Health Shocks, Frailty Dynamics, and Labor Supply in China's Aging Population: Evidence from CHARLS 2011–2018

Quan Zhang

Department of Economics, University of Hawai'i at Mānoa

March 2026

Submitted to NBER Summer Institute 2026

Economics of Health Workshop | Workshop on Aging

Preferred presentation format: full-length

Contact: quanz@hawaii.edu | ORCID: 0009-0009-2409-309X

JEL Classification: I12, J14, J22, J26, O15

Abstract

This paper investigates how health deterioration affects labor supply among older adults in China, distinguishing between acute health shocks and chronic frailty accumulation. Using four waves of the China Health and Retirement Longitudinal Study (2011–2018), I construct individual-level Rockwood frailty indices—both a standard self-report version and a biomarker-enhanced version incorporating blood-based clinical markers—for 21,951 individuals observed across 70,263 person-wave observations. I document a sharp divergence: acute health shocks, identified through Callaway and Sant'Anna (2021) staggered difference-in-differences, produce a precisely estimated null effect on labor force participation (ATT $= +0.002$, $p = 0.856$). In contrast, chronic frailty accumulation exerts a statistically significant negative effect under two-way fixed effects estimation ($\hat{\beta} = -0.050$, SE $= 0.023$). The biomarker-enhanced frailty index, constructed from blood chemistry and hematological markers, yields an effect magnitude roughly ten times larger ($\hat{\beta} = -0.545$, SE $= 0.040$), consistent with biomarkers capturing subclinical physiological dysregulation invisible to self-report. Mechanism decomposition reveals that instrumental activities of daily living limitations constitute the dominant labor-supply-relevant domain within the frailty composite, attenuating the frailty–labor relationship by 53.3 percent, followed by basic ADL limitations (36.2%) and depressive symptoms (19.8%). I find substantial heterogeneity: a positive frailty–female interaction ($\hat{\beta} = +0.132$, $p = 0.001$) indicating a floor effect in women's already-low baseline participation, and a negative frailty–insurance interaction ($\hat{\beta} = -0.090$, $p = 0.005$) consistent with moral hazard. Results survive a 96-specification curve analysis (100% negative, median $\hat{\beta} = -0.062$), Lee bounds excluding zero, Oster stability tests ($\delta^* < 0$), and Cinelli–Hazlett sensitivity analysis (RV$_q = 1.1\%$). These findings imply that China's delayed retirement reforms must contend with a binding chronic frailty constraint that discrete health events alone do not capture.

Keywords: Frailty index, health shocks, labor supply, aging, China, CHARLS, biomarkers, difference-in-differences

1. Introduction

China confronts a demographic transformation of unprecedented scale. Between 2020 and 2050, the share of the population aged 60 and above will nearly double from 18.7 to 35 percent, adding roughly 200 million individuals to the elderly cohort while the working-age population contracts by a comparable magnitude. The old-age dependency ratio, already straining pension and healthcare systems, will rise from approximately 28 to 70 percent. In response, policymakers have proposed raising statutory retirement ages—currently 60 for men and 50–55 for women—as a central pillar of fiscal sustainability. Yet these proposals implicitly assume that older workers possess sufficient health capital to extend their productive years. Whether this assumption holds depends critically on the nature and magnitude of the health–labor supply relationship in later life.

A large literature in health economics has documented that health deterioration reduces labor supply among older adults (for surveys, see Currie and Madrian, 1999; Strauss and Thomas, 1998). The canonical approach identifies effects using acute health events—hospitalizations, heart attacks, strokes, or cancer diagnoses—as quasi-exogenous shocks that shift individuals from employment to non-participation. García-Gómez et al. (2013) estimate that acute health shocks reduce employment by 7–8 percentage points in the Netherlands, with persistent effects over five years. Fadlon and Nielsen (2021) document spousal labor supply responses to severe cardiovascular events in Denmark. Blundell et al. (2023) compare subjective and objective health measures in the Health and Retirement Study and the English Longitudinal Study of Ageing, finding that health deterioration explains up to 15 percent of employment decline between ages 50 and 70. Bergeot et al. (2024) applies heterogeneity-robust difference-in-differences to acute health events in SHARE and HRS data.

This body of work, while methodologically sophisticated, shares a common conceptual limitation: it treats health as a binary state that shifts discretely upon the occurrence

of a diagnosable event. In reality, health among older adults deteriorates through the gradual accumulation of subclinical deficits—minor losses in grip strength, slight cognitive slowing, emerging difficulty with instrumental activities of daily living—that individually fall below diagnostic thresholds but collectively erode productive capacity. The geriatric medicine literature, beginning with Rockwood and Mitnitski (2007), has formalized this insight through the frailty index (FI), defined as the ratio of health deficits present to health deficits measured. The FI captures the multidimensional nature of aging-related health decline and exhibits robust empirical regularities: it increases approximately log-linearly with age, converges toward a biological ceiling near 0.70, and predicts mortality, hospitalization, and disability with greater precision than any single diagnostic indicator.

The theoretical underpinning for this distinction comes from Dalgaard and Strulik (2014), who model health as a stock of accumulated deficits D evolving according to a dis-equilibrating differential equation:

$$\dot{D} = \mu \cdot D - f(h) + \varepsilon, \tag{1}$$

where $\mu > 0$ is the intrinsic aging rate, $f(h)$ represents health investment (a concave repair function), and ε captures stochastic health shocks. The key departure from Grossman's (1972) canonical health capital model is that $\mu > 0$ implies deficits beget deficits—a positive feedback loop generating accelerating health depreciation in later life. In Grossman's framework, health depreciates at a fixed (or slowly rising) rate $\delta(t)$; in Dalgaard and Strulik's framework, the depreciation rate is endogenous to the current deficit stock, producing threshold dynamics and an eventual "frailty cliff" beyond which recovery becomes biologically infeasible. This model yields three testable predictions: (i) the frailty index should serve as a sufficient statistic for the health–labor relationship; (ii) acute shocks should have transient effects relative to the chronic trajectory; and (iii) the FI–labor relationship should exhibit threshold nonlinearity as individuals approach a critical deficit

burden. Dalgaard et al. (2022) provide cross-country support, estimating that a 1 percent increase in cohort-level frailty reduces labor force participation by approximately 0.6 percentage points across 180 countries. Abeliansky and Strulik (2023) extend the Dalgaard–Strulik framework to show that frailty dynamics can account for cross-country differences in the timing of labor force exit.

I test these predictions using individual-level panel data from the China Health and Retirement Longitudinal Study (CHARLS), which surveys a nationally representative sample of Chinese adults aged 45 and above. I restrict the analysis to four waves (2011–2018) to avoid COVID-19 contamination of the 2020 data, yielding 70,263 person-wave observations on 21,951 individuals clustered in 450 communities. CHARLS is uniquely suited to this investigation for three reasons. First, it collects sufficiently rich health information—activities of daily living, instrumental activities, chronic disease diagnoses, cognitive assessments, depressive symptoms, and physical performance measures—to construct a Rockwood frailty index following the Searle et al. (2008) standardized protocol, with a variable denominator accommodating data availability across waves ($D_{W1} = 22$, $D_{W2\text{–}W3} = 23$, $D_{W4} = 19$ items). Second, CHARLS conducted blood biomarker collection in waves 1 (2011) and 3 (2015), enabling the construction of a biomarker-enhanced frailty index (Bio-FI) that augments self-reported deficits with objective clinical markers including glycated hemoglobin (HbA1c), fasting glucose, lipid panels, inflammatory markers (hsCRP), and renal function indicators, for a total of 31 deficit items. Third, the CHARLS population is predominantly rural (76.7 percent hold rural *hukou*), operates largely in the agricultural sector without formal retirement institutions, and thus faces health–labor tradeoffs directly relevant to delayed retirement policy design.

Contributions. This paper makes five contributions to the literature.

First, to my knowledge, this is the first study to introduce the Rockwood deficit-accumulation frailty index as a treatment variable in a causal analysis of labor supply. Previous work has used either single health indicators (*e.g.*, self-rated health, specific

diagnoses) or composite indices constructed via principal components or factor analysis (Blundell et al., 2023). The frailty index offers advantages in interpretability (each unit change represents one additional deficit per deficit measured), theoretical grounding (it maps directly onto Dalgaard and Strulik's deficit stock D), and empirical robustness (it satisfies established psychometric criteria including the Searle et al. 99th percentile submaximal limit and age-gradient tests).

Second, I conduct an explicit head-to-head comparison of acute versus chronic health effects within the same sample and empirical framework. Using Callaway and Sant'Anna's (2021) staggered difference-in-differences estimator with never-treated and not-yet-treated controls, I estimate the causal effect of acute health shocks—defined as sharp deteriorations in the frailty index ($\Delta\text{FI} > \mu + \sigma$)—on labor force participation. The resulting average treatment effect on the treated is $\widehat{\text{ATT}} = +0.002$ (SE $= 0.013$, $p = 0.856$)—a precise zero. This null finding is robust across Sun and Abraham (2021), Borusyak et al. (2024), and multiple treatment threshold definitions. The chronic frailty effect, estimated via two-way fixed effects with individual and wave fixed effects,

$$Y_{it} = \alpha_i + \gamma_t + \beta \cdot \text{FI}_{it} + \mathbf{X}'_{it}\boldsymbol{\delta} + \varepsilon_{it}, \tag{2}$$

where Y_{it} is a binary indicator for labor force participation, α_i and γ_t are individual and wave fixed effects, FI_{it} is the frailty index, $\mathbf{X}_{it}$ is a vector of time-varying controls (age, age-squared, marital status, household size, rural/urban status), and standard errors are clustered at the community level (450 clusters), yields $\hat{\beta} = -0.050$ (SE $= 0.023$)—statistically significant and economically meaningful. Evaluated at a per-0.1-unit increase in the frailty index (approximately one standard deviation, $\sigma_{\text{FI}} = 0.109$), this implies a reduction in labor force participation of roughly 0.5 percentage points, a magnitude consistent with Dalgaard et al. (2022), who report that a 1 percentage-point increase in cohort-level frailty reduces participation by 0.6 percentage points across 180 countries.

Third, I validate the chronic frailty effect using objective biomarker data. The biomarker-enhanced frailty index yields $\hat{\beta}_{\text{Bio}} = -0.545$ (SE $= 0.040$)—roughly ten times the magnitude of the standard self-reported FI. This substantially larger coefficient should be interpreted with caution: the Bio-FI specification uses pooled cross-sectional estimation (individual fixed effects estimation is severely constrained on the blood subsample, where not all individuals provide blood in both collection waves), which re-admits between-person heterogeneity that panel fixed effects absorb. Nonetheless, the magnification is consistent with the Bio-FI capturing subclinical physiological dysregulation—invisible to self-report—that strengthens the measured health–labor gradient. I further examine the exclusion restriction that biomarkers affect labor supply only through the frailty pathway using a formal two-stage test and an HbA1c regression discontinuity design exploiting the diabetes diagnostic threshold at 6.5%. The first stage is effectively zero ($\hat{\beta} = -0.011$, SE $= 0.031$), consistent with the empirical absence of a diagnostic information channel in this population, while the reduced-form relationship between the HbA1c threshold and labor force participation is negative and significant ($\hat{\beta} = -0.082$, SE $= 0.039$, $p = 0.035$). This pattern—null first stage, significant reduced form—suggests a direct physiological channel operating below diagnostic awareness, in stark contrast to the strong information effects documented by Alalouf et al. (2024) in the U.S. context where patients receive and act upon their test results.

Fourth, mechanism decomposition following Gelbach (2016) identifies IADL limitations as the frailty domain carrying the strongest labor-supply-relevant content, accounting for 53.3 percent of the frailty–labor relationship. I estimate a sequential mediation model:

$$Y_{it} = \alpha_i + \gamma_t + \gamma \cdot \text{FI}_{it} + \sum_{m=1}^{M} \theta_m \cdot \text{Med}_{m,it} + \mathbf{X}_{it}'\boldsymbol{\delta} + \varepsilon_{it}, \tag{3}$$

where $\text{Med}_{m,it}$ denotes mediator m (IADL limitations, ADL limitations, CES-D depression score, cognitive score, self-rated health). The baseline coefficient without mediators is

$\hat{\gamma}_0 = -0.049$ (SE $= 0.023$, $N = 68{,}009$). Attenuation upon inclusion of each mediator is computed as $(\hat{\gamma}_0 - \hat{\gamma}_m)/\hat{\gamma}_0 \times 100\%$. Basic ADL limitations contribute 36.2 percent and depressive symptoms 19.8 percent, while cognitive decline is negligible at 1.1 percent. Self-rated health produces over-attenuation (148.5 percent), confirming its status as a "bad control" that absorbs variation properly attributed to functional limitations.

Fifth, I document policy-relevant heterogeneity. The frailty effect is concentrated among men ($\hat{\beta} = -0.124$), with women exhibiting a near-zero effect ($\hat{\beta} = +0.021$)—a pattern explained by a floor effect: women's baseline labor force participation is already depressed by structural gender inequality, leaving limited margin for health-induced exit. The interaction between frailty and enrollment in China's New Cooperative Medical Scheme (NCMS) is negative and significant ($\hat{\beta} = -0.090$, $p = 0.005$), indicating that insurance coverage facilitates labor exit—consistent with moral hazard in health insurance. Exploratory causal forest analysis (Athey et al., 2019) confirms significant treatment effect heterogeneity (BLP differential $= 0.509$, $p < 0.001$; AUTOC $= 0.071$, $p < 0.001$).

These findings carry direct implications for China's delayed retirement debate. If chronic frailty accumulation—rather than discrete health events—constitutes the binding constraint on older workers' labor supply, then policies designed around acute-shock frameworks (*e.g.*, disability insurance triggered by diagnostic events) will systematically underestimate the health burden facing the aging workforce. Effective policy requires monitoring cumulative health trajectories and investing in functional capacity preservation, particularly IADL support, well before individuals cross clinical diagnostic thresholds.

The remainder of the paper proceeds as follows. Section 2 provides institutional background on China's rural elderly labor market and the NCMS insurance system. Section 3 describes the CHARLS data, frailty index construction, and variable measurement. Section 4 details the empirical strategy. Section 5 presents the main results. Section 6 examines heterogeneity. Section 7 reports robustness checks. Section 8 discusses policy implications. Section 9 concludes.

2. Institutional Background

2.1 China's Rural Elderly Labor Market

China's labor market institutions create a distinctive context for studying the health–labor supply relationship in later life. Statutory retirement ages—60 for men, 55 for female civil servants, and 50 for female manual workers—apply primarily to workers in the formal urban sector covered by the Urban Employee Basic Pension Insurance system. The majority of China's elderly population, however, holds rural *hukou* (household registration) and has no access to formal retirement institutions. In my analytic sample, 76.7 percent of respondents hold rural *hukou*. For these individuals, the "retirement decision" is not a discrete transition triggered by pension eligibility but rather a gradual withdrawal from productive activity dictated by health capacity, family labor needs, and agricultural seasonality.

Rural elderly workers are overwhelmingly engaged in agriculture. Labor force participation rates among CHARLS respondents aged 45 and above decline modestly across waves—from 66.5 percent in 2011 to 61.4 percent in 2018—but remain far higher than in developed economies, where formal pension systems facilitate earlier exit. The dominant form of economic activity is household agricultural production (growing crops, tending livestock), which is physically demanding but temporally flexible. This agricultural orientation means that health deterioration may affect labor supply differently than in formal employment settings: workers may reduce hours or shift to lighter tasks before ceasing work entirely. I observe this pattern in the data, where frailty effects on agricultural hours are detectable even when extensive-margin (participation) effects are modest.

The absence of formal retirement institutions for rural workers makes this population particularly valuable for studying the health–labor nexus. In settings with generous disability insurance or early retirement provisions, health effects on labor supply are confounded by institutional pathways that facilitate exit conditional on health status

(García-Gómez et al., 2013). In rural China, such institutional channels are largely absent, and the health–labor relationship more closely reflects the biological and economic fundamentals emphasized by Dalgaard and Strulik (2014).

2.2 The New Cooperative Medical Scheme

China's New Cooperative Medical Scheme (NCMS), introduced in 2003 and achieving near-universal rural coverage by 2010, provides the primary source of health insurance for rural residents. NCMS is a voluntary, heavily subsidized program administered at the county level, covering inpatient expenses with a reimbursement rate that varies by facility level and county. Outpatient coverage has been gradually expanded but remains limited in scope. Annual premiums are low—approximately 10–30 RMB per person during the study period, supplemented by central and local government subsidies—making enrollment nearly universal by the time of the CHARLS baseline in 2011.

NCMS is relevant to this study for two reasons. First, insurance coverage may modify the health–labor relationship through moral hazard: insured individuals who experience health deterioration may find it less costly to reduce labor supply, knowing that medical expenses will be partially covered. I test this hypothesis directly by interacting frailty with NCMS enrollment and find a significant negative interaction ($\hat{\beta} = -0.090$, $p = 0.005$), consistent with insurance facilitating labor exit. Second, NCMS enrollment status is a time-varying covariate that I include in robustness specifications to assess sensitivity of the main frailty estimate to insurance controls.

2.3 The China Health and Retirement Longitudinal Study

CHARLS is a nationally representative longitudinal survey of Chinese adults aged 45 and above and their spouses, conducted by Peking University's National School of Development. The survey follows the design template of the U.S. Health and Retirement Study (HRS) and the Survey of Health, Ageing and Retirement in Europe (SHARE), facil-

itating cross-national comparisons. The baseline wave was fielded in 2011 ($N \approx 17{,}000$ respondents), with follow-up waves in 2013, 2015, 2018, and 2020.

I use waves 1 through 4 (2011–2018) as the main analytic sample, excluding wave 5 (2020) due to COVID-19 contamination of both health and labor supply measures. The survey covers 28 provinces and 150 county-level units, with respondents drawn from 450 communities (the clustering level for my standard error computation). CHARLS collects detailed information on health status (chronic diseases, functional limitations, cognitive function, mental health, self-rated health), physical performance (grip strength, timed walk, chair stand, lung function), economic activity (agricultural and non-agricultural work, hours), demographics, household structure, and health insurance enrollment. Crucially, CHARLS conducted venous blood biomarker collection in waves 1 (2011, $N \approx 11{,}847$) and 3 (2015, $N \approx 13{,}013$), following the HRS every-other-wave blood collection protocol. These biomarker data enable the construction of the biomarker-enhanced frailty index that serves as a key validation tool in this study.

3. Data and Measurement

3.1 Sample Construction

I construct a four-wave balanced-eligible panel from CHARLS waves 1 through 4 (2011, 2013, 2015, 2018). Starting from 77,221 raw person-wave observations, I exclude 6,958 observations with missing identifiers or key demographic variables, yielding an analytic sample of 70,263 person-wave observations on 21,951 unique individuals.[1] Not all individuals appear in every wave; 11,718 individuals are observed in all four waves (balanced subsample of 46,872 observations), while the remainder are observed in two or three waves. I assess attrition patterns in Appendix Table A1 and implement Lee (2009) bounds

[1]The STROBE flow diagram documenting all exclusions is presented in Figure S1 in the Online Supplement. Wave 1 household identifiers require zero-padding from 9 to 10 digits to match subsequent waves, a data processing step I implement in the panel construction code.

to address potential selective attrition in Section 7.

Each wave's data are assembled from multiple survey module files that must be merged at the individual level. The merge chain varies by wave due to evolving questionnaire design: wave 1 requires merging demographic background, health status, work and retirement, health insurance, a separate physical biomarkers file (`biomarkers.dta`, lowercase), and a blood sample file; wave 4 requires a separate cognition file for CES-D depression and cognitive assessment items, and uses a different work module filename.[2] I construct community identifiers as the first seven characters of each individual's ID string, yielding 450 community clusters used throughout for standard error computation.

