Yes)				
No			0.385 (0.367-0.403) ***	0.385 (0.367-0.404) ***
Missing			0.816 (0.563-1.181)	0.815 (0.563-1.178)
Hypertension (ref = Yes)				
No			0.882 (0.834-0.933) ***	0.882 (0.833-0.932) ***
Missing			0.920 (0.670-1.265)	0.919 (0.669-1.262)
Diabetes (ref = Yes)				
No			0.810 (0.730-0.900) ***	0.808 (0.727-0.899) ***
Missing			0.859 (0.650-1.134)	0.857 (0.649-1.133)
Dyslipidemia (ref = Yes)				
No			0.776 (0.714-0.843) ***	0.775 (0.713-0.842) ***
Missing			0.746 (0.612-0.911) **	0.744 (0.610-0.908) **
Aging process * Urban-rural classification (ref = Semi-urban)				
Aging process * Rural				0.982 (0.969-0.995) **
Aging process * Urban				1.027 (1.009-1.044) **
Observations	54327	54327	54327	54327
Number of individuals	16479	16479	16479	16479
QIC	69843	69609	67097	67061

Note. † $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table A.1.2. Generalized estimating equations for pain prevalence (Exclude age group 70+ in 2011)

	Model 1	Model 2	Model 3	Model 4
Intercept	0.911 (0.818-1.015) †	0.965 (0.852-1.092)	3.026 (2.548-3.595) ***	3.032 (2.552-3.603) ***
Aging process	1.151 (1.146-1.157) ***	1.152 (1.146-1.158) ***	1.161 (1.155-1.167) ***	1.167 (1.154-1.180) ***
Urban-rural classification (ref = Semi-urban)				
Rural	1.272 (1.198-1.350) ***	1.246 (1.172-1.324) ***	1.219 (1.147-1.294) ***	1.219 (1.148-1.295) ***
Urban	0.762 (0.706-0.823) ***	0.887 (0.818-0.962) **	0.883 (0.814-0.958) **	0.878 (0.809-0.953) **
Gender (ref = Female)				
Male	0.539 (0.514-0.565) ***	0.578 (0.550-0.609) ***	0.607 (0.577-0.638) ***	0.607 (0.577-0.638) ***
Marital Status (ref = Single)				
Married or cohabiting	0.857 (0.785-0.935) ***	0.868 (0.796-0.947) **	0.891 (0.817-0.972) **	0.890 (0.816-0.971) **
Age Group in 2011 (ref = 45 - 49)				
50 - 54	1.082 (1.005-1.164) *	1.066 (0.990-1.148) †	1.016 (0.944-1.094)	1.016 (0.944-1.093)
55 -59	1.127 (1.054-1.206) ***	1.037 (0.967-1.112)	0.960 (0.896-1.030)	0.960 (0.896-1.029)
60-64	1.248 (1.162-1.339) ***	1.108 (1.029-1.194) **	0.981 (0.911-1.056)	0.981 (0.911-1.056)
65-69	1.301 (1.198-1.413) ***	1.170 (1.075-1.274) ***	1.013 (0.930-1.104)	1.012 (0.929-1.103)
Education (ref = Illiterate)				
Less than elementary school		1.054 (0.978-1.135)	1.027 (0.954-1.105)	1.029 (0.956-1.107)
Up to elementary school		0.937 (0.871-1.007) †	0.948 (0.882-1.018)	0.950 (0.884-1.021)
Middle school		0.763 (0.707-0.824) ***	0.789 (0.731-0.851) ***	0.790 (0.733-0.853) ***
High school or beyond		0.651 (0.590-0.718) ***	0.682 (0.618-0.752) ***	0.682 (0.618-0.752) ***
Arthritis (ref = Yes)				
No			0.392 (0.373-0.412) ***	0.392 (0.373-0.412) ***
Missing			0.822 (0.558-1.212)	0.821 (0.557-1.209)
Hypertension (ref = Yes)				
No			0.878 (0.828-0.931) ***	0.878 (0.828-0.931) ***
Missing			0.879 (0.641-1.205)	0.878 (0.640-1.203)
Diabetes (ref = Yes)				
No			0.833 (0.748-0.928) ***	0.832 (0.747-0.927) ***
Missing			0.889 (0.669-1.182)	0.888 (0.668-1.181)
Dyslipidemia (ref = Yes)				
No			0.765 (0.703-0.833) ***	0.764 (0.702-0.832) ***
Missing			0.756 (0.615-0.930) **	0.755 (0.613-0.929) **
Aging process * Urban-rural classification (ref = Semi-urban)				
Aging process * Rural				0.986 (0.973-0.999) *
Aging process * Urban				1.025 (1.008-1.042) **
Observations	58477	58477	58477	58477
Number of individuals	13799	13799	13799	13799
QIC	75109	74858	72243	72220