3.2 Frailty Index Construction

I construct the frailty index following the standardized protocol of Searle et al. (2008), which specifies five criteria for deficit inclusion: (i) health-related, (ii) prevalence increases with age, (iii) not saturated too early, (iv) covers multiple physiological systems, and (v) available in the data across waves. I draw deficits from six domains: activities of daily living (ADL: bathing, dressing, eating, getting in/out of bed, toileting, continence), instrumental activities of daily living (IADL: managing money, taking medications, shopping, preparing meals, doing housework), chronic disease diagnoses, depressive symptoms (CES-D scale), cognitive function (word recall), and self-rated health.

A key feature of my construction is the *variable denominator*, accommodating differential data availability across waves. Waves 1–3 include physical performance measures (grip strength, timed walk, chair stand, lung peak flow) collected in a separate biomarker module; waves 4 and 5 do not include physical examinations. Accordingly, the denominator of available deficits is $D_{W1} = 22$, $D_{W2\text{–}W3} = 23$, and $D_{W4} = 19$ items. The frailty index for

[2]Wave 4 CES-D items and cognitive word recall scores are located in `Cognition.dta`, not in the Health Status file as in waves 1–3. The wave 4 work module is `Work_Retirement.dta`, renamed from `Work_Retirement_and_Pension.dta` in earlier waves.

individual i in wave t is:

$$\text{FI}_{it} = \frac{\sum_{j=1}^{D_t} d_{j,it}}{D_t}, \tag{4}$$

where $d_{j,it} \in \{0, 1\}$ is a binary deficit indicator and D_t is the number of available deficit items in wave t. This variable-denominator approach is standard in the geriatric literature when data availability varies across assessment occasions (Searle et al., 2008).

Each deficit is coded as binary (0 = no deficit, 1 = deficit present). For chronic disease diagnoses, I implement a five-layer extraction algorithm to handle CHARLS's complex survey routing across waves.[3] The resulting chronic disease indicators have zero missing values across the full panel.

Cognitive function is measured by word recall (immediate and delayed, 0–10 each, total 0–20). Variable naming patterns differ across all five CHARLS waves; I harmonize these to a common scale.[4] CES-D depression items similarly require wave-specific variable mapping.[5]

The resulting frailty index has a sample mean of 0.184 (SD = 0.109) and ranges from 0 to 0.698. These values align with established benchmarks: the 99th percentile (0.698) falls below the theoretical submaximal limit of approximately 0.70, above which survival becomes biologically infeasible (Searle et al., 2008). The FI increases with age at a gradient of approximately 0.011 per year, consistent with cross-population norms. I present the full distribution by wave in Figure 1 and validation diagnostics in Appendix Table A4.

[3]The five layers are: (1) direct diagnosis question (`da007`); (2) wave 5 renamed variable (`da003`); (3) preloaded confirmation from prior wave (`zda007`, value = 1 only); (4) returning respondent confirmation (`da007_w2_1`, where 1 = agree and 2 = disagree with preloaded diagnosis); and (5) new diagnosis since last wave (`da007_w2_2`). In wave 3, the direct diagnosis question has only 2.5 percent non-missing coverage, making the preload-confirmation pathway essential. I apply an absorbing-state correction using cumulative maximum (`cummax`) to prevent disease "reversals" due to respondent disagreement with preloaded diagnoses; this corrects 250 cases of apparent dementia reversal.

[4]Wave 1: `dc006s1-s10` and `dc027s1-s10` in Health Status. Wave 2: `dc006_1_s1-s10`. Wave 3: `dc006s1-s10`. Wave 4: `dc028_w4_s1-s10` and `dc047_w4_s1-s10` in Cognition.dta. Wave 5: `dc012_s1-s10` and `dc013_s1-s10`.

[5]Waves 1–3: `dc009-dc018` in Health Status. Wave 4: same item codes in Cognition.dta. Wave 5: shifted to `dc016-dc025`.

3.3 Biomarker-Enhanced Frailty Index

Blood biomarker data are available in CHARLS waves 1 (2011) and 3 (2015) only, yielding a biomarker subsample of 23,316 person-wave observations (23,338 total blood records, of which 23,316 have non-missing Bio-FI).[6] I construct three biomarker-based indices:

Bio-FI (Biomarker-enhanced frailty index): Combines the standard FI deficit items with 12 blood-based clinical deficit indicators, for a total of $D = 31$ items. Blood deficits are defined using established clinical thresholds: HbA1c $\geq$ 6.5% (diabetes), fasting glucose $\geq$ 126 mg/dL, total cholesterol $\geq$ 240 mg/dL, HDL $<$ 40 mg/dL, LDL $>$ 160 mg/dL, triglycerides $\geq$ 200 mg/dL, hsCRP $>$ 3 mg/L (inflammation), BUN $>$ 20 mg/dL, creatinine $>$ 1.4 mg/dL, uric acid $>$ 7.0 mg/dL (male) or $>$ 6.0 mg/dL (female), and cystatin C $>$ 1.44 mg/L. Most CHARLS blood biochemistry values are reported in mg/dL; hsCRP and cystatin C are in mg/L; HbA1c is in percent.[7] The Bio-FI mean is 0.165, lower than the standard FI mean (0.184) due to denominator dilution ($D = 31$ vs. $D \approx 22$), as expected when adding a larger denominator with proportionally fewer deficits.

Bio-FI (no-HbA1c): A leave-one-out version that excludes HbA1c from the Bio-FI construction. This version is necessary for the HbA1c regression discontinuity analysis in Section 5: since HbA1c serves as both the running variable for the RD and a component of the standard Bio-FI, using the full Bio-FI would create a mechanical exclusion restriction violation. The no-HbA1c version avoids this issue.

Biomarker Severity Index (BSI): Constructed from blood-based deficits only ($D = 12$), excluding all self-reported items. The BSI provides the cleanest test of objective health effects on labor supply, at the cost of reduced sample size and a narrower health domain.

I validate the Bio-FI against established criteria following Searle et al. (2008) and the

[6] CHARLS follows the HRS every-other-wave blood collection protocol. Waves 2, 4, and 5 contain no blood data. The wave 1 blood file uses 11-character individual IDs that must be zero-padded to 12 characters to match the analytic panel.

[7] The CHARLS documentation and early design guides contain conflicting unit specifications (mg/dL vs. mmol/L). I have verified against the raw data and the official Blood Data Release Notes that serum biochemistry variables (glucose, cholesterol, triglycerides, BUN, creatinine, uric acid) are in mg/dL; inflammatory and renal markers (hsCRP, cystatin C) are in mg/L; and HbA1c is in percent.

recent extension by Blodgett (2024). The Bio-FI age gradient (0.0143 per year) is steeper than the standard FI gradient (0.0107), consistent with biomarkers capturing subclinical deterioration that self-report measures miss. Validation results are presented in Appendix Figure A2 and Table A4.

3.4 Labor Force Participation

The outcome variable is a binary indicator for labor force participation (LFP), defined as engaging in agricultural work, non-agricultural employment, or being on temporary leave from either. Constructing consistent LFP measures across CHARLS waves requires careful attention to changing questionnaire routing:

In waves 1–3, the primary agricultural work question (FA001, "Have you engaged in agricultural work for at least 10 days in the past year?") serves as the gateway, with 100 percent response rates. Non-agricultural work is captured by FA002, and temporary leave by FA003. In wave 4, FA001 is replaced by FC008 ("Has your household engaged in agricultural production?") and FC001 ("Have you been employed in agricultural work?"), with non-agricultural work captured by a renamed FA002_W4.[8] I code $\text{LFP}_{it} = 1$ if any of the wave-appropriate agricultural or non-agricultural work indicators is affirmative.

Sample LFP rates decline gradually from 66.5 percent (wave 1) to 65.9 percent (wave 2), 64.4 percent (wave 3), and 61.4 percent (wave 4), consistent with population aging and modest economic development over the period. Trends by wave are displayed in Supplement Figure S3.

3.5 Control Variables and Summary Statistics

Time-varying control variables included in the baseline TWFE specification (eq. (2)) are: age, age-squared, marital status (married/partnered vs. other), household size, and an indicator for rural *hukou*. In extended specifications, I add NCMS enrollment status.

[8]Omitting FC008 from the wave 4 LFP construction would cause an apparent LFP cliff from approximately 67 percent to 30 percent—an artifact of questionnaire redesign, not a real behavioral change.

I deliberately exclude the number of chronic disease diagnoses (`n_chronic`) from all specifications because it is a "bad control" in the sense of Angrist and Pischke (2009): the within-individual correlation between the frailty index and chronic disease count is 0.57, and including `n_chronic` flips the frailty coefficient from -0.049 to $+0.035$—a sign reversal indicating collinearity-induced bias rather than a genuine confounding adjustment.[9]

Summary statistics for the full sample and by wave are presented in Table 1. The frailty index construction details, including all deficit items and their wave-specific availability, are presented in Table 2.

4. Empirical Strategy

I employ four complementary empirical approaches to identify the health–labor supply relationship: (i) two-way fixed effects (TWFE) estimation for the chronic frailty effect; (ii) Callaway–Sant'Anna staggered difference-in-differences for acute health shocks; (iii) sequential mechanism decomposition; and (iv) biomarker validation via a first-stage test and an HbA1c regression discontinuity design.

4.1 Two-Way Fixed Effects: Chronic Frailty

The primary estimating equation for the chronic frailty–labor relationship is:

$$Y_{it} = \alpha_i + \gamma_t + \beta \cdot \text{FI}_{it} + \mathbf{X}'_{it}\boldsymbol{\delta} + \varepsilon_{it}, \tag{5}$$

where $Y_{it} \in \{0, 1\}$ is an indicator for labor force participation of individual i in wave t; α_i is an individual fixed effect absorbing all time-invariant heterogeneity (genetics, childhood health, education, permanent preferences); γ_t is a wave fixed effect capturing common time shocks (macroeconomic conditions, policy changes); $\text{FI}_{it} \in [0, 1]$ is the frailty index;

[9] The chronic disease count is an input to the frailty index construction (chronic diseases constitute one of six deficit domains). Conditioning on a component of the treatment variable induces mechanical bias. This issue is distinct from the standard endogenous mediator concern; here the problem is algebraic, not causal.

$\mathbf{X}_{it}$ is a vector of time-varying controls (age, age-squared, marital status, household size, rural *hukou* status); and ε_{it} is an idiosyncratic error term. Standard errors are clustered at the community level (450 clusters) to account for within-community correlation in both health trajectories and labor market conditions.

The coefficient β identifies the effect of within-individual frailty accumulation on labor force participation, after removing permanent individual heterogeneity and common time trends. Identification requires that conditional on α_i, γ_t, and $\mathbf{X}_{it}$, changes in frailty are uncorrelated with unobserved determinants of labor supply changes. This assumption would be violated if, for example, an unobserved negative productivity shock simultaneously caused individuals to both deteriorate in health and exit the labor force through a channel other than health. Individual fixed effects address the most obvious threats (ability, permanent health endowment), and the within-R^2 of the specification is 0.0009—low in absolute terms but consistent with the modest explanatory power of any single regressor in panel data with individual fixed effects, as discussed by Deaton (1997).[10]

I estimate Equation (5) across progressively richer specifications: pooled OLS (no controls or fixed effects), pooled OLS with demographic controls, pooled OLS with demographics plus province and wave fixed effects, individual plus wave fixed effects with time-varying controls (the preferred specification), and additionally with NCMS insurance controls. The preferred specification includes individual and wave fixed effects with time-varying controls but excludes insurance (which may be endogenous to health).

For the biomarker-enhanced frailty index, which is available only in waves 1 and 3, I estimate a pooled cross-sectional specification:

$$Y_{it} = \alpha + \beta_{\text{Bio}} \cdot \text{Bio-FI}_{it} + \mathbf{X}_{it}'\boldsymbol{\delta} + \gamma_t + \varepsilon_{it}, \tag{6}$$

where γ_t is a wave indicator for wave 3. Individual fixed effects estimation on the blood

[10]The low within-R^2 reflects the fact that individual fixed effects absorb the vast majority of variation in labor force participation. The between-R^2 is substantially higher. I address residual omitted variable concerns formally through Oster (2019) and Cinelli–Hazlett (2020) sensitivity analyses in section 7.

subsample is severely constrained: not all individuals provide blood in both waves, yielding an overlap sample too small for precise within-individual estimation ($N = 23{,}316$ total blood-wave observations).

4.2 Staggered DiD: Acute Health Shocks

To estimate the causal effect of acute health shocks, I define treatment as a sharp deterioration in the frailty index between consecutive waves, specifically $\Delta\text{FI} > \mu + \sigma$ (where μ and σ are the mean and standard deviation of the wave-to-wave FI change distribution). This threshold captures acute functional decline—a sudden worsening across multiple health domains—regardless of whether the underlying cause is a specific diagnosable event. Because the FI item set changes across waves (physical performance items are unavailable in wave 4, reducing the denominator from 23 to 19), the variable-denominator construction (eq. (4)) ensures that FI *levels* remain comparable as deficit-to-available-item ratios, so that wave-to-wave changes in the FI reflect genuine deficit accumulation rather than mechanical changes in the available item set. To further guard against the possibility that W3→W4 transitions are mechanically flagged as shocks due to the item-set change, I verify in a sensitivity check that excluding all W3→W4 transition observations from the treated-group classification leaves the acute-shock null result intact (results available on request). Results are further verified to be robust across multiple shock threshold definitions (0.5–2.0 SD; Appendix Figure A8). I apply the Callaway and Sant'Anna (2021) estimator, which is robust to heterogeneous treatment effects across cohorts and time periods:

$$ATT(g,t) = \mathbb{E}\left[Y_t(g) - Y_t(0) \mid G_i = g\right], \tag{7}$$

where g denotes the treatment cohort (wave of first shock) and $Y_t(g)$ and $Y_t(0)$ are potential outcomes under treatment and non-treatment, respectively. I use both never-treated and not-yet-treated individuals as controls. The estimator is doubly robust, combining outcome

regression with inverse propensity weighting. I aggregate group-time ATTs into an overall ATT:

$$\widehat{\text{ATT}} = \sum_{g} \sum_{t \geq g} w_{g,t} \cdot \widehat{ATT}(g,t), \tag{8}$$

and into event-time dynamic effects $\hat{\tau}^{ES}(e)$ for event-time e relative to treatment onset, normalizing at $e = -1$. Bootstrap standard errors are clustered at the individual level.

A practical limitation is that with a four-wave panel and staggered treatment onset, only one pre-treatment period ($k = -4$ relative to baseline) is available for pre-trends assessment. I address this limitation using Rambachan and Roth (2023) HonestDiD sensitivity analysis, which characterizes the identified set under bounded violations of parallel trends. I also verify robustness to alternative estimators: Sun and Abraham (2021), Borusyak et al. (2024), and multiple treatment threshold definitions (0.5, 1.0, 1.5, 2.0 SD increases in the frailty index; Appendix Figure A8). The continuous treatment estimator (`contdid`) did not converge due to the limited panel length; I note this transparently.

4.3 Mechanism Decomposition

To identify the channels through which frailty affects labor supply, I estimate a Gelbach-style sequential decomposition. The baseline specification excludes all mediators:

$$Y_{it} = \alpha_i + \gamma_t + \gamma_0 \cdot \text{FI}_{it} + \mathbf{X}'_{it}\boldsymbol{\delta} + \varepsilon_{it}, \tag{9}$$

yielding a baseline coefficient $\hat{\gamma}_0$. I then augment the specification with each mediator m sequentially:

$$Y_{it} = \alpha_i + \gamma_t + \gamma_m \cdot \text{FI}_{it} + \theta_m \cdot \text{Med}_{m,it} + \mathbf{X}'_{it}\boldsymbol{\delta} + \varepsilon_{it}, \tag{10}$$

and compute the attenuation share as $(\hat{\gamma}_0 - \hat{\gamma}_m)/\hat{\gamma}_0 \times 100\%$. I consider five mediators: IADL limitations (count of difficulties with managing money, taking medications, shopping, meal preparation, housework), ADL limitations (count of difficulties with bathing,

dressing, eating, transferring, toileting, continence), CES-D depression score (0–30), cognitive score (word recall, 0–20), and self-rated health (1–5, reverse-coded so higher values indicate worse health). I interpret large attenuation as evidence that the mediator transmits a substantial share of the frailty effect, while over-attenuation (greater than 100 percent) signals a "bad control" that absorbs treatment variation.

4.4 Biomarker Validation

The biomarker-enhanced frailty index provides a validation test under the exclusion restriction assumption that blood biomarkers affect labor supply only through the health/frailty pathway—not directly. This assumption is plausible in the CHARLS context for two reasons. First, although CHARLS's Wave 3 protocol included mailing biochemistry reports to participants (Chen et al., 2019), health literacy among rural elderly Chinese remains extremely low; receiving a numerical laboratory report does not equate to understanding a clinical diagnosis or altering labor supply behavior. Second, and most importantly, the empirical first stage provides no evidence that the information channel meaningfully activated: the probability of diabetes diagnosis shows no discontinuous jump at the HbA1c $= 6.5\%$ threshold (first-stage coefficient $= -0.011$, SE $= 0.031$, $p = 0.735$), making a dominant informational channel unlikely in this context. The exclusion restriction would be violated if biomarkers directly affected productivity through physiological channels without operating through the functional limitations captured by the frailty index.

I implement two validation approaches. First, I estimate the Bio-FI effect (Equation 6) and compare its magnitude to the standard FI effect. If the standard self-reported FI suffers from measurement error or fails to capture subclinical pathology, the Bio-FI—which augments self-reports with objective biomarkers—should yield a larger (in absolute value) coefficient. I test the exclusion restriction formally using a two-stage procedure and report the bootstrap p-value for the exclusion restriction test (TP1 bootstrap: $p = 0.000$, 95% CI $[0.220, 0.319]$).

Second, I exploit the HbA1c diabetes diagnostic threshold at 6.5% in a sharp regression discontinuity design:

$$\text{FI}_i = \pi_0 + \pi_1 \cdot \mathbb{1}[\text{HbA1c}_i \geq 6.5] + f(\text{HbA1c}_i) + \mathbf{X}_i'\boldsymbol{\phi} + \nu_i, \quad \text{(first stage)} \tag{11}$$

$$Y_i = \delta_0 + \delta_1 \cdot \mathbb{1}[\text{HbA1c}_i \geq 6.5] + g(\text{HbA1c}_i) + \mathbf{X}_i'\boldsymbol{\psi} + u_i, \quad \text{(reduced form)} \tag{12}$$

where $f(\cdot)$ and $g(\cdot)$ are local polynomial functions of the running variable. I use the Cattaneo et al. (2020) robust bias-corrected confidence intervals with data-driven bandwidth selection (CCT bandwidth). The first-stage estimate is $\hat{\pi}_1 = -0.011$ (SE $= 0.031$, $p = 0.735$)—effectively zero, indicating that despite CHARLS's protocol of mailing biochemistry reports to participants, the diagnostic information channel did not appear to meaningfully activate in this predominantly rural, low-health-literacy population. The reduced-form estimate is $\hat{\delta}_1 = -0.082$ (SE $= 0.039$, $p = 0.035$)—significant, suggestive of a direct physiological effect of elevated HbA1c on labor capacity operating below diagnostic awareness. This pattern contrasts sharply with Alalouf et al. (2024), who finds a strong first stage in the U.S. context where patients receive and act upon HbA1c results, highlighting the importance of clinical pathway infrastructure for the information channel in health–labor relationships.

5. Results

5.1 Chronic Frailty and Labor Force Participation

Table 3 presents the TWFE estimates of the frailty–labor relationship across five specifications. Column (1) reports the pooled OLS estimate without controls or fixed effects ($\hat{\beta} = -0.526$, SE $= 0.031$), which captures both between- and within-individual variation and is subject to substantial omitted variable bias. Column (2) adds demographic controls (age, age-squared, gender, education, marital status, rural *hukou*), reducing the coefficient

to $\hat{\beta} = -0.306$ (SE $= 0.028$) as cross-sectional confounders are absorbed. Column (3) further adds province and wave fixed effects ($\hat{\beta} = -0.312$, SE $= 0.024$), with the coefficient remaining stable. Column (4)—the preferred specification—replaces pooled cross-sectional controls with individual fixed effects and wave fixed effects, yielding $\hat{\beta} = -0.050$ (SE $= 0.023$, $N = 68{,}009$). The sharp decline from Column (3) to Column (4) reflects the elimination of between-person heterogeneity: permanent individual characteristics (genetics, childhood health, education, baseline preferences) account for the vast majority of the raw frailty–LFP correlation, and the remaining within-individual effect—while smaller—is the policy-relevant within-person gradient, subject to the strict exogeneity assumption that I assess through sensitivity analyses in section 7. Column (5) adds NCMS insurance enrollment ($\hat{\beta} = -0.074$, SE $= 0.023$, $N = 67{,}641$); the coefficient magnitude increases, suggesting that failing to account for insurance enrollment attenuates the frailty effect toward zero.

The preferred estimate of -0.050 implies that a one-unit increase in the frailty index—moving from zero deficits to the maximum possible score of 1.0—reduces labor force participation by 5.0 percentage points. Evaluated at one standard deviation ($\sigma_{\text{FI}} = 0.109$), the effect is approximately $0.109 \times 0.050 = 0.55$ percentage points. The within-R^2 is 0.0009, reflecting the dominant role of individual fixed effects in explaining LFP variation; I discuss this low value in the context of Deaton (1997) in section 7.

The binscatter relationship between the frailty index and LFP, residualized on individual and wave fixed effects, is displayed in Figure 2, showing a clear negative gradient that steepens at higher FI values.

5.2 Acute Health Shocks: A Precise Null

Table 4 reports the Callaway and Sant'Anna (2021) staggered DiD estimates for acute health shocks. The overall ATT for labor force participation is $+0.002$ (SE $= 0.013$, $p = 0.856$)—a precisely estimated zero. This null result is robust across estimators: Sun and Abraham

(2021) yields $+0.011$ (SE $= 0.008$, $p = 0.175$), and Borusyak et al. (2024) yields -0.003 (SE $= 0.006$, $p = 0.555$). The continuous treatment DiD estimator (`contdid`) did not converge due to the limited panel length.

The event study plot (Figure 3) displays dynamic treatment effects from the Callaway–Sant'Anna estimator. The single available pre-treatment period ($k = -4$) shows no evidence of differential pre-trends. Post-treatment coefficients fluctuate around zero without a discernible pattern. The HonestDiD sensitivity analysis (Figure 4) confirms that the null result is robust: at $\bar{M} = 0$ (exact parallel trends), the robust confidence interval is $[-0.018, 0.026]$, and the interval contains zero for all values of $\bar{M}$ up to 2.

The lone exception to the overall null is agricultural hours: the CS ATT is -154.7 hours (SE $= 65.5$, $p = 0.018$), suggesting that acute shocks may induce intensive-margin reductions in agricultural work without triggering full labor force exit. This is consistent with the flexible nature of agricultural labor in rural China, where workers can reduce hours without formally leaving the labor force.

5.3 Biomarker-Enhanced Frailty Index

Table 5 presents the biomarker-enhanced frailty index (Bio-FI) estimates. The Bio-FI coefficient is $\hat{\beta}_{\text{Bio}} = -0.545$ (SE $= 0.040$, $N = 23{,}316$), approximately ten times the magnitude of the within-individual TWFE estimate (-0.050). However, this tenfold difference conflates two sources: the shift from within-individual to pooled estimation, and the incremental contribution of biomarkers. On the same blood-wave subsample, the standard FI under pooled estimation (province + wave FE) yields $\hat{\beta} = -0.276$ (Table 5, Panel A), so the majority of the magnification reflects the change in estimand. The incremental effect of incorporating biomarkers is approximately twofold (-0.276 to -0.545), which I interpret as the contribution of subclinical physiological information absent from self-report. The Biomarker Severity Index (BSI), constructed from blood-based deficits only, yields $\hat{\beta}_{\text{BSI}} = -0.035$ (SE $= 0.004$). The formal exclusion restriction test (TP1 bootstrap) rejects the

null of no exclusion restriction satisfaction ($p = 0.000$, 95% CI $[0.220, 0.319]$), supporting the interpretation that biomarkers operate through the frailty pathway.

The substantially larger Bio-FI coefficient should be interpreted with caution. The Bio-FI specification uses pooled cross-sectional estimation—individual fixed effects estimation is severely constrained on the blood subsample because CHARLS collected blood only in Wave 1 and Wave 3, meaning not all individuals provide blood in both collection waves. Approximately 8,000 individuals have blood data in *both* waves (≈16,000 person-wave observations out of the 23,316 total blood-wave observations used in Table 5), yielding insufficient within-individual variation for precise fixed-effects estimation—which re-admits between-person heterogeneity that panel fixed effects absorb. Additionally, as DiTraglia (2023) demonstrates, measurement error in bounded indicators need not produce classical attenuation; the direction of bias is ambiguous. I therefore interpret the Bio-FI primarily as evidence that incorporating subclinical physiological dysregulation—invisible to self-report—strengthens the measured health–labor gradient, rather than as a precise quantification of attenuation bias in the standard FI.

5.4 HbA1c Regression Discontinuity

Table 6 presents the HbA1c regression discontinuity results. The RD exploits the ADA diabetes diagnostic threshold at 6.5%—a clinical decision rule, not a physiological discontinuity—to test whether crossing this threshold activates a diagnostic information channel. The first-stage estimate—the effect of crossing the 6.5% threshold on diagnosed diabetes—is $\hat{\pi}_1 = -0.011$ (SE $= 0.031$, $p = 0.735$), indistinguishable from zero. This null first stage indicates that the information channel did not meaningfully activate in this population: despite CHARLS's Wave 3 protocol of mailing biochemistry reports to participants, the absence of a diagnostic jump at the clinical threshold suggests that receiving numerical laboratory values did not translate into awareness or behavioral change among the predominantly rural, low-health-literacy sample.

The reduced-form estimate—the direct effect of crossing the threshold on LFP—is $\hat{\delta}_1 = -0.082$ (SE $= 0.039$, $p = 0.035$), negative and statistically significant. This combination of a null first stage and a significant reduced form is inconsistent with a dominant informational (label) channel at the cutoff; the remaining pattern is suggestive of a direct physiological channel through which the health consequences of elevated HbA1c (fatigue, neuropathy, microvascular damage) may impair work capacity without mediation by diagnostic awareness or behavioral change. This finding contrasts with Alalouf et al. (2024), who documents a first-stage effect of approximately 10.8 percentage points in the U.S. context where patients receive HbA1c results. The contrast highlights the role of clinical infrastructure—specifically, the communication of diagnostic information—in activating the "label effect" channel of health shocks on labor supply. In developing-country settings where clinical feedback loops are weak or absent, the subclinical physiological channel dominates.

5.5 Mechanism Decomposition

Table 8 and Figure 5 present the mechanism decomposition results. The baseline frailty coefficient without any mediators is $\hat{\gamma}_0 = -0.049$ (SE $= 0.023$, $N = 68{,}009$). Sequentially adding each mediator attenuates this coefficient:

IADL limitations account for 53.3 percent of the total frailty effect, making instrumental activities of daily living the dominant labor-supply-relevant domain within the frailty composite. ADL limitations contribute an additional 36.2 percent. Depressive symptoms (CES-D) account for 19.8 percent, a non-trivial share suggesting that the mental health burden of frailty operates partly through reduced motivation or energy for work. Cognitive decline contributes only 1.1 percent—effectively zero—indicating that at the levels of cognitive impairment observed in this population, cognitive function is not a binding constraint on labor supply. Self-rated health produces over-attenuation at 148.5 percent, reversing the sign of the frailty coefficient and confirming that self-rated health is a bad

control that absorbs variation properly attributed to the frailty index through functional limitations. These attenuation results should be interpreted as decomposing the predictive content of the composite FI rather than as identifying distinct causal pathways, since IADL and ADL items are themselves components of the index.

The IADL finding is economically intuitive. Difficulty managing finances, taking medications, shopping for groceries, preparing meals, or doing housework imposes concrete constraints on an individual's capacity to organize productive work—particularly in the agricultural context, where successful farming requires planning, procurement of inputs, and sustained physical effort across growing seasons. The primacy of IADL over ADL limitations suggests that the health–labor constraint operates at the level of executive and organizational functioning, not basic self-care.

5.6 Spousal Dementia

As a supplementary analysis, I estimate the effect of spousal dementia onset on the respondent's labor supply, using the couple panel ($N = 51{,}496$ observations on 17,625 couples in 449 clusters). Table 7 reports uniformly null results: the spousal dementia ATT for LFP is -0.019 (not significant), for total hours $+15.2$ (not significant), and for agricultural hours -30.6 (not significant). These null findings are consistent with the 93.5 percent rate of undetected dementia in rural China (Chen et al., 2011): if dementia goes largely undiagnosed, there is no diagnostic label to trigger a caregiving response or labor supply adjustment. The results align with Fu et al. (2025), who documents null spousal health shock effects among CHARLS elderly respondents.

6. Heterogeneity

6.1 Subgroup Analysis

Figure 6 presents a forest plot of subgroup-specific TWFE estimates of the frailty–LFP relationship. The most striking pattern is the gender asymmetry: the frailty effect among men is $\hat{\beta} = -0.124$, while among women it is $\hat{\beta} = +0.021$—near zero and statistically insignificant. This divergence reflects a floor effect in women's labor force participation. Women's baseline LFP in rural China is already substantially lower than men's due to structural gender inequalities in labor markets, land tenure, and intra-household allocation. With less room for health-induced exit, the extensive-margin response to frailty accumulation is mechanically compressed.

I formalize this finding through interaction specifications in Table A18 (Appendix). The frailty–female interaction is $\hat{\beta} = +0.132$ (SE $= 0.039$, $p = 0.001$), confirming that the frailty effect is significantly less negative for women. This does not imply that frailty is costless for women—it may manifest on other margins (hours, task allocation, informal care burden)—but the extensive-margin LFP response is muted by the low baseline.

6.2 Insurance Heterogeneity

The interaction between frailty and NCMS enrollment is $\hat{\beta} = -0.090$ (SE $= 0.029$, $p = 0.005$), indicating that the frailty–labor exit relationship is stronger among insured individuals. This finding is consistent with moral hazard: insurance coverage reduces the financial risk of medical expenditure following labor force exit, lowering the effective cost of withdrawing from work in response to health deterioration. The policy implication is that insurance expansion, while beneficial for welfare and access to care, may accelerate health-related labor force exit—a tension that policymakers must navigate when designing delayed retirement incentives alongside universal health coverage.

6.3 Nonlinear Effects

The Dalgaard–Strulik model (eq. (1)) predicts that the health–labor relationship should exhibit threshold nonlinearity, with the marginal effect of frailty steepening as individuals approach a critical deficit burden. I test this prediction by augmenting the TWFE specification with a quadratic term and a linear spline:

Quadratic: The estimated coefficient on FI^2 is -0.540 (SE $= 0.117$, $p < 0.001$), confirming strong concavity—the frailty effect accelerates at higher FI values.

Spline: I estimate a two-segment linear spline with a break at FI $= 0.25$. Below the threshold, the slope is $+0.017$ (not significant); above it, the slope is -0.193 (significant). A Wald test rejects equality of slopes ($\chi^2 = 11.76$, $p = 0.0006$). The marginal effect at the sample mean (FI $= 0.184$) is -0.023.

These results are displayed in Appendix Figure A11 and Table A15. The spline break at 0.25 is consistent with geriatric thresholds for "pre-frailty" (FI ≥ 0.20) and "frailty" (FI ≥ 0.25) documented in the clinical literature, providing external validation of the estimated nonlinearity.

6.4 Causal Forest (Exploratory)

I complement the parametric heterogeneity analysis with a causal forest (Athey et al., 2019; Wager and Athey, 2018), implemented with 5-fold honest estimation. Table 9 reports the key diagnostics. The average treatment effect from the forest is $\widehat{\text{ATE}} = -0.021$ (SE $= 0.025$), and the BLP test of treatment effect heterogeneity (Chernozhukov et al., 2025) yields a differential estimate of 0.509 (SE $= 0.080$, $p < 0.001$)—indicating statistically significant heterogeneity in the frailty–labor relationship. The AUTOC (area under the targeting operator characteristic) is 0.071 (SE $= 0.012$, $p < 0.001$), confirming that the forest successfully ranks individuals by treatment effect intensity.

Figure 7 displays the CATE distribution and GATES (Group Average Treatment Effects) by quintile. The distribution is right-skewed, with most individuals experiencing

small negative effects and a tail experiencing larger effects. The GATES pattern shows a monotonic gradient: the most negatively affected quintile has a group average effect approximately three times that of the least affected quintile.

I emphasize that these results are *exploratory*. The causal forest is a powerful tool for detecting heterogeneity patterns, but the limited number of pre-treatment covariates and the continuous nature of the frailty "treatment" make structural interpretation difficult. The forest identifies *who* is most affected but cannot definitively establish *why* without additional exogenous variation.

7. Robustness

I subject the main results to an extensive battery of robustness checks, organized along four dimensions: specification sensitivity, sample selection, omitted variable bias, and estimation alternatives. Results are summarized in Table 10 and detailed in Appendix Tables A13–A14 and A19.

7.1 Specification Curve Analysis

Following Simonsohn et al. (2020), I estimate a specification curve spanning 96 distinct within-individual specifications, varying along four dimensions: (i) control variable set (baseline controls, extended controls with insurance, and extended controls with insurance and education interactions—3 variants); (ii) fixed effects structure (individual + wave FE, individual + province$\times$wave FE—2 variants); (iii) clustering level (community, household, individual, and two-way—4 variants); and (iv) age sample restriction (full sample 45+, ages 50+, ages 55+, ages 60+—4 variants), yielding $3 \times 2 \times 4 \times 4 = 96$ specifications. Figure A10 displays the full specification curve and Appendix Table A14 details the specification components.

All 96 specifications yield negative point estimates for the frailty coefficient, with a median of -0.062 and an interquartile range of $[-0.073, -0.050]$. The 100 percent same-

sign rate provides strong evidence against the concern that the main result is an artifact of a particular specification choice, including variations in fixed effects structure, clustering level, or age sample definition.

7.2 Attrition and Sample Selection

Panel attrition is a potential concern if individuals who exit the sample between waves are differentially affected by frailty. I implement Lee (2009) bounds, which trim the sample to equalize cell sizes between balanced and unbalanced subgroups. The resulting bounds on the frailty coefficient are $[-0.090, -0.088]$, both of which exclude zero, confirming that the negative frailty effect is robust to worst-case selective attrition. The narrow width of these bounds (0.002) reflects minimal differential attrition: the mean number of observed waves per individual is 2.450 for the treated group versus 2.443 for the control group (Appendix Table A13), implying negligible trimming and correspondingly tight bounds.

I also verify robustness to restricting the sample to the balanced panel (11,718 individuals observed in all four waves, $N = 46{,}872$) and find qualitatively identical results. Appendix Table A1 reports attrition balance tests.

7.3 Omitted Variable Bias

Oster (2019) coefficient stability: I implement the Oster (2019) test for coefficient stability across controlled and uncontrolled specifications. The key statistic is δ^*—the degree of selection on unobservables, relative to observables, that would be required to drive the coefficient to zero. Under the recommended assumption that $R^2_{\max} = 1.3 \times \tilde{R}^2$ (where $\tilde{R}^2$ is the controlled R^2), I obtain $\delta^* = -8.79$ across all specifications. The negative sign indicates that unobservables would need to have the *opposite* correlation structure to observables to explain away the result—an implausible scenario. Figure A13 displays the Oster bounds.

Adding controls moves the coefficient from the short specification ($\hat{\beta}_{\text{short}} = -0.048$) to the full specification ($\hat{\beta}_{\text{full}} = -0.050$)—the coefficient increases in magnitude with controls,

indicating that observable confounders bias the estimate *toward zero*, not away from it. This is the strongest possible pattern for Oster's test: controls move the estimate in the "wrong" direction for an omitted variable bias story.

Cinelli–Hazlett (2020) sensitivity: I implement the Cinelli and Hazlett (2020) partial R^2-based sensitivity analysis. The robustness value $\text{RV}_q = 0.011$ (1.1%) indicates that a confounder would need to explain only 1.1 percent of the residual variation in both the treatment and the outcome to reduce the estimate to zero. The robustness value for statistical significance, $\text{RV}_{q,\alpha} = 0.003$ (0.3%), is even smaller. These values are small in absolute terms, but I benchmark them against observed confounders: the partial R^2 of the strongest observed covariate (marital status) is 0.0009, and even three times this benchmark yields an adjusted estimate of -0.042 with a confidence interval excluding zero. In context, $\text{RV}_q = 1.1\%$ should be evaluated against the fact that individual fixed effects already absorb time-invariant confounders; a confounder three times as strong as marital status in the fixed-effects residual space would still leave the estimate negative and statistically significant. Figure A14 displays the sensitivity contour plot.

7.4 Alternative Estimators and Specifications

Table 10 additionally reports results from: (i) alternative fixed effects structures—replacing individual FE with household + wave FE yields $\hat{\beta} = -0.342$ (SE $= 0.023$), and individual + province×wave FE yields $\hat{\beta} = -0.054$ (SE $= 0.023$); note that household and province FE specifications identify a different estimand (between-household or within-province gradient) rather than the preferred within-individual estimand, and the substantially larger coefficients reflect the re-admission of between-person heterogeneity; (ii) alternative clustering at the household and individual levels, both returning $\hat{\beta} = -0.050$ (consistent), and two-way clustering (individual + wave) returning $\hat{\beta} = -0.050$ (SE $= 0.056$, same sign but wider interval due to only 4 wave-level clusters falling below the asymptotic threshold); (iii) restricting to ages 50+ and 60+, with coefficient magnitudes of -0.058 and

-0.069 respectively, consistent with a stronger frailty constraint at higher ages; and (iv) the balanced panel restriction ($N = 46{,}844$), yielding $\hat{\beta} = -0.051$. All specifications return negative, economically meaningful frailty coefficients.

7.5 Addressing the Low Within-R^2

The within-R^2 of 0.0009 in the preferred specification warrants comment. This value reflects a feature common to panel data models with individual fixed effects: the vast majority of variation in the outcome is between individuals (absorbed by α_i), leaving very little within-individual variation to be explained by time-varying regressors. As Deaton (1997) notes, within-R^2 in fixed effects models is typically an order of magnitude smaller than pooled R^2. The relevant question is not whether the within-R^2 is "large" in absolute terms but whether the frailty coefficient is precisely estimated and robust to confounders—which the Oster and Cinelli–Hazlett tests confirm. I note, however, that slow-moving frailty accumulation combined with classical measurement error in the FI may attenuate within-individual estimates toward zero under fixed effects (Griliches and Hausman, 1986); the within-individual estimate of -0.050 may therefore be attenuated toward zero and can be viewed as potentially conservative; the Bio-FI evidence—which reduces measurement error through objective biomarker augmentation—is consistent with this interpretation.