Note. † $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

A.2 R syntax for statistical modelling

```
# Part A: GEE  ------------------------------------------------------
# Full sample
full_1 <- geeglm(
  pain01 ~ years + urban + years * urban + gender  + marital + age_group +

    education + pche_quartile + arthritis_pre_2011 +
    hypertension_pre_2011 + dyslipidemia_pre_2011 + diabetes_pre_2011,
  id = id,
  data = full_data,
  family = binomial(link = "logit"),
  corstr = "unstructured"
)
summary(full_1)
QIC(full_1)

# OR format
summary_gee_or(full_1)

# Interaction effect plot
slopes_1 <- ggpredict(full_1, terms = c("years", "urban"))
slopes_1$x <- slopes_1$x + 4.11
slopes_1$x

group_levels <- unique(slopes_1$group)

plot_simple_slopes_1 <- ggplot(slopes_1, aes(x = x, y = predicted, linetyp
e = factor(group))) +
  geom_line(size = 0.35, color = "black") +
  scale_y_continuous(labels = percent) +
  labs(x = "Years after baseline survey", y = "Predicted Pain Prevalence",
 linetype = "Urban-rural classification") +
  ggtitle("Predicted pain prevalence by aging process and urban-rural clas
sification") +
  scale_linetype_manual(values = c( "solid", "twodash", "dotted")) +
  scale_x_continuous(breaks = c(0:9)) +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5),
    legend.position = "bottom"
  )

print(plot_simple_slopes_1)


# Sensitivity Analysis-----------------------------------------------
# no 2013 wave ------------------------------------------------------
full_2 <- geeglm(
  pain01 ~ years + urban + years * urban + gender  + marital + age_group +

    education + pche_quartile + arthritis_pre_2011 +
    hypertension_pre_2011 + dyslipidemia_pre_2011 + diabetes_pre_2011,
  id = id,
```

```
  data = no2013_data,
  family = binomial(link = "logit"),
  corstr = "unstructured"
)
summary(full_2)
QIC(full_2)



# no age_group >= 70-----------------------------------------------------
full_3 <- geeglm(
  pain01 ~ years + urban + years * urban + gender  + marital + age_group +

    education + pche_quartile + arthritis_pre_2011 +
    hypertension_pre_2011 + dyslipidemia_pre_2011 + diabetes_pre_2011,
  id = id,
  data = no70_data,
  family = binomial(link = "logit"),
  corstr = "unstructured"
)
summary(full_3)
QIC(full_3)
# ----------------------------------------------------------------------




# ----------------------------------------------------------------------
# Part B: MLM ----------------------------------------------------------
library(lme4)

## Model 0: Null Model
null_model <- glmer(
  pain01 ~ 1 + (1 | id),
  data = full_data,
  family = binomial(link = "logit")
)

summary(null_model)

var_random_intercept <- as.numeric(VarCorr(null_model)$id[1])

#### ICC
icc <- var_random_intercept / (var_random_intercept + (pi^2 / 3))
icc

# ----------------------------------------------------------------
## Model 1: random intercept with predictor and covariates
model_1 <- glmer(
  pain01 ~ urban +
    gender + marital + age_group +
    education +
    arthritis_pre_2011 + hypertension_pre_2011 +
    diabetes_pre_2011 + dyslipidemia_pre_2011 +
```

```
    (1 | id),
  data = full_data,
  family = binomial(link = "logit")
)
summary(model_1)

# lr test
model_1_lmerTest <- lmerTest::lmer(
  pain01 ~ urban +
    gender + marital + age_group +
    education +
    arthritis_pre_2011 + hypertension_pre_2011 +
    diabetes_pre_2011 + dyslipidemia_pre_2011 +
    (1 | id),
  data = full_data,
  family = binomial(link = "logit")
)

## model 2: adding aging process
model_2 <- glmer(
  pain01 ~ years + urban +
    gender + marital + age_group +
    education +
    arthritis_pre_2011 + hypertension_pre_2011 +
    diabetes_pre_2011 + dyslipidemia_pre_2011 +
    (1 + years | id),
  data = full_data,
  family = binomial(link = "logit")
)
summary(model_2)

## model 3: interaction term
model_3 <- glmer(
  pain01 ~ years + urban +
    years * urban +
    gender + marital + age_group +
    education +
    arthritis_pre_2011 + hypertension_pre_2011 +
    diabetes_pre_2011 + dyslipidemia_pre_2011 +
    (1 + years | id),
  data = full_data,
  family = binomial(link = "logit")
)
summary(model_3)
```