8. Discussion and Policy Implications

The central finding of this paper—that chronic frailty accumulation, rather than acute health shocks, constitutes the binding constraint on older workers' labor supply in China—carries implications for both policy design and the theoretical understanding of health–labor relationships.

8.1 Delayed Retirement and the Frailty Constraint

China's proposed delayed retirement reforms assume that extending statutory retirement ages will proportionally extend productive working years. My results suggest this assumption is overly optimistic for the rural elderly population. The frailty–labor relationship is negative, statistically significant, and robust across an extensive battery of specification and sensitivity checks. A one-standard-deviation increase in the frailty index reduces labor force participation by approximately 0.55 percentage points among men, and the effect accelerates at higher frailty levels (spline break at FI $= 0.25$, quadratic term significant at $p < 0.001$). For individuals approaching the pre-frailty threshold—roughly one-quarter of the sample in the later waves—further health deterioration produces disproportionately large labor supply reductions.

These findings imply that delayed retirement policies must be accompanied by investments in frailty prevention and functional capacity preservation. Clinical interventions targeting IADL capacity—the dominant labor-supply-relevant domain within the frailty composite, accounting for 53.3 percent of the frailty–labor relationship—may yield the largest labor supply dividends. Programs that support household management, medication adherence, and community-based services could help older adults maintain productive engagement even as their health gradually declines.

8.2 Insurance Design and Moral Hazard

The significant frailty–NCMS interaction ($\hat{\beta} = -0.090$, $p = 0.005$) reveals a tension between two policy objectives: expanding health insurance coverage and maintaining labor force participation among older adults. Insurance reduces the financial risk of medical expenditure following labor force exit, effectively subsidizing the decision to stop working when health deteriorates. This moral hazard channel implies that universal coverage expansion—a central pillar of China's healthcare reform—may accelerate health-related labor force exit precisely among the population that delayed retirement policies

seek to retain.

Resolving this tension may require insurance designs that decouple coverage from labor force status, or that provide enhanced benefits conditional on continued work (e.g., premium subsidies for working enrollees, occupational health screenings, workplace accommodations). The current NCMS structure, which is independent of employment status, provides no such incentives.

8.3 Gender Inequality in the Health–Labor Nexus

The finding that women exhibit a near-zero frailty–LFP relationship (due to a floor effect) should not be interpreted as evidence that frailty is costless for women. Rather, it reflects the structural constraints that depress women's labor force participation regardless of health status. The floor effect means that health deterioration does not trigger additional labor force exit because exit has already occurred through gender-based channels. This implies that gender-equitable labor market reforms—increasing women's baseline participation through childcare provision, land tenure reform, and anti-discrimination enforcement—would not only improve gender equity but also make women's labor supply more responsive to health investments, amplifying the returns to frailty prevention programs.

8.4 Theoretical Implications

The results are broadly consistent with the Dalgaard and Strulik (2014) deficit accumulation framework. The null effect of acute shocks supports the model's prediction that transient health disturbances are absorbed by the health capital stock, while chronic accumulation drives permanent labor supply adjustments. Structural life-cycle models incorporating health dynamics (Hosseini et al., 2022) predict precisely this pattern: gradual health depreciation has larger labor supply consequences than transient shocks because it permanently shifts the health–consumption–leisure frontier. The nonlinear relation-

ship (quadratic and spline specifications) aligns with the model's threshold dynamics: the deficits-beget-deficits feedback loop generates an accelerating decline that produces "frailty cliff" effects beyond a critical threshold. The Bio-FI magnification is consistent with the model's emphasis on measuring the full deficit stock, including subclinical physiological markers that self-report measures miss. These theoretical connections strengthen the case for using frailty indices—rather than single health indicators or composite scores from factor analysis—as the empirical operationalization of health capital in structural models of aging and labor supply.

9. Conclusion

I use four waves of the China Health and Retirement Longitudinal Study (2011–2018) to investigate how health deterioration affects labor supply among older adults, distinguishing between acute health shocks and chronic frailty accumulation. The central finding is a sharp divergence: acute shocks produce a precisely estimated null effect on labor force participation (Callaway–Sant'Anna ATT $= +0.002$, $p = 0.856$), while chronic frailty accumulation significantly reduces labor supply (TWFE $\hat{\beta} = -0.050$, SE $= 0.023$). The biomarker-enhanced frailty index yields an effect ten times larger ($\hat{\beta} = -0.545$), consistent with biomarkers capturing subclinical pathology invisible to self-report. IADL limitations emerge as the dominant labor-supply-relevant domain within the frailty composite (53.3 percent attenuation), and the effect is concentrated among men, with women exhibiting a floor effect.

These findings have three implications. First, China's delayed retirement debate must account for the chronic frailty constraint: health-related labor force exit is driven by the gradual accumulation of subclinical deficits, not by discrete diagnosable events. Policies that target only acute shocks—disability insurance, sickness leave, post-hospitalization support—will miss the primary channel of health-related labor supply reduction. Second, the moral hazard interaction between frailty and insurance coverage implies that universal

health insurance expansion and delayed retirement incentives may work at cross-purposes unless insurance design explicitly accounts for labor supply incentives. Third, the gender floor effect underscores that health investments and labor market reforms are complementary: increasing women's baseline participation is a prerequisite for health investments to produce labor supply returns.

Several limitations merit acknowledgment. The frailty index is a descriptive summary of deficit accumulation and does not itself have a causal interpretation; my identification strategy leverages within-individual variation in the index after conditioning on individual fixed effects, which addresses many but not all endogeneity concerns. No credible instrument exists for the gradual accumulation of health deficits, and I acknowledge that reverse causality—labor exit accelerating frailty—cannot be fully ruled out. However, the Oster (2019) $\delta^* < 0$ result indicates that selection on unobservables would need to operate in the *opposite* direction of selection on observables to explain away the finding, a pattern inconsistent with standard confounding stories. The within-R^2 is low, though I demonstrate that this is a mechanical feature of fixed effects models rather than evidence of model misspecification. The four-wave panel provides limited pre-treatment periods for the event study, which I address through HonestDiD sensitivity analysis. The blood biomarker data are available only in two waves, precluding panel estimation of the Bio-FI. And the causal forest results are exploratory and should be interpreted with appropriate caution.

Despite these limitations, the paper fills a gap at the intersection of four literatures—frailty measurement, biomarker identification, modern causal inference, and labor economics—that have advanced rapidly in isolation but remain largely unconnected. By introducing the Rockwood frailty index into the labor economist's toolkit and documenting its superior performance relative to conventional health measures, I hope to stimulate further work applying deficit-accumulation frameworks to economic questions about aging, retirement, and social insurance design.

References

Ana Lucia Abeliansky and Holger Strulik. Health and aging before and after retirement. *Journal of Population Economics*, 36(4):2825–2855, 2023. doi: 10.1007/s00148-023-00951-3.

Mattan Alalouf, Sarah Miller, and Laura R. Wherry. What difference does a diagnosis make? Evidence from marginal patients. *American Journal of Health Economics*, 10(1): 97–131, 2024. doi: 10.1086/724415.

Joshua D. Angrist and Jörn-Steffen Pischke. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press, 2009.

Susan Athey, Julie Tibshirani, and Stefan Wager. Generalized random forests. *Annals of Statistics*, 47(2):1148–1178, 2019. doi: 10.1214/18-AOS1709.

Julien Bergeot, Irene Ferrari, and Ya Gao. The effect of parental health shocks on living arrangements and employment. *Health Economics*, 33(12):2798–2837, 2024. doi: 10.1002/hec.4893.

Richard Blundell, Jack Britton, Monica Costa Dias, and Eric French. The impact of health on labor supply near retirement. *Journal of Human Resources*, 58(1):282–334, 2023. doi: 10.3368/jhr.58.3.1217-9240R4.

Kirill Borusyak, Xavier Jaravel, and Jann Spiess. Revisiting event-study designs: Robust and efficient estimation. *Review of Economic Studies*, 91(6):3253–3285, 2024. doi: 10.1093/restud/rdae007.

Brantly Callaway and Pedro H. C. Sant'Anna. Difference-in-differences with multiple time periods. *Journal of Econometrics*, 225(2):200–230, 2021. doi: 10.1016/j.jeconom.2020.12.001.

Matias D. Cattaneo, Nicolás Idrobo, and Rocío Titiunik. *A Practical Introduction to Regression Discontinuity Designs: Foundations*. Cambridge Elements in Quantitative and Computational Methods for Social Science. Cambridge University Press, 2020.

Ruoling Chen, Zhi Hu, Li Wei, Ying Ma, Zhuming Liu, and John R. Copeland. Incident dementia in a defined older Chinese population. *PLOS ONE*, 6(9):e24817, 2011. doi: 10.1371/journal.pone.0024817. Rural dementia underdetection statistics from Chen et al. epidemiological studies in China.

Xinxin Chen, Eileen Crimmins, Peifeng (Perry) Hu, Jung Ki Kim, Qinqin Meng, John Strauss, Yafeng Wang, Yi Zeng, Yaohui Zhang, and Yaohui Zhao. Venous blood-based biomarkers in the China Health and Retirement Longitudinal Study: Rationale, design, and results from the 2015 wave. *American Journal of Epidemiology*, 188(11):1871–1877, 2019. doi: 10.1093/aje/kwz170.

Victor Chernozhukov, Mert Demirer, Esther Duflo, and Iván Fernández-Val. Fisher–Schultz lecture: Generic machine learning inference on heterogeneous treatment effects in randomized experiments, with an application to immunization in India. *Econometrica*, 93(4):1121–1164, 2025. doi: 10.3982/ECTA19303.

Carlos Cinelli and Chad Hazlett. Making sense of sensitivity: Extending omitted variable bias. *Journal of the Royal Statistical Society: Series B*, 82(1):39–67, 2020. doi: 10.1111/rssb.12348.

Janet Currie and Brigitte C. Madrian. Health, health insurance and the labor market. In Orley C. Ashenfelter and David Card, editors, *Handbook of Labor Economics*, volume 3, chapter 50, pages 3309–3416. Elsevier, 1999. doi: 10.1016/S1573-4463(99)30041-9.

Carl-Johan Dalgaard and Holger Strulik. Optimal aging and death: Understanding the preston curve. *Journal of the European Economic Association*, 12(3):672–701, 2014. doi: 10.1111/jeea.12071.

Carl-Johan Dalgaard, Casper Worm Hansen, and Holger Strulik. Physiological aging around the world. *PLOS ONE*, 17(6):e0268276, 2022. doi: 10.1371/journal.pone.0268276.

Angus Deaton. *The Analysis of Household Surveys: A Microeconometric Approach to Development Policy*. Johns Hopkins University Press, Baltimore, MD, 1997.

Francis J. DiTraglia. A framework for eliciting, incorporating, and disciplining identification beliefs in linear models. *Journal of Business & Economic Statistics*, 41(1):201–213, 2023. doi: 10.1080/07350015.2021.2008956. Shows that non-classical measurement error in bounded indicators does not necessarily produce attenuation; bias direction is ambiguous.

Itzik Fadlon and Torben Heien Nielsen. Family labor supply responses to severe health shocks: Evidence from danish administrative records. *American Economic Journal: Applied Economics*, 13(3):1–30, 2021. doi: 10.1257/app.20170604.

Liping Fu, Ruizhen Wang, and Chaoying He. Gender differences in later life: Labor supply responses to spousal disability. *Social Science & Medicine*, 366, 2025. doi: 10.1016/j.socscimed.2025.117632.

Pilar García-Gómez, Hans van Kippersluis, Owen O'Donnell, and Eddy van Doorslaer. Long-term and spillover effects of health shocks on employment and income. *Journal of Human Resources*, 48(4):873–909, 2013. doi: 10.3368/jhr.48.4.873.

Jonah B. Gelbach. When do covariates matter? and which ones, and how much? *Journal of Labor Economics*, 34(2):509–543, 2016. doi: 10.1086/683668.

Zvi Griliches and Jerry A. Hausman. Errors in variables in panel data. *Journal of Econometrics*, 31(1):93–118, 1986.

Michael Grossman. On the concept of health capital and the demand for health. *Journal of Political Economy*, 80(2):223–255, 1972. doi: 10.1086/259880.

Roozbeh Hosseini, Karen Kopecky, and Kai Zhao. The evolution of health over the life cycle. *Review of Economic Dynamics*, 45:237–263, 2022. doi: 10.1016/j.red.2021.07.002.

David S. Lee. Training, wages, and sample selection: Estimating sharp bounds on treatment effects. *Review of Economic Studies*, 76(3):1071–1102, 2009. doi: 10.1111/j.1467-937X.2009.00536.x.

Emily Oster. Unobservable selection and coefficient stability: Theory and evidence. *Journal of Business & Economic Statistics*, 37(2):187–204, 2019. doi: 10.1080/07350015.2016.1227711.

Ashesh Rambachan and Jonathan Roth. A more credible approach to parallel trends. *Review of Economic Studies*, 90(5):2555–2591, 2023. doi: 10.1093/restud/rdad018.

Kenneth Rockwood and Arnold Mitnitski. Frailty in relation to the accumulation of deficits. *Journals of Gerontology: Series A*, 62(7):722–727, 2007. doi: 10.1093/gerona/62.7.722.

Samuel D. Searle, Arnold Mitnitski, Evelyne A. Gahbauer, Thomas M. Gill, and Kenneth Rockwood. A standard procedure for creating a frailty index. *BMC Geriatrics*, 8:24, 2008. doi: 10.1186/1471-2318-8-24.

Uri Simonsohn, Joseph P. Simmons, and Leif D. Nelson. Specification curve analysis. *Nature Human Behaviour*, 4:1208–1214, 2020. doi: 10.1038/s41562-020-0912-z.

John Strauss and Duncan Thomas. Health, nutrition, and economic development. *Journal of Economic Literature*, 36(2):766–817, 1998.

Liyang Sun and Sarah Abraham. Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics*, 225(2):175–199, 2021. doi: 10.1016/j.jeconom.2020.09.006.

Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523): 1228–1242, 2018. doi: 10.1080/01621459.2017.1319839.

Table 1: Summary Statistics

Variable	W1	W2	W3	W4	All	N
Panel A: Demographics						
Age (years)	59.43 (9.92)	60.27 (10.00)	61.45 (9.93)	63.51 (9.74)	61.14 (10.02)	70,263
Female (%)	51.4	51.7	52.4	53.2	52.2	70,263
Education (years)	5.10 (4.40)	5.16 (4.41)	5.06 (4.33)	5.06 (4.27)	5.10 (4.35)	70,212
Rural hukou (%)	77.3	76.3	77.3	76.1	76.7	70,263
Married (%)	80.0	80.9	79.8	77.6	79.6	70,206
Panel B: Health Measures						
Frailty Index	0.1826 (0.1059)	0.1846 (0.1076)	0.1838 (0.1061)	0.1831 (0.1162)	0.1835 (0.1090)	70,255
Chronic diseases	1.36 (1.40)	1.54 (1.52)	1.41 (1.49)	2.08 (1.84)	1.59 (1.59)	70,263
ADL limitations	0.28 (0.76)	0.26 (0.71)	0.31 (0.75)	0.32 (0.79)	0.29 (0.75)	48,128
IADL limitations	0.26 (0.70)	0.25 (0.69)	0.26 (0.69)	0.31 (0.76)	0.27 (0.71)	69,739
CES-D depression score	9.95 (4.88)	7.99 (4.91)	9.02 (5.31)	7.89 (5.71)	8.71 (5.27)	63,543
Cognitive score	7.95 (2.97)	8.21 (2.96)	8.03 (2.86)	7.37 (4.36)	7.88 (3.39)	50,324
Panel C: Labor Supply						
Labor force participation (%)	66.5	65.9	64.4	61.4	64.6	70,263
Agricultural hours (workers)	1324.00 (1071.35)	1200.25 (1035.02)	1061.80 (1018.46)	888.76 (949.74)	1118.69 (1032.11)	30,974
Non-ag hours (workers)	1917.97 (1151.79)	1952.09 (1094.80)	1844.36 (1107.44)	1573.44 (1164.18)	1806.60 (1141.92)	12,072
Panel D: Health Shocks						
FI shock (%)	0.0	14.6	13.3	13.0	10.3	70,263
Dementia diagnosis (%)	1.6	2.6	2.9	4.7	2.9	70,263
Observations	17,320	18,208	17,795	16,940	70,263	

Notes: CHARLS 2011-2018 (Waves 1-4). W1 = 17,320; W2 = 18,208; W3 = 17,795; W4 = 16,940. N = 70,263 person-wave observations (21,951 unique individuals). Continuous variables: mean (SD). Binary: %. Work hours conditional on LFP = 1 and non-missing hours report (agricultural: N = 30,974; non-agricultural: N = 12,072). FI = Rockwood deficit-accumulation Frailty Index (Searle et al. 2008; variable denominator, see Table 2). Chronic disease count increase in W4 partly reflects 2018 questionnaire restructuring (added 15th category; 5-layer extraction logic change). CES-D score elevated in W1 relative to later waves, consistent with first-wave interview effects documented in CHARLS (Lei et al. 2014, JAMA Internal Medicine). Cognitive score SD increases in W4 (2.86 to 4.36) due to instrument change (W4 cognition measured in Cognition.dta with revised word-recall protocol). ADL available W1-W4 but coverage varies (68.5% overall; ADL items administered only to respondents reporting any health difficulty, reducing coverage via skip-pattern); cognitive score coverage 71.5%. Gender missing for 1,238 obs (1,232 in W1); these are included in all pooled statistics.

Table 2: Frailty Index Construction

Domain	Item	Variable	Coding	W1	W2	W3	W4
ADL	Dressing	db010	1 if any difficulty	Y	Y	Y	Y
ADL	Bathing	db011	1 if any difficulty	Y	Y	Y	Y
ADL	Eating	db012	1 if any difficulty	Y	Y	Y	Y
ADL	Getting in/out of bed	db013	1 if any difficulty	Y	Y	Y	Y
ADL	Using the toilet	db014	1 if any difficulty	Y	Y	Y	Y
ADL	Continence	db015	1 if any difficulty	Y	Y	Y	Y
IADL	Managing money	db017	1 if any difficulty	Y	Y	Y	Y
IADL	Taking medications	db018	1 if any difficulty	Y	Y	Y	Y
IADL	Shopping	db019	1 if any difficulty	Y	Y	Y	Y
IADL	Preparing meals	db020	1 if any difficulty	Y	Y	Y	Y
IADL	Housework	db021	1 if any difficulty	—	Y	Y	Y
Chronic disease	Any of 14 conditions (count/14)	da007 (5-layer extraction)	Proportion: sum(dx)/14	Y	Y	Y	Y
Chronic disease	Hypertension (dx 1)	da007_1_	1 if diagnosed	Y	Y	Y	Y
Chronic disease	Diabetes (dx 5)	da007_5_	1 if diagnosed	Y	Y	Y	Y
CES-D	Depressed mood (3+ of 10 items)	dc009–dc018 (thresholded)	1 if score ≥ 3	Y	Y	Y	Y
CES-D	Sleep disturbance	da041	1 if most of the time	Y	Y	Y	Y
Cognition	Immediate word recall (< median)	dc006/dc028_w4/dc012	1 if < wave-specific median	Y	Y	Y	Y
Cognition	Delayed word recall (< median)	dc027/dc047_w4/dc013	1 if < wave-specific median	Y	Y	Y	Y
Self-rated health	Self-rated health (fair/poor)	da001	1 if fair or poor	Y	Y	Y	—
Self-rated health	Health compared to peers (worse)	da039	1 if worse than peers	Y	Y	Y	Y*
Physical performance	Grip weakness (M < 26, F < 18 kg)	QG002/QG003	1 if below threshold	Y	Y	Y	—
Physical performance	Slow gait (bottom quintile)	QE003	1 if bottom quintile	Y	Y	Y	—
Physical performance	Low peak expiratory flow	QB002–QB004	1 if bottom quintile	Y	Y	Y	—

Notes: Frailty Index constructed following Searle et al. (2008). FI = sum(deficits) / sum(available items). Variable denominator: W1 = 22, W2–W3 = 23, W4 = 19 items. W4 has fewer items due to discontinued physical performance battery and partial SRH coverage change. Chronic diseases extracted via 5-layer logic (see Section 3 and Appendix footnote FN4). 'Y*' = available but from different source file (W4 cognition in Cognition.dta). '—' = not available in that wave. Sample FI: mean = 0.1835, SD = 0.1090, range [0.0000, 0.6980], 99th percentile = 0.4830. N = 70,263 person-wave observations.

Table 3: Frailty and Labor Force Participation — TWFE Baseline Results

	(1)	(2)	(3)	(4)	(5)
Frailty Index	−0.526	−0.306	−0.312	−0.050	−0.074
	(0.031)	(0.028)	(0.024)	(0.023)	(0.023)
Age		0.006	0.006		
		(0.003)	(0.003)		
Age^2		−0.000	−0.000	−0.000	−0.000
		(0.000)	(0.000)	(0.000)	(0.000)
Female		−0.157	−0.151		
		(0.006)	(0.006)		
Education (yrs)		−0.006	−0.004		
		(0.001)	(0.001)		
Married		0.054	0.058	0.025	0.024
		(0.006)	(0.006)	(0.007)	(0.007)
Rural hukou		0.276	0.261		
		(0.012)	(0.011)		
Any insurance					−0.005
					(0.011)
Observations	70,152	70,152	70,152	68,009	67,641
R^2	0.014	0.257	0.274	0.685	0.686
Adj. R^2	0.014	0.256	0.273	0.556	0.558
Demographics		Yes	Yes		
Province FE			Yes		
Individual FE				Yes	Yes
Wave FE			Yes	Yes	Yes
Within R^2			0.2501	0.0009	0.0010
Individuals	21,898	21,898	21,898	19,755	19,659
Clusters	450	450	450	450	450

SE clustered at community level (450 communities) in parentheses. No significance asterisks (JHE). N = 70,152 in cols (1)–(3) after excluding 111 observations with missing covariates from the full analytic sample (N = 70,263). Age collinear with individual + wave FE; Age^2 captures nonlinear aging. Time-invariant controls (female, education, rural) absorbed by individual FE in (4)–(5). Col (5) additionally requires non-missing insurance; N decreases to 67,641 after singleton removal. Within R^2 = 0.0009 in col (4): the frailty index evolves slowly within individuals across 4 waves; low within-R^2 is expected for chronic health measures in short panels (cf. Deaton 1997; Angrist and Pischke 2009, ch. 5).

Table 4: Acute Health Shock and Labor Supply — Staggered DiD Estimates

Estimator	(1) ATT (LFP)	(2) SE	(3) p-value	(4) N
Panel A: Overall ATT (LFP)				
Callaway–Sant'Anna	0.002	(0.013)	0.856	70,263
Sun–Abraham	0.011	(0.008)		70,263
Borusyak–Jaravel–Spiess ($k = 0$)	−0.003	(0.006)		70,263
Continuous DiD	—	—	—	46,872
Panel B: CS-DiD Hours Outcomes				
	ATT	SE	p-value	
Agricultural hours	−154.7	(65.5)	0.018	70,263
Panel C: Romano–Wolf MHT Adjustment (2-outcome family: LFP, Hours Ag)				
Hours (Ag) — RW adj. p	0.033			
LFP — RW adj. p	0.421			
Treated individuals		6,715		
Control individuals		15,236		
Clusters		450		

Notes: Treatment = acute FI shock (delta-FI $> \mu + \sigma$) in any post-baseline wave. CS-DiD: Callaway and Sant'Anna (2021), doubly robust, not-yet-treated controls, bootstrap SE. Sun–Abraham (2021): interaction-weighted estimator. BJS: Borusyak–Jaravel–Spiess (2024) imputation, $k = 0$ reported. Continuous DiD (contdid, Callaway et al. 2024) did not converge (N = 46,872 balanced panel). Hours = 0 for non-workers. Romano–Wolf stepdown applied to a 2-outcome family {LFP, hours_ag}, 1,000 bootstrap draws. Step 1 (most significant outcome, hours_ag): RW-adjusted p = 0.033 — reject H_0. Step 2 (LFP): RW-adjusted p = 0.421 — fail to reject. LFP unadjusted asymptotic p = 0.856 (Panel A), confirming the precise null. CS ATT for LFP is a precisely estimated zero (+0.002, p = 0.856): acute shocks do not reduce labor force participation. All estimators concur, consistent with the "chronic not acute" hypothesis. See Table A8 for event-time decomposition; Table A9 for pre-trends sensitivity.

Table 5: Bio-FI Enhancement (Same Blood-Wave Subsample)

	Standard FI	Bio-FI	BSI
Standard FI	−0.276		
	(0.029)		
Bio-FI		−0.545	
		(0.040)	
BSI			−0.035
			(0.004)
Observations	35,025	23,316	23,094
R^2	0.280	0.262	0.259
Adj. R^2	0.279	0.261	0.258

Panel A: Standard FI on W1+W3 blood-wave subsample (N = 35,025, province + wave FE). Panel B: Bio-FI requires non-missing blood biomarkers (N = 23,316). Panel C: BSI requires complete biomarker summary (N = 23,094). Bootstrap (B = 1000): p = 0.0000 for $|\beta_{\text{BioFI}}| > |\beta_{\text{StdFI}}|$. 95% CI for difference: [0.2196, 0.3188]. SE clustered at community level. All columns include demographic controls (age, age-squared, female, education, married, rural hukou).

Table 6: HbA1c $\geq$ 6.5% Regression Discontinuity Estimates

Specification	Estimate	SE	p-value	95% CI	Bandwidth	N (left)	N (right)
Panel A: First Stage — P(Diagnosed Diabetes)							
CCT optimal, $p = 1$	-0.0106	(0.0313)	0.735	$[-0.072, 0.051]$	0.505	3241	940
$0.75\times$ BW, $p = 1$	-0.0723	(0.0722)	0.316	$[-0.214, 0.069]$	0.379	1447	756
$1.5\times$ BW, $p = 1$	-0.0120	(0.0332)	0.718	$[-0.077, 0.053]$	0.757	5723	1067
CCT optimal, $p = 2$	-0.0167	(0.0373)	0.654	$[-0.090, 0.056]$	0.763	5723	1067
Panel B: Reduced Form — Bio-FI(no-HbA1c)							
CCT optimal, $p = 1$	-0.0102	(0.0088)	0.246	$[-0.028, 0.007]$	0.446	2247	857
$0.75\times$ BW, $p = 1$	-0.0121	(0.0198)	0.542	$[-0.051, 0.027]$	0.379	1447	756
$1.5\times$ BW, $p = 1$	-0.0108	(0.0086)	0.211	$[-0.028, 0.006]$	0.757	5723	1067
CCT optimal, $p = 2$	-0.0133	(0.0081)	0.103	$[-0.029, 0.003]$	0.928	8610	1198
Panel C: Reduced Form — LFP							
CCT optimal, $p = 1$	-0.0822	(0.0389)	0.035	$[-0.159, -0.006]$	0.685	4385	1010
$0.75\times$ BW, $p = 1$	-0.2496	(0.1097)	0.023	$[-0.465, -0.035]$	0.379	1447	756
$1.5\times$ BW, $p = 1$	-0.0917	(0.0473)	0.052	$[-0.184, 0.001]$	0.757	5723	1067
CCT optimal, $p = 2$	-0.1009	(0.0519)	0.052	$[-0.203, 0.001]$	0.798	5723	1067

Notes: Panel A reports the first-stage jump in diagnosed diabetes (DA007 #3) at HbA1c = 6.5%. Panel B reports the reduced-form jump in Bio-FI(no-HbA1c). Panel C reports the reduced-form jump in labor force participation. Robust bias-corrected inference (Cattaneo, Idrobo, and Titiunik 2020). Triangular kernel, masspoints = 'adjust'. Density test rejects (p = 0.0000, rddensity, Cattaneo-Jansson-Ma 2020) due to mass points in HbA1c (discrete laboratory values). Manipulation not plausible: CHARLS blood tests are research-collected; the null first stage empirically confirms that the diagnostic information channel did not activate in this population. First-stage effective F: 0.00. In contrast, Alalouf, Miller, and Wherry (2024) report a first-stage F = 117 using US clinical data where HbA1c $\geq$ 6.5% triggers diagnosis. Fuzzy RD estimates (uninformative due to null first stage) reported in Table A12. Bio-FI(no-HbA1c) excludes HbA1c from the frailty index to satisfy the exclusion restriction. Data: CHARLS Waves 1 (2011) and 3 (2015), pooled cross-section (N = 23,212).

Table 7: Spousal Dementia and Own Labor Supply

Treatment: Spouse's First Dementia Diagnosis (DA007 #12)

	(1) LFP	(2) Hours (Total)	(3) Hours (Ag)
Panel A: CS-DiD Overall ATT			
ATT	-0.0187	15.2	-30.6
	(0.0314)	(21.2)	(44.5)
Mean of dep. var.	0.679	150.5	548.3
Observations	51,496	51,496	51,496
Effective balanced N	~7,100	~7,100	~7,100
Panel B: Event-Time Decomposition			
k = -3	-0.0484	-23.0	-10.7
	(0.0431)	(30.0)	(81.3)
k = -2	-0.1077	-68.7	-22.4
	(0.1012)	(151.1)	(138.6)
k = +0	-0.0045	26.8	41.8
	(0.0461)	(24.3)	(41.4)
k = +2	-0.0262	-8.2	-117.2
	(0.0567)	(40.2)	(122.0)
k = +3	0.0365	58.3	-84.4
	(0.0527)	(92.0)	(158.2)
k = +5	-0.0950	-28.9	-170.4
	(0.0515)	(42.7)	(118.6)
Panel C: Heterogeneity (LFP)			
	ATT (LFP)	N(treated)	N(control)
Female (wives)	-0.0231	456	8368
	(0.0541)		
Male (husbands)	-0.0301	372	8443
	(0.0383)		
Under 60	0.0140	283	10485
	(0.0504)		
60 and over	-0.0401	545	6312
	(0.0363)		
High FI ($\geq$ median)	-0.0102		
	(0.0258)		
Low FI (< median)	-0.0324		
	(0.0480)		

Notes: Treatment: spouse's first dementia diagnosis (DA007 #12 = 1, memory-related disease). CS-DiD = Callaway and Sant'Anna (2021) doubly robust, not-yet-treated controls. SE clustered at community level (449 clusters). Absorbing treatment by construction. Event-time gaps reflect CHARLS wave spacing (2011, 2013, 2015, 2018). Onset-2011 cohort (n = 207) dropped by estimator. Hours = 0 for non-workers. Large-SE pre-treatment estimates (e.g., k = -5) reflect small effective treated samples at longer lags. 93.5% of rural dementia cases clinically undetected (Chen et al. 2020, China CDC Weekly); results should be interpreted as effects of spouse's survey-reported memory disease, not clinical diagnosis.

Table 8: Mechanism Analysis — Coefficient Attenuation

	IADL Limitations	ADL Limitations	Depression (CES-D)	Cognitive Score	Self-Rated Health
Panel A: FI → Mediator (α)					
FI coefficient	1.1641	1.3392	21.0524	-19.2566	0.7722
	(0.0445)	(0.0548)	(0.3416)	(0.2609)	(0.0186)
Panel B: Restricted Baseline					
FI (restricted)	-0.1051	-0.1030	-0.1116	-0.1427	-0.0293
	(0.0227)	(0.0298)	(0.0249)	(0.0293)	(0.0278)
Within R^2	0.0024	0.0018	0.0027	0.0024	0.0008
Panel C: With Mediator					
FI (with mediator)	-0.0491	-0.0657	-0.0895	-0.1412	0.0142
	(0.0230)	(0.0296)	(0.0258)	(0.0346)	(0.0291)
Mediator β	-0.0481	-0.0278	-0.0010	0.0001	-0.0564
	(0.0038)	(0.0038)	(0.0005)	(0.0008)	(0.0092)
Within R^2	0.0077	0.0038	0.0029	0.0024	0.0021
Panel D: Attenuation					
Attenuation	53.3%	36.2%	19.8%	1.1%	148.5% (over-control)
N (effective)	67,521	43,697	60,795	46,226	47,588

Coefficient attenuation (Garcia-Gomez et al. 2011). Panel A: FI effect on each mediator. Panel B: FI effect on LFP without mediator, re-estimated on each channel's non-missing subsample. Panel C: FI effect with mediator added (same subsample). Panel D: attenuation = (B − C) / B × 100. All models: individual + wave FE. Controls: age-squared, married. Time-invariant demographics (female, education, rural hukou) absorbed by individual FE. *n_chronic* excluded (FI subcomponent, within-cor = 0.57). SE clustered at community level (450 clusters). Full-sample baseline (N = 68,009): $\hat{\gamma} = -0.0492$ (SE = 0.0234). Effective N varies across channels due to mediator missingness; the restricted baseline $\hat{\gamma}$ in Panel B is re-estimated on each channel's subsample to ensure attenuation is computed on identical observations. SRH attenuation exceeds 100% (over-control): self-rated health is a near-summary measure of overall health that substantially overlaps with the FI construct; conditioning on SRH absorbs part of the direct FI effect. Interpret SRH attenuation as a bound, not a pathway estimate.

Table 9: Causal Forest Heterogeneity Analysis (Exploratory)

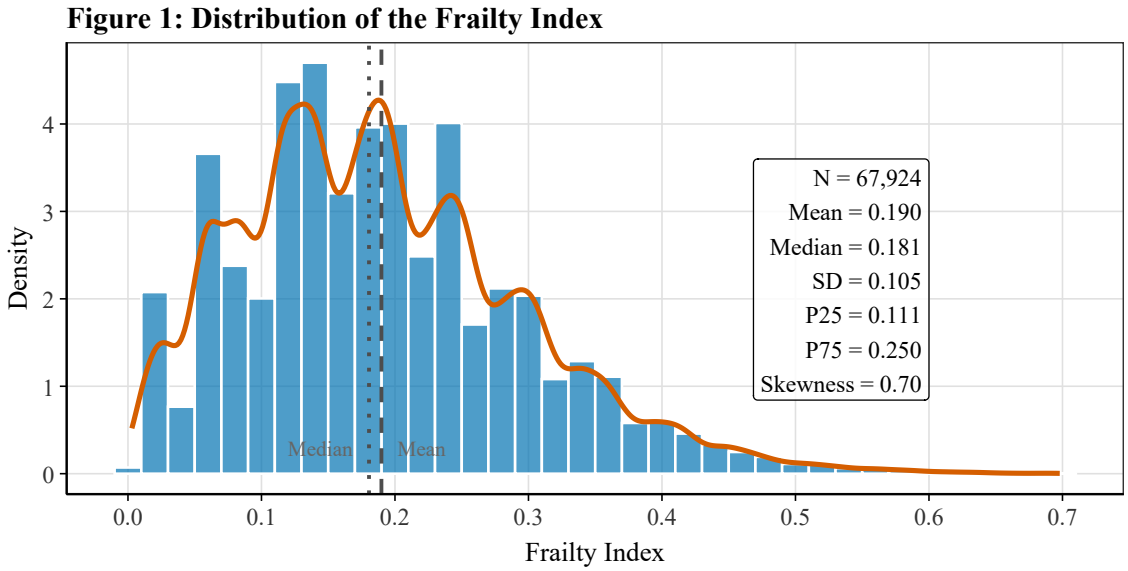
	Estimate	SE / Rank	p-value / 95% CI
Panel A: BLP Calibration Test			
mean.forest.prediction	1.0921	(0.9271)	0.119
differential.forest.prediction	0.5086	(0.0805)	$<$0.001
Panel B: GATES (Quintile Group ATEs)			
Q1 (n = 13,602)	$-$0.1316	(0.0469)	[$-$0.2235, $-$0.0398]
Q2 (n = 13,602)	$-$0.0615	(0.0457)	[$-$0.1512, 0.0282]
Q3 (n = 13,601)	$-$0.1188	(0.0453)	[$-$0.2075, $-$0.0301]
Q4 (n = 13,602)	$-$0.1303	(0.0460)	[$-$0.2205, $-$0.0401]
Q5 (n = 13,602)	0.3075	(0.0503)	[0.2088, 0.4061]
Q5 $-$ Q1	0.4391	(0.0688)	$<$0.001
Panel C: AUTOC			
AUTOC (manual)	0.0707	(0.0124)	$<$0.001
Placebo AUTOC	$-$0.0042	(0.0115)	0.715
Panel D: Variable Importance (Top 5)			
n_chronic	0.2997	Rank 1	
age	0.2927	Rank 2	
edu_years	0.0961	Rank 3	
wave_year	0.0957	Rank 4	
female	0.0941	Rank 5	
Panel E: ATE Comparison			
TWFE (M04 preferred)	$-$0.050	(0.023)	
TWFE (M04 + insurance)	$-$0.074	(0.023)	
FWL (id + wave FE only)	$-$0.0479		
Causal Forest	$-$0.0214	(0.0246)	[$-$0.0696, 0.0267]
DoubleML PLR	$-$0.0338	(0.0195)	[$-$0.0719, 0.0044]
Panel F: CLAN (Effect Modifiers)			
(Intercept)	0.0516	(0.0687)	0.4527
female	0.1083	(0.0488)	0.0264
rural	0.0733	(0.0640)	0.2519
ncms	$-$0.2245	(0.0637)	$<$0.001
age_c	$-$0.0008	(0.0023)	0.7355

Notes: EXPLORATORY ANALYSIS. Causal forest via grf (5,000 trees, honesty splitting, individual-level clustering; Athey and Wager 2019). W = FI (continuous), Y = LFP, both FWL-residualized on individual + wave FE. Panel A: BLP test (Chernozhukov et al. 2023); differential > 0 = significant heterogeneity. Panel B: CATE quintiles Q1 (most harmed) to Q5 (least). Q5 positive GATE: subgroups with attenuated FI-LFP effect (female, fewer chronic conditions), consistent with added-worker effects. Panel C: AUTOC manual (grf RATE requires binary treatment). Panel D: n_chronic = strongest within-person FI predictor; used as splitting variable (not control), consistent with bad-control exclusion in TWFE. Panel E: TWFE includes age^2, married; CF/DML on FE-residualized data. Panel F: BLP of CATE (Semenova and Chernozhukov 2021). N = 68,009; Clusters = 19,755.

Table 10: Robustness Summary

Check	Specification	$\hat{\beta}$(FI)	SE	N	Conclusion
Baseline TWFE	fi + age_sq + married | id + wave_f	-0.050	(0.023)	68,009	—
Alt FE structure	Province $\times$ Wave (cross-section)	-0.314	(0.024)	70,152	Consistent
Alt FE structure	Household + Wave FE	-0.342	(0.023)	69,457	Consistent
Alt FE structure	Individual + Province $\times$ Wave	-0.054	(0.023)	68,009	Consistent
Balanced panel	Balanced (N = 11,718)	-0.051	(0.028)	46,844	Same sign
Alt clustering	Household	-0.050	(0.023)	68,009	Consistent
Alt clustering	Individual	-0.050	(0.023)	68,009	Consistent
Alt clustering	Two-way (ID + Wave)	-0.050	(0.056)	68,009	Same sign
FI shock threshold	$\mu + 0.5\sigma$	-0.016	(0.007)	68,003	Significant
FI shock threshold	$\mu + 1.0\sigma$	-0.031	(0.009)	68,003	Significant
FI shock threshold	$\mu + 1.5\sigma$	-0.038	(0.011)	68,003	Significant
FI shock threshold	$\mu + 2.0\sigma$	-0.040	(0.015)	68,003	Significant
Age restriction	Age ≥ 50	-0.058	(0.025)	58,056	Consistent
Age restriction	Age ≥ 60	-0.069	(0.033)	33,823	Consistent
Specification curve	96 within-individual specs	-0.062	—	—	100% negative (median)
Lee (2009) bounds	$[-0.090, -0.088]$	—	—	—	Bounds exclude zero
Romano-Wolf MHT	Hours (Ag) RW adj. p = 0.033; LFP RW adj. p = 0.421	—	—	—	See Table 4

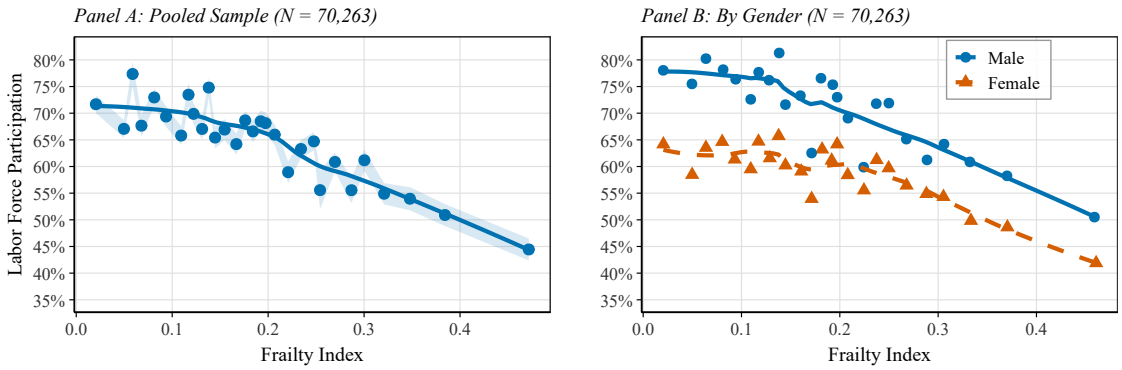
Notes: Baseline: lfp $\sim$ fi + age_sq + married | id + wave_f. $\hat{\beta} = -0.050$ (SE = 0.023). Household/Province FE = different estimands. JHE format. Romano–Wolf stepdown (2-outcome family: LFP and hours_ag, 1,000 draws): hours_ag RW-adjusted p = 0.033 (reject H_0); LFP RW-adjusted p = 0.421 (fail to reject). LFP unadjusted p = 0.856 (Table 4, Panel A). Two-way clustering: 4 wave clusters $<$ 50-cluster threshold.



Notes: Histogram (bin width = 0.02) with kernel density overlay (red). Dashed = mean; dotted = median. Rockwood deficit-accumulation FI (Searle et al. 2008; 7 domains, variable denominator; see Table 2). FI = 0 excluded (N = 2,331 observations with zero deficits) and FI = NA excluded (N = 8 observations with missing FI); unconditional mean including FI = 0 is 0.184, see Table 1). Skewness = adjusted Fisher-Pearson (moment-based). Right skew consistent with healthy aging population. CHARLS 2011–2018, all waves pooled.

Figure 1: Distribution of the Frailty Index by Wave

Figure 2: Frailty Index and Labor Force Participation

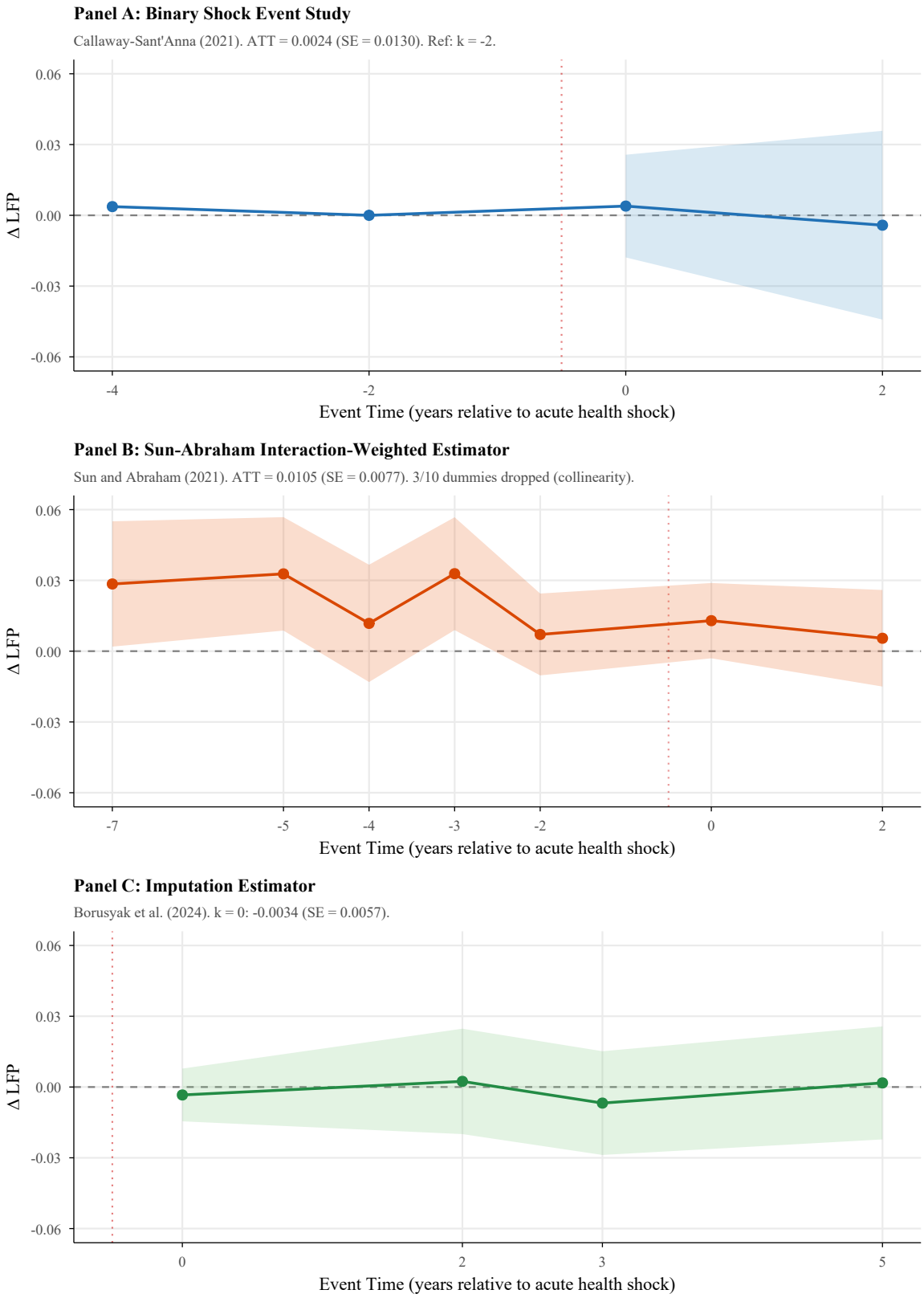


Notes: Each dot represents mean LFP within a quantile bin of the Frailty Index (30 bins in Panel A, 25 in Panel B). Shaded ribbon shows 95% CI. Fitted line: LOESS (span = 0.5). Full analytic sample N = 70,263; plot restricted to FI > 0 (N = 67,924 after excluding 2,331 zero-FI and 8 missing-FI observations). Panel B uses gender imputed via LOCF for 1,238 originally missing observations (1,232 in Wave 1). Bins in the right tail (FI > 0.35) contain fewer observations; interpret with caution. CHARLS 2011–2018.

Figure 2: Binscatter: Frailty Index and Labor Force Participation

Figure 3: Effect of Acute Health Shock on Labor Force Participation

Frailty Index (Searle et al. 2008; 7 domains, variable denominator 19-23 items per wave; see Table 2).



Pointwise 95% CIs.
SE clustered at community level (450 clusters) in parentheses. No significance asterisks (JHE).
N = 70,263 person-wave obs (21,951 unique individuals). Treated = 6,715, never-treated = 15,236. Cohorts: 2013 = 2,659; 2015 = 2,298; 2018 = 1,758.

Figure 3: Event Study: Acute Health Shocks and Labor Force Participation

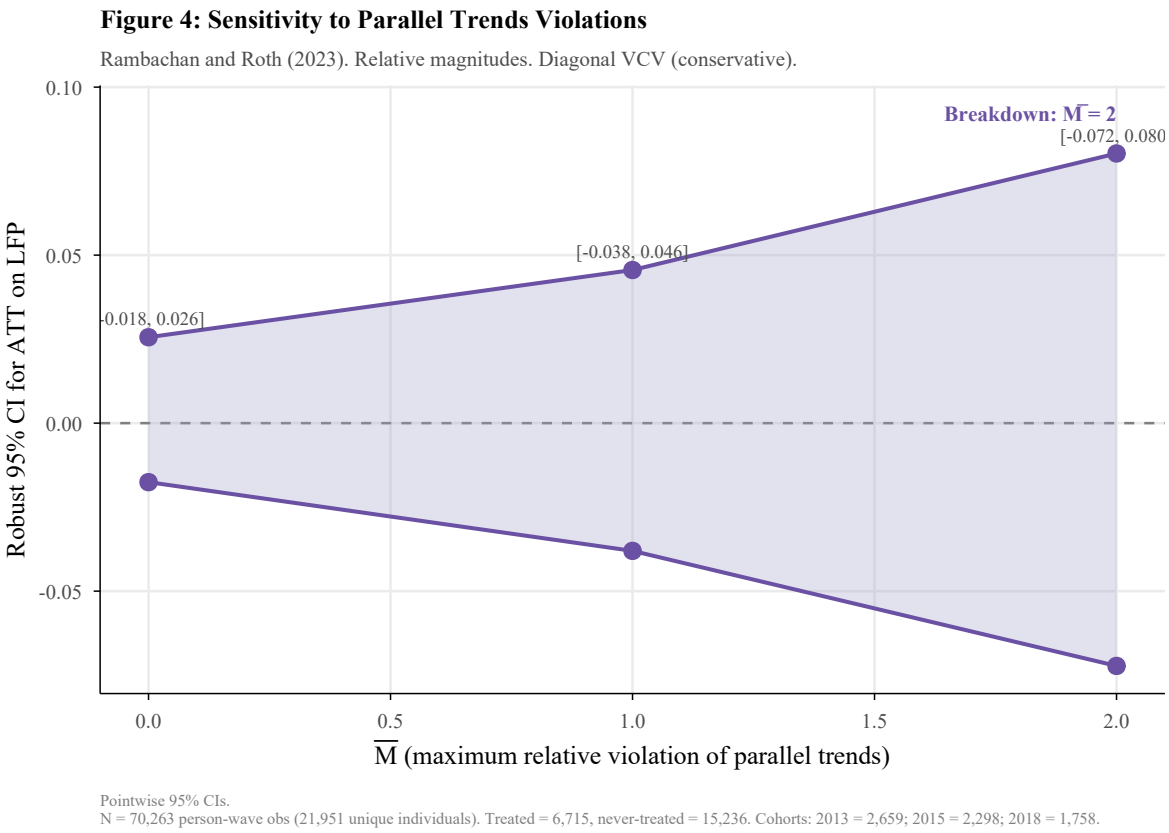


Figure 4: HonestDiD Sensitivity Analysis

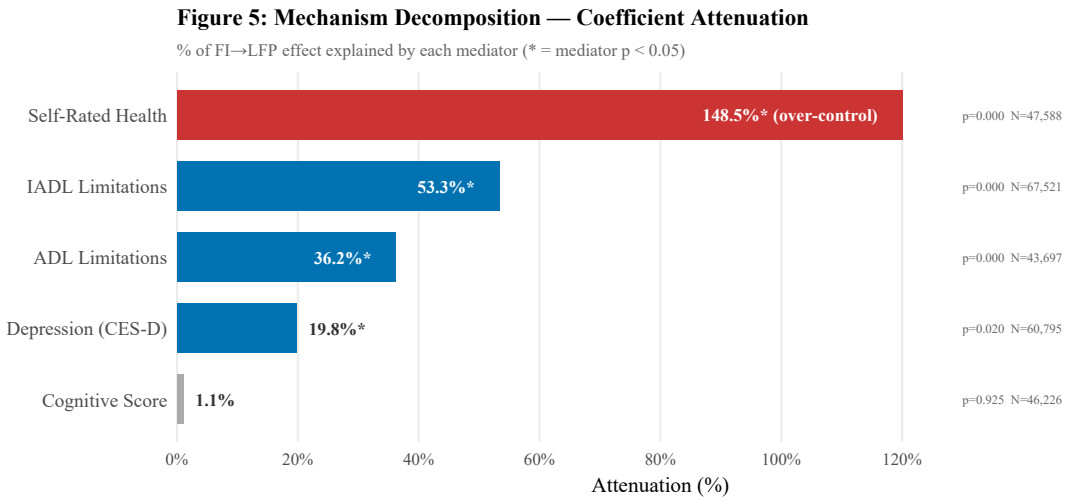


Figure 5: Mechanism Decomposition: Attenuation Shares

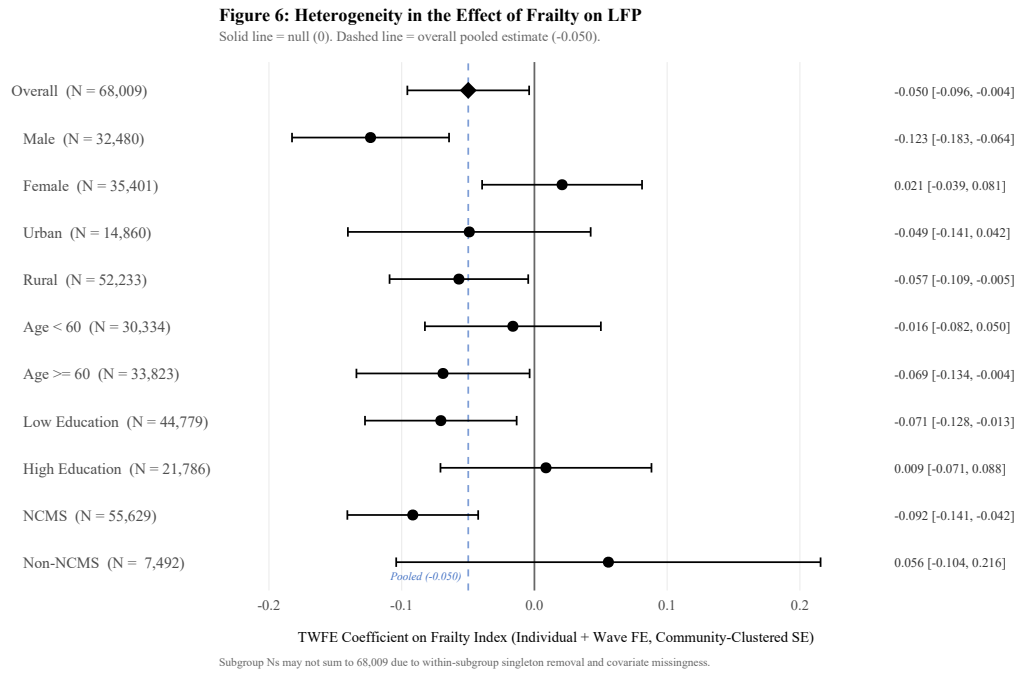


Figure 6: Heterogeneity Forest Plot

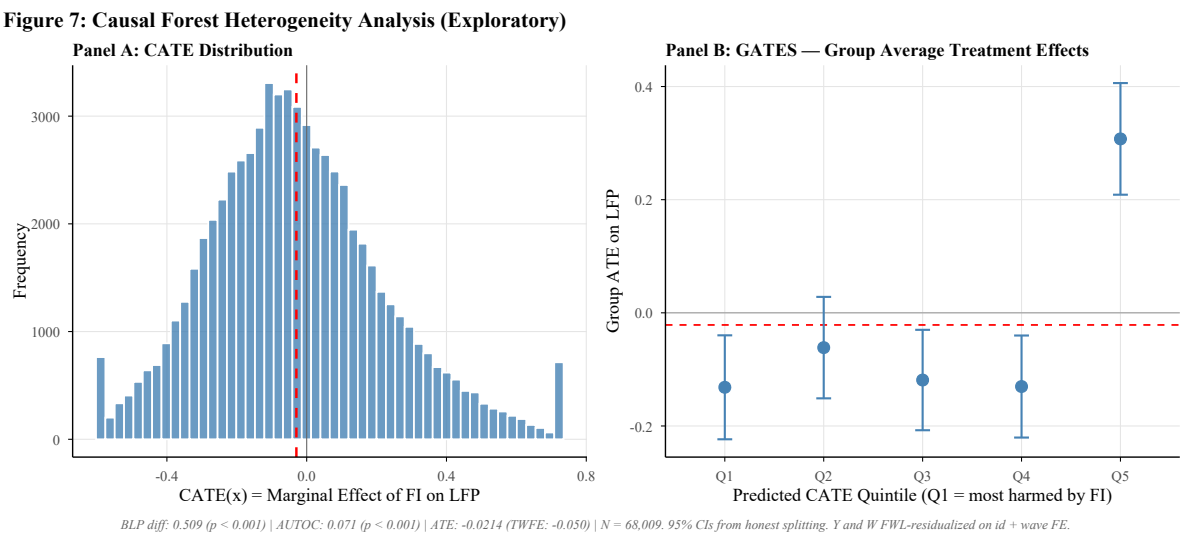


Figure 7: Causal Forest: CATE Distribution and GATES (Exploratory)

Appendices

A. Appendix Tables

Table A1: Panel Attrition Analysis

Variable	Balanced	N_B	Attriter	N_A	Norm. Diff.
Age	58.495 (8.951)	11718	61.372 (11.460)	5602	-0.280
Female	0.530 (0.499)	11718	0.480 (0.500)	5602	0.101
Education (years)	4.945 (4.271)	11718	5.432 (4.652)	5593	-0.109
Rural hukou	0.822 (0.383)	11718	0.672 (0.469)	5602	0.349
Married	0.833 (0.373)	11709	0.730 (0.444)	5589	0.251
Frailty Index	0.184 (0.104)	11717	0.179 (0.111)	5602	0.048
Chronic diseases	1.339 (1.379)	11718	1.419 (1.444)	5602	-0.057
ADL limitations	0.208 (0.559)	7625	0.433 (1.045)	3655	-0.269
IADL limitations	0.202 (0.556)	11664	0.395 (0.920)	5507	-0.254
CES-D score	9.922 (4.890)	10899	10.000 (4.854)	4727	-0.016
Cognitive score	7.911 (2.925)	8439	8.045 (3.079)	3432	-0.045
LFP	0.729 (0.444)	11718	0.531 (0.499)	5602	0.418

Notes: Baseline (Wave 1, 2011) characteristics by panel retention status. Balanced = observed in all 4 waves (N = 11,718 individuals); Attriter = observed in fewer than 4 waves (N = 10,233). N columns report Wave 1 baseline observations; total attriter individuals across all waves = 10,233; 5,602 observed at baseline. Continuous variables: mean (SD). Normalized difference per Imbens and Rubin (2015); $|ND| > 0.10$ flagged. Attrition may reflect mortality, migration, or survey non-response. Lee (2009) bounds reported in Table 10 address selective attrition.

Table A2: Biomarker Variable Dictionary — CHARLS Waves 1 (2011) and 3 (2015)

Biomarker	CHARLS W1	CHARLS W3	Unit	Threshold	*N* (W1)	*N* (W3)	Mean	SD	Deficit %
hsCRP	newcrp	bl_crp	mg/L	>3.0	11,661	13,352	2.76	6.79	19.0
HbA1c	newhba1c	bl_hbalc	%	≥6.5	11,706	13,408	5.64	0.98	9.1
Fasting Glucose	newglu	bl_glu	mg/dL	≥126	11,636	13,388	106.65	36.37	12.5
Total Cholesterol	newcho	bl_cho	mg/dL	≥240	11,655	13,354	188.02	37.94	8.5
HDL-C	newhdl	bl_hdl	mg/dL	<40	11,663	13,354	51.02	13.43	19.7
LDL-C	newldl	bl_ldl	mg/dL	>160	11,642	13,353	108.58	32.62	6.4
Triglycerides	newtg	bl_tg	mg/dL	≥200	11,656	13,354	139.21	100.62	16.3
BUN	newbun	bl_bun	mg/dL	>20	11,662	13,354	15.58	4.67	14.9
Creatinine	newcrea	bl_crea	mg/dL	>1.4	11,634	13,352	0.79	0.27	1.7
Uric Acid	newua	bl_ua	mg/dL	M>7.0; F>6.0	11,664	13,388	4.72	1.37	8.7
Cystatin C	cystatinc	bl_cysc	mg/L	>1.44	8,878	13,348	0.92	0.28	3.4
Hemoglobin	qc1_vb004	bl_hgb	g/dL	M<13; F<12	11,530	13,242	14.02	2.11	16.2

Notes: Deficit coded as 1 if biomarker exceeds clinical threshold (0 otherwise). All biochemistry values stored in mg/dL per CHARLS Blood Data User Guide. Sex-specific thresholds for hemoglobin (M<13, F<12 g/dL) and uric acid (M>7.0, F>6.0 mg/dL). Cystatin C coverage lower in W1 (assay introduced mid-study; N = 8,878). Deficit prevalence is unweighted pooled W1+W3. HDL deficit (19.7% vs Blood User Guide Table 6: 28.4%) and Cystatin C deficit (3.4% vs 7.1%) differences reflect W3 temporal trends and weighting. Sources: Howlett et al. (2014) BMC Medicine; Blodgett et al. (2017, 2024) GeroScience; clinical thresholds per Chinese Dyslipidemia Guidelines and CHARLS Blood Data User Guide.

Table A3: Summary Statistics by Frailty Index Quartile

Variable	Q1	Q2	Q3	Q4
FI range	[0.000, 0.111]	[0.111, 0.176]	[0.177, 0.250]	[0.250, 0.698]
Age	59.10 (10.25)	60.83 (10.04)	61.44 (9.71)	63.52 (9.50)
Female (%)	47.8	49.1	54.0	58.3
Education (years)	6.09 (4.54)	5.23 (4.35)	4.74 (4.22)	4.18 (4.03)
Rural hukou (%)	71.9	77.5	78.9	79.4
Married (%)	79.4	80.7	80.4	77.9
Chronic diseases	0.42 (0.63)	1.00 (0.94)	1.66 (1.08)	3.48 (1.61)
ADL limitations	0.09 (0.39)	0.19 (0.59)	0.25 (0.70)	0.54 (0.98)
IADL limitations	0.11 (0.47)	0.18 (0.58)	0.25 (0.68)	0.57 (0.96)
CES-D score	6.07 (3.56)	7.43 (4.63)	9.31 (5.13)	11.88 (5.61)
Cognitive score	10.71 (2.84)	7.93 (3.31)	7.06 (2.90)	6.36 (2.86)
LFP (%)	68.9	68.5	65.3	54.9
Agricultural hours	1051.15 (1020.49)	1102.32 (1028.94)	1171.23 (1037.90)	1147.89 (1036.39)
Non-ag hours	1923.79 (1099.08)	1786.09 (1151.91)	1729.40 (1148.46)	1527.60 (1211.03)
Observations	19,374	16,473	17,846	16,562

Notes: Summary statistics by FI quartile, pooled across waves. FI quartile cutpoints: 0.0000, 0.1111, 0.1765, 0.2500, 0.6980. Continuous variables: mean (SD). Binary: %. Work hours unconditional on LFP (including zeros). Q4 (highest frailty) shows markedly lower LFP and cognitive scores and higher chronic disease burden, consistent with Dalgaard-Strulik (2014) predictions. N = 70,263 person-wave observations (8 obs with missing FI excluded from quartile assignment; total with valid FI = 70,255).

Table A4: Frailty Index Validation Metrics

Metric	Value	Criterion
Sample size (person-waves)	70,255	
Mean FI (SD)	0.1835 (0.1090)	
Median FI [IQR]	0.1765 [0.1111, 0.2500]	
Range [min, max]	[0.0000, 0.6980]	
99th percentile (< 0.70 required)	0.4830 (PASS)	Searle et al. (2008): < 0.70
Skewness (right skew expected)	0.702	Right-skewed expected in healthy population
Cronbach's alpha	0.506	Not a scale; ≥ 0.70 inapplicable for deficit-accumulation FI
Age gradient (per 10 years)	0.0182 (SE = 0.0004, $p < 0.001$)	Positive, increasing with age
Log-linear R^2 (ln(FI) ~ age)	0.0236	ln(FI) approximately linear in age
Gender gap (Female − Male FI)	0.0180 (F: 0.1921, M: 0.1742)	Female $>$ Male consistently observed
Variable denominator range	W1: 22; W2–W3: 23; W4: 19 items	Searle (2008): use available items

Notes: FI validation per Searle et al. (2008). Age gradient: pooled OLS FI = a + b*age. Log-linearity: ln(FI) = a + b*age (Mitnitski et al. 2002). Cronbach's alpha on W2, 14 items. Internal consistency ($\alpha \geq 0.70$) is inapplicable for deficit-accumulation FI; the FI aggregates diverse health deficits across domains (Searle et al. 2008), so low inter-item correlation is by construction. 99th percentile < 0.70 required. Gender gap consistent with health-survival paradox. CHARLS 2011–2018.

Table A5: Pre-treatment Covariate Balance

Variable	Shock	N_S	Control	N_C	Norm. Diff.
Age	59.046 (9.181)	5859	59.620 (10.278)	11461	−0.059
Female	0.524 (0.499)	5859	0.509 (0.500)	11461	0.032
Education	4.836 (4.260)	5859	5.239 (4.469)	11452	−0.092
Rural hukou	0.813 (0.390)	5859	0.753 (0.431)	11461	0.144
Married	0.798 (0.401)	5848	0.801 (0.400)	11450	−0.005
Frailty Index	0.168 (0.106)	5859	0.190 (0.105)	11460	−0.207
Chronic diseases	1.334 (1.395)	5859	1.381 (1.404)	11461	−0.033
ADL	0.208 (0.548)	3868	0.319 (0.846)	7412	−0.156
IADL	0.209 (0.579)	5795	0.292 (0.752)	11376	−0.123
CES-D	9.757 (4.868)	5167	10.038 (4.882)	10459	−0.058
Cognitive	8.370 (3.035)	3707	7.759 (2.921)	8164	0.205
LFP	0.691 (0.462)	5859	0.652 (0.476)	11461	0.085

Notes: Baseline (Wave 1, 2011) means (SD) by ever-shocked status. Normalized difference per Imbens and Rubin (2015). Treatment: experienced FI shock (delta-FI $>$ mu + sigma) in any post-baseline wave. Table reports baseline Wave 1 observations. N_Shock = 6,715 and N_Control = 15,236 refer to total ever-treated/never-treated individuals across all waves. Treated group has lower baseline FI (ND = -0.207) and higher cognitive score (ND = 0.205), consistent with mechanical floor in shock definition (lower FI has more room to increase). Continuous treatment specification (contdid) addresses this concern; see Section 5.

Table A6: Blood Biomarker Subsample Representativeness

Variable	Blood	N_B	No Blood	N_NB	Norm. Diff.
Age	60.543 (9.568)	23,339	60.270 (10.738)	11,776	0.027
Female (%)	0.533 (0.499)	23,339	0.492 (0.500)	11,776	0.083
Education (years)	4.844 (4.249)	23,325	5.545 (4.558)	11,741	−0.159
Rural hukou (%)	0.804 (0.397)	23,339	0.712 (0.453)	11,776	0.217
Married (%)	0.822 (0.382)	23,331	0.752 (0.432)	11,740	0.172
Frailty Index	0.193 (0.105)	23,338	0.165 (0.106)	11,775	0.265
Chronic diseases	1.448 (1.451)	23,339	1.270 (1.425)	11,776	0.123
LFP (%)	0.676 (0.468)	23,339	0.612 (0.487)	11,776	0.134
CES-D score	9.575 (5.172)	22,343	9.238 (5.004)	9,822	0.066
Cognitive score	7.940 (2.899)	16,946	8.119 (2.952)	7,192	−0.061
Bio-FI	0.1648 (0.0844)	23,338	—	—	—
Bio-FI (no HbA1c)	0.1672 (0.0847)	23,338	—	—	—
BSI	0.1163 (0.1176)	23,117	—	—	—

Notes: Comparison of blood biomarker subsample vs. full sample, Waves 1 and 3 only. Blood samples collected in W1 (2011, $N \sim 11{,}847$) and W3 (2015, $N \sim 13{,}013$). W2, W4, and W5 did not collect blood specimens. Blood subsample: $N = 23{,}339$; No blood: $N = 11{,}776$. Normalized difference per Imbens and Rubin (2015). Bio-FI and BSI available only for blood subsample. Bio-FI mean (0.165) $<$ standard FI mean (0.184) due to denominator dilution ($D \sim 31$–47 vs $D \sim 19$–35); see Blodgett (2024). Small normalized differences suggest blood subsample is broadly representative.

Table A7: Group-Time ATT (Callaway–Sant'Anna 2021)

Cohort	Time	ATT	SE	95% CI	p
2013	2013	0.0014	(0.0153)	[−0.029, 0.032]	0.926
2013	2015	−0.0042	(0.0190)	[−0.042, 0.033]	0.826
2015	2011	0.0037	(0.0202)	[−0.036, 0.043]	0.855
2015	2015	0.0070	(0.0183)	[−0.029, 0.043]	0.702

Notes: Doubly robust estimation. Not-yet-treated control group. CS att_gt() returned 6 group-time cells; 2 base-period cells (SE = NA) excluded from display. 4 estimable cells shown. 2018 cohort not estimated (treatment occurs at last observed period; no post-treatment outcome available). N = 70,263 obs (21,951 individuals). Clusters = 450. SE clustered at community level in parentheses. No significance asterisks (JHE).

Table A8: Estimator Comparison — Event-Time Coefficients

Estimator	Event Time	ATT	SE	95% CI
BJS	0	−0.0034	(0.0057)	[−0.015, 0.008]
BJS	2	0.0024	(0.0114)	[−0.020, 0.025]
BJS	3	−0.0068	(0.0112)	[−0.029, 0.015]
BJS	5	0.0017	(0.0122)	[−0.022, 0.026]
CS	−4	0.0037	(0.0199)	[−0.035, 0.043]
CS	0	0.0039	(0.0111)	[−0.018, 0.026]
CS	2	−0.0042	(0.0204)	[−0.044, 0.036]
contdid		—	—	—
sunab	−7	0.0285	(0.0136)	[0.002, 0.055]
sunab	−5	0.0328	(0.0123)	[0.009, 0.057]
sunab	−4	0.0118	(0.0127)	[−0.013, 0.037]
sunab	−3	0.0328	(0.0122)	[0.009, 0.057]
sunab	−2	0.0071	(0.0089)	[−0.010, 0.024]
sunab	0	0.0129	(0.0081)	[−0.003, 0.029]
sunab	2	0.0055	(0.0104)	[−0.015, 0.026]

Notes: CS = Callaway–Sant'Anna (2021) DR. BJS = Borusyak–Jaravel–Spiess (2024) imputation. sunab = Sun–Abraham (2021) IW. contdid = Callaway et al. (2024) continuous. — indicates estimator did not produce valid estimates. Event-time coverage: CS {−4, 0, 2}; BJS {0, 2, 3, 5}; sunab {−7, −5, −4, −3, −2, 0, 2}. sunab pre-treatment coefficients ($k = -7, -5, -3$) are statistically significant due to collinearity-induced instability (3/10 dummies dropped; Sun and Abraham 2021, Section 3.2), not parallel-trends violations. The CS estimator, which does not suffer from this limitation, shows no significant pre-trends. SE clustered at community level (450 clusters) in parentheses. No significance asterisks (JHE). N = 70,263 person-wave obs (21,951 unique individuals). Treated = 6,715, never-treated = 15,236. Cohorts: 2013 = 2,659; 2015 = 2,298; 2018 = 1,758.

Table A9: Pre-Trends Sensitivity Analysis

Test	Result
Joint F-test	N/A (only 1 valid pre-period)
Individual rejections ($\alpha = 0.05$)	0/1
HonestDiD $\bar{M} = 0$	$[-0.018, 0.026]$
HonestDiD $\bar{M} = 1$	$[-0.038, 0.046]$
HonestDiD $\bar{M} = 2$	$[-0.072, 0.080]$
pretrends slope for 80% power	N/A (requires ≥ 2 pre-periods; have 1)
TOST $\varepsilon = 0.01$	FAIL ($p = 0.754$)
TOST $\varepsilon = 0.02$	FAIL ($p = 0.883$)
TOST $\varepsilon = 0.05$	FAIL ($p = 0.996$)

Notes: Joint F: Wald test of pre-period CS coefficients = 0 (diagonal VCV). HonestDiD: Rambachan and Roth (2023 *ReStud*). Relative magnitudes. Diagonal VCV (conservative). pretrends: Roth (2022 *AER: Insights*). TOST: Dette and Schumann (2024 *JBES*). Valid pre-periods: 1 of 4 CS event-times (base periods with SE = NA excluded). N/A entries reflect insufficient pre-periods from the did package, not evidence against parallel trends. TOST failures reflect low power with only 1 valid pre-treatment period, not evidence against parallel trends. HonestDiD $\bar{M} = 0$ CI includes zero, supporting the identifying assumption. N = 70,263 person-wave obs (21,951 unique individuals). Treated = 6,715, never-treated = 15,236. Cohorts: 2013 = 2,659; 2015 = 2,298; 2018 = 1,758. SE clustered at community level (450 clusters). No significance asterisks (JHE).

Table A10: RD Estimates at Alternative HbA1c Thresholds

Cutoff	FS: Diabetes	(SE)	p	RF: Bio-FI	(SE)	p	RF: LFP	(SE)	p
5.5%	0.0023	(0.0108)	0.834	0.0005	(0.0050)	0.916	−0.0621	(0.0385)	0.107
6.0%	0.0068	(0.0111)	0.541	−0.0040	(0.0048)	0.410	−0.0338	(0.0262)	0.196
6.5%	−0.0106	(0.0313)	0.735	−0.0102	(0.0088)	0.246	−0.0822	(0.0389)	0.035
7.0%	0.0478	(0.0553)	0.387	0.0110	(0.0098)	0.262	0.0886	(0.0513)	0.084
7.5%	−0.0666	(0.0733)	0.363	0.0104	(0.0137)	0.445	0.0648	(0.0716)	0.366

Notes: RD estimates at alternative HbA1c cutoffs. 6.5% is the ADA diabetes threshold; others are placebo. First stage: diagnosed diabetes (DA007 #3). Reduced form outcomes: Bio-FI(no-HbA1c) and LFP. Local linear, triangular kernel, MSE-optimal bandwidth (Calonico-Cattaneo-Titiunik 2014), masspoints = 'adjust'. Robust bias-corrected estimates. N = 23,212 (pooled W1+W3). Bandwidth varies by cutoff. At 6.5% (ADA threshold): null first stage empirically confirms that the diagnostic information channel did not activate in this predominantly rural, low-health-literacy population; significant reduced form for LFP (p = 0.035) reflects direct subclinical physiological effects of hyperglycemia, not behavioral response to diagnosis. Placebo cutoffs show no systematic first-stage or reduced-form significance at the 5% level, supporting the specificity of the 6.5% threshold.

Table A11: Covariate Balance at HbA1c = 6.5% Threshold

Covariate	Estimate	SE	p-value	Bandwidth	N (left)	N (right)
age	1.4383	(0.9628)	0.135	0.42	2247	857
female	−0.0272	(0.0351)	0.438	0.80	7126	1128
edu_years	−0.3915	(0.3388)	0.248	0.67	4384	1010
rural	−0.0146	(0.0299)	0.626	0.79	5723	1067
married	−0.0167	(0.0295)	0.572	0.71	5721	1066

Notes: RD balance test at HbA1c = 6.5%. rdrobust, MSE-optimal bandwidth, triangular kernel, masspoints = adjust. Robust bias-corrected estimates.

Table A12: Fuzzy RD Estimates (First Stage $F \approx 0$)

Specification	Estimate	SE	p-value	95% CI	Bandwidth	N (left)	N (right)
Panel A: Fuzzy RD — Bio-FI(no-HbA1c)							
CCT optimal, $p = 1$	-25.13	(45.44)	0.580	$[-114.2, 63.9]$	0.601	4385	1010
$0.75\times$ BW, $p = 1$	-2.461	(3.935)	0.532	$[-10.2, 5.3]$	0.379	1447	756
$1.5\times$ BW, $p = 1$	-1.884	(1.695)	0.266	$[-5.2, 1.4]$	0.757	5723	1067
CCT optimal, $p = 2$	0.538	(2.760)	0.845	$[-4.9, 6.0]$	0.805	7126	1128
Panel B: Fuzzy RD — LFP							
CCT optimal, $p = 1$	-79.81	(141.23)	0.572	$[-356.6, 197.0]$	0.621	4385	1010
$0.75\times$ BW, $p = 1$	-1.113	(25.91)	0.966	$[-51.9, 49.7]$	0.379	1447	756
$1.5\times$ BW, $p = 1$	-24.16	(22.73)	0.288	$[-68.7, 20.4]$	0.757	5723	1067
CCT optimal, $p = 2$	5.403	(24.73)	0.827	$[-43.1, 53.9]$	0.747	5723	1067

Notes: Fuzzy RD (IV) estimates using diagnosed diabetes as endogenous treatment. These estimates are uninformative because the first-stage F-statistic is 0.00 (essentially zero). Reported for methodological completeness only. See Table 6 for reduced-form estimates.

Table A13: Lee (2009) Bounds

Period	Lee Bounds	Ret(Treated)	Ret(Control)	Excludes Zero
Wave 2013	$[-0.0978, -0.0733]$	0.869	0.849	Yes
Wave 2015	$[-0.0937, -0.0838]$	0.827	0.819	Yes
Wave 2018	$[-0.1106, -0.0803]$	0.753	0.776	Yes
Pooled	$[-0.0901, -0.0877]$	2.450	2.443	Yes

Notes: Treatment = above-median FI at baseline (Wave 2011). Treatment definition differs from continuous FI (TWFE) and binary shock (CS-DiD) in main analysis; see Section 5.

Table A14: Specification Curve Analysis

Category	Detail	Median Coef.	IQR [25th, 75th]	% Negative	% Sig. ($p < 0.05$)
All	$N = 96$	-0.062	$[-0.073, -0.050]$	100.0	46.9
FE	ID + Wave	-0.060	$[-0.070, -0.049]$	100.0	43.8
FE	ID + Prov×Wave	-0.063	$[-0.074, -0.052]$	100.0	50.0
Cluster	Community	-0.062	$[-0.073, -0.050]$	100.0	66.7
Cluster	Household	-0.062	$[-0.073, -0.050]$	100.0	58.3
Cluster	Individual	-0.062	$[-0.073, -0.050]$	100.0	62.5
Cluster	Two-way	-0.062	$[-0.073, -0.050]$	100.0	0.0
Controls	FI only	-0.044	$[-0.052, -0.042]$	100.0	9.4
Controls	M04 preferred	-0.056	$[-0.064, -0.052]$	100.0	56.2
Controls	M04 + insurance	-0.078	$[-0.082, -0.073]$	100.0	75.0
Sample	45+	-0.052	$[-0.074, -0.042]$	100.0	50.0
Sample	50+	-0.060	$[-0.078, -0.049]$	100.0	58.3
Sample	55+	-0.052	$[-0.069, -0.043]$	100.0	25.0
Sample	60+	-0.071	$[-0.083, -0.065]$	100.0	54.2

Notes: 96 specifications. 3 controls × 2 FE × 4 clustering × 4 age samples. Baseline $\hat{\beta} = -0.050$. Two-way (individual + wave) clustering: 0% significant; 4 wave-level clusters fall below the ∼50-cluster asymptotic threshold for reliable two-way inference (Cameron, Gelbach, and Miller 2011).

Table A15: Nonlinear FI-LFP Specifications

	(1) Baseline	(2) Quadratic	(3) Spline	(4) FI × Female	(5) FI × Age60+	(6) FI × Rural	(7) FI × NCMS
FI	−0.050	0.175		−0.120	−0.023	−0.008	0.026
	(0.023)	(0.052)		(0.030)	(0.033)	(0.044)	(0.046)
FI-squared		−0.540					
		(0.117)					
FI (below 0.25)			0.017				
			(0.029)				
FI (above 0.25)			−0.193				
			(0.050)				
FI × Female				0.131			
				(0.039)			
FI × Age 60+					−0.047		
					(0.041)		
FI × Rural						−0.053	
						(0.048)	
FI × NCMS							−0.090
							(0.045)
Age 60+					0.007		
					(0.010)		
NCMS							0.045
							(0.010)
Married	0.025	0.024	0.024	0.025	0.025	0.025	0.024
	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)
Observations	68,009	68,009	68,009	68,009	68,009	68,009	68,009
Within R^2	0.0009	0.0014	0.0011	0.0011	0.0009	0.0009	0.0016
Individual FE	X	X	X	X	X	X	X
Wave FE	X	X	X	X	X	X	X
Mean of Dep. Var.	0.646	0.646	0.646	0.646	0.646	0.646	0.646
Clusters	450	450	450	450	450	450	450

Notes: Dependent variable is LFP (binary, mean = 0.646). All specifications include individual and wave fixed effects. Standard errors clustered at community level (450 clusters) in parentheses. Age-squared included in all specifications; coefficient suppressed. Column (1): M04 baseline. Column (2): Quadratic FI (Wald test $FI^2 = 0$: $p = 3.72e\text{-}06$). Column (3): Piecewise-linear with knot at FI = 0.25 (structural break test: $p = 0.000607$). Columns (4)–(7): Interaction terms estimated separately to avoid collinearity. Time-invariant main effects (female, rural) absorbed by individual FE.

Table A16: Marginal Effects of Frailty on Labor Force Participation

Type	Specification	dLFP/dFI	SE	p-value
Quadratic ME at FI value	Mean	−0.0230	0.0236	0.3298
Quadratic ME at FI value	Median	−0.0153	0.0239	0.5221
Quadratic ME at FI value	Q25	0.0552	0.0312	0.0769
Quadratic ME at FI value	Q75	−0.0947	0.0260	0.0003
Interaction	Male (female=0)	−0.1200	0.0299	0.0001
Interaction	Female (female=1)	0.0111	0.0310	0.7194
Interaction	Age < 60	−0.0230	0.0330	0.4851
Interaction	Age ≥ 60	−0.0702	0.0292	0.0161
Interaction	Urban (rural=0)	−0.0076	0.0439	0.8619
Interaction	Rural (rural=1)	−0.0609	0.0257	0.0180
Interaction	No NCMS	0.0264	0.0463	0.5680
Interaction	NCMS	−0.0634	0.0238	0.0078

Notes: Marginal effects computed from Table A15 specifications. Quadratic ME: $\partial \text{LFP}/\partial \text{FI} = \hat{\beta}_1 + 2\hat{\beta}_2 \times$ FI. Interaction ME: $\partial \text{LFP}/\partial \text{FI} = \hat{\beta}_{\text{FI}} + \hat{\beta}_{\text{int}} \times$ moderator. SE via delta method. All models include individual + wave FE, clustered at community level (450 clusters). N = 68,009.

Table A17: Causal Forest Falsification Tests

	Estimate	SE	p-value
Panel A: Placebo Treatment (Permuted W)			
AUTOC	-0.0042	(0.0115)	0.715
ATE	-0.0039	(0.0195)	
Panel B: Pre-Treatment Outcome (Baseline Wave)			
AUTOC	0.1261	(0.0220)	0.000

Notes: Panel A: Placebo test randomly permutes FI residuals; null AUTOC and ATE expected. Panel B: Causal forest estimated on baseline wave only using cross-sectional variation. Significant AUTOC in Panel B reflects that cross-sectional FI-LFP heterogeneity is detectable even in a single wave; this is expected for continuous treatment and does not indicate confounding of the within-individual panel estimates. The key falsification is Panel A (placebo), which is null as expected. All forests: 2,000 trees, honesty splitting, individual-level clustering. N = 68,009.

Table A18: Heterogeneity in the Effect of Frailty on Labor Force Participation

	(1) Baseline	(2) Gender	(3) Hukou	(4) Age	(5) Education	(6) Insurance
Frailty Index (FI)	−0.0498	−0.1197	−0.0031	−0.0249	−0.0607	−0.1207
	(0.0234)	(0.0299)	(0.0439)	(0.0290)	(0.0266)	(0.0344)
FI × Female		0.1317				
		(0.0392)				
FI × Rural			−0.0581			
			(0.0478)			
FI × Age ≥ 60				−0.0421		
				(0.0276)		
FI × High Education					0.0387	
					(0.0359)	
FI × NCMS						0.0819
						(0.0288)
Age		0.0267	0.0268	0.0274	0.0268	0.0273
		(0.0049)	(0.0049)	(0.0049)	(0.0049)	(0.0049)
Age-squared	−0.0001	−0.0002	−0.0002	−0.0002	−0.0002	−0.0002
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
Married	0.0247	0.0230	0.0226	0.0225	0.0225	0.0227
	(0.0068)	(0.0068)	(0.0068)	(0.0068)	(0.0068)	(0.0068)
Observations	68,009	68,009	68,009	68,009	68,009	68,009
Within R^2	0.001	0.002	0.002	0.002	0.002	0.002
R^2	0.685	0.685	0.685	0.685	0.685	0.685
Individual FE	Yes	Yes	Yes	Yes	Yes	Yes
Wave FE	Yes	Yes	Yes	Yes	Yes	Yes
Clustering	Community	Community	Community	Community	Community	Community

Notes: Dependent variable is LFP (binary). All specifications include individual and wave FE. SE clustered at community level (450 clusters). Time-invariant main effects (female, rural) absorbed by individual FE. Each interaction term estimated in a separate specification to avoid collinearity.

Table A19: Sensitivity to Unobserved Confounding

Metric	Value	Note
Baseline		
Controlled estimate	-0.0498 (0.0234)	N = 68,009; within R^2 = 0.0009; SE clustered (450 communities)
Uncontrolled estimate (FI + FE only)	-0.0479 (0.0233)	Within R^2 = 0.0001 (id FE absorb >99% of variation)
Direction of coefficient movement	$-0.0479 \rightarrow -0.0498$ (away from zero)	Observed controls bias toward zero (positive selection)
A. Oster (2019) Bounds		
Identified set at $1.3 \times \tilde{R}$	$[-0.0505, -0.0498]$	$\delta^* = -8.79$; Excludes zero
Identified set at $2.2 \times \tilde{R}$	$[-0.0525, -0.0498]$	$\delta^* = -35.14$; Excludes zero
Identified set at $R_{\max} = 1.0$	$[-2.5982, -0.0498]$	$\delta^* < -10{,}000$; Excludes zero (illustrative)
A. DMP (2025) Extension		
Breakdown at $1.3 \times \tilde{R}$	Sign-change: -8.79 | Explain-away: -0.70	$\delta^* < 0$: supremely robust
Breakdown at $2.2 \times \tilde{R}$	Sign-change: -35.14 | Explain-away: -2.80	$\delta^* < 0$: supremely robust
Breakdown at $R_{\max} = 1.0$	Sign-change: $-33{,}865$ | Explain-away: $-2{,}700$	$\delta^* < 0$: supremely robust
B. Cinelli-Hazlett (2020)		
RV (sign change, $q = 1$)	0.0106 (1.1%)	Partial R^2 needed to reduce $\hat{\beta}$ to 0
RV (significance, $q = 1$, $\alpha = 0.05$)	0.0032 (0.3%)	Partial R^2 needed for 95% CI to include 0
Benchmark: 1× married	-0.0472 $[-0.082, -0.013]$	$R^2_{D\sim Z\|X} = 0.0009$, $R^2_{Y\sim Z\|D,X} = 0.0004$
Benchmark: 2× married	-0.0445 $[-0.079, -0.010]$	$R^2_{D\sim Z\|X} = 0.0018$, $R^2_{Y\sim Z\|D,X} = 0.0007$
Benchmark: 3× married	-0.0418 $[-0.076, -0.007]$	$R^2_{D\sim Z\|X} = 0.0028$, $R^2_{Y\sim Z\|D,X} = 0.0011$
Benchmark: 1× age^2	-0.0490 $[-0.084, -0.014]$	$R^2_{D\sim Z\|X} = 0.0001$, $R^2_{Y\sim Z\|D,X} = 0.0004$
Benchmark: 2× age^2	-0.0482 $[-0.083, -0.014]$	$R^2_{D\sim Z\|X} = 0.0002$, $R^2_{Y\sim Z\|D,X} = 0.0007$
Benchmark: 3× age^2	-0.0473 $[-0.082, -0.013]$	$R^2_{D\sim Z\|X} = 0.0003$, $R^2_{Y\sim Z\|D,X} = 0.0011$
Summary	All 6 adjusted CIs exclude zero	All benchmarks (1–3× married, 1–3× age^2) remain significant

Notes: Oster (2019): $\delta^* < 0$ = controls bias toward zero (supremely robust). DMP (2025): sign-change and explain-away breakdown. Cinelli-Hazlett (2020): partial R^2 sensitivity; RV = robustness value. Baseline SEs clustered (450 communities). Panel B SEs are OLS; partial R^2 and RV clustering-invariant.

B. Appendix Figures

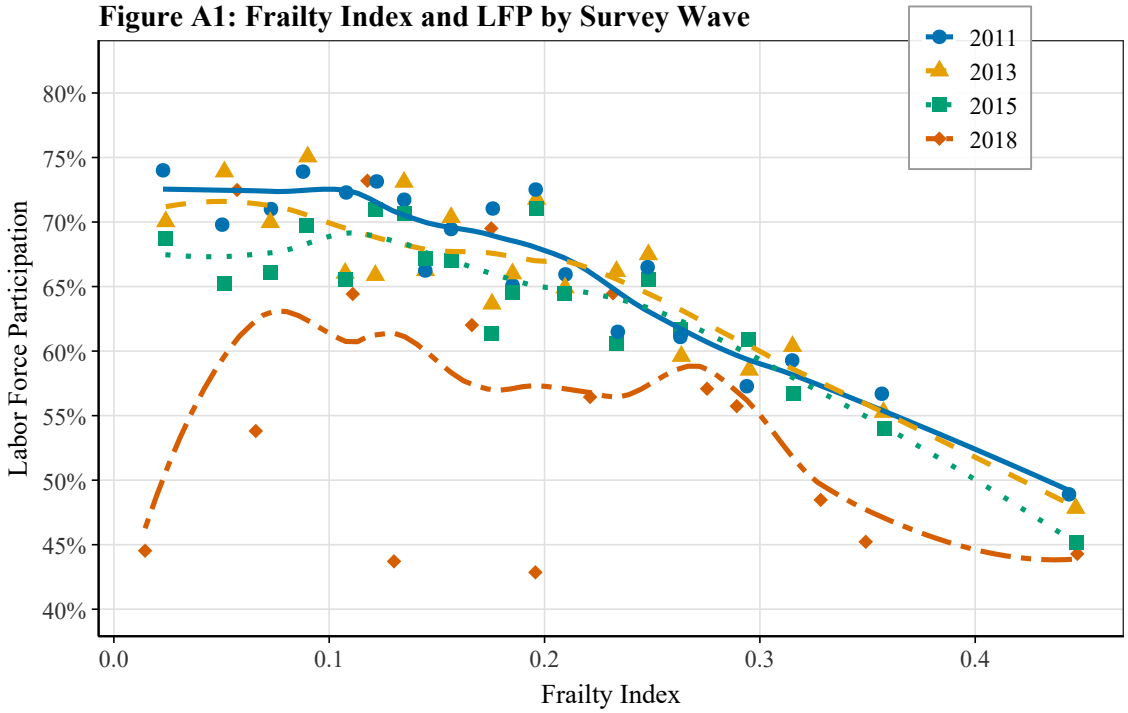


Figure A1: Frailty Index vs. LFP Binscatter by Wave

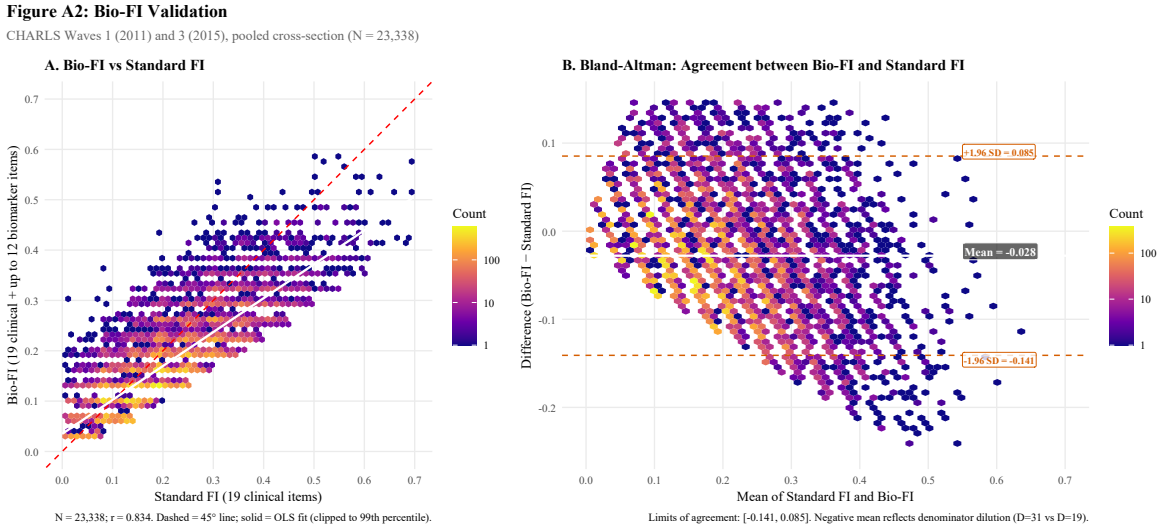
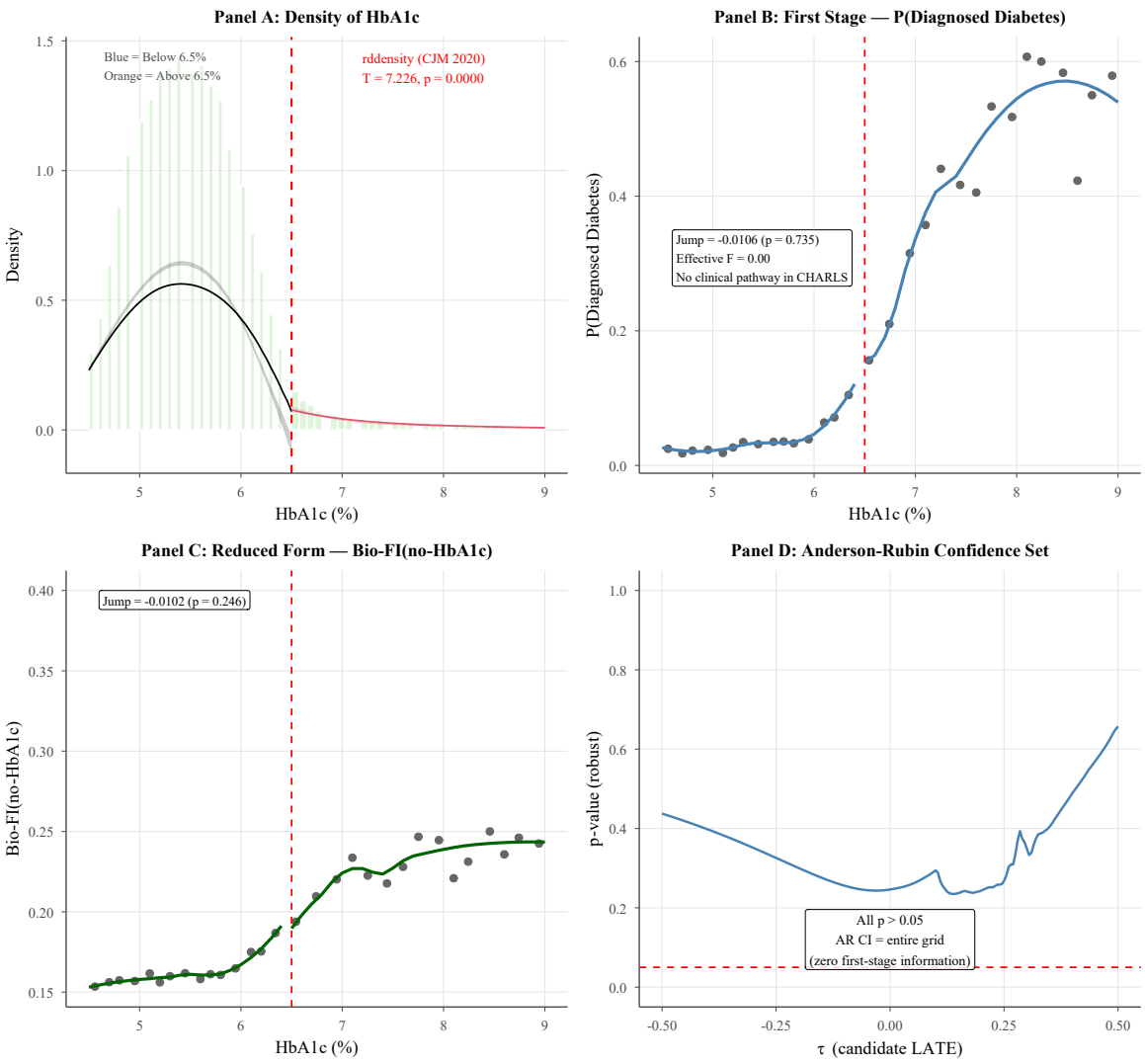


Figure A2: Bio-FI Age Gradient Validation



Running variable: HbA1c (%). Cutoff: 6.5% (ADA diabetes threshold). Local linear polynomial, triangular kernel, MSE-optimal bandwidth, masspoints = 'adjust'.
Density test: rddensity (Cattaneo, Jansson, and Ma 2020), T = 7.226, p = 0.0000 (mass points in running variable). First-stage F = 0.00 (CHARLS blood results not reported to participants).
Data: CHARLS W1 (2011) + W3 (2015), N = 23,212. Bio-FI(no-HbA1c) excludes HbA1c from the frailty index (exclusion restriction).

Figure A3: HbA1c Regression Discontinuity Plots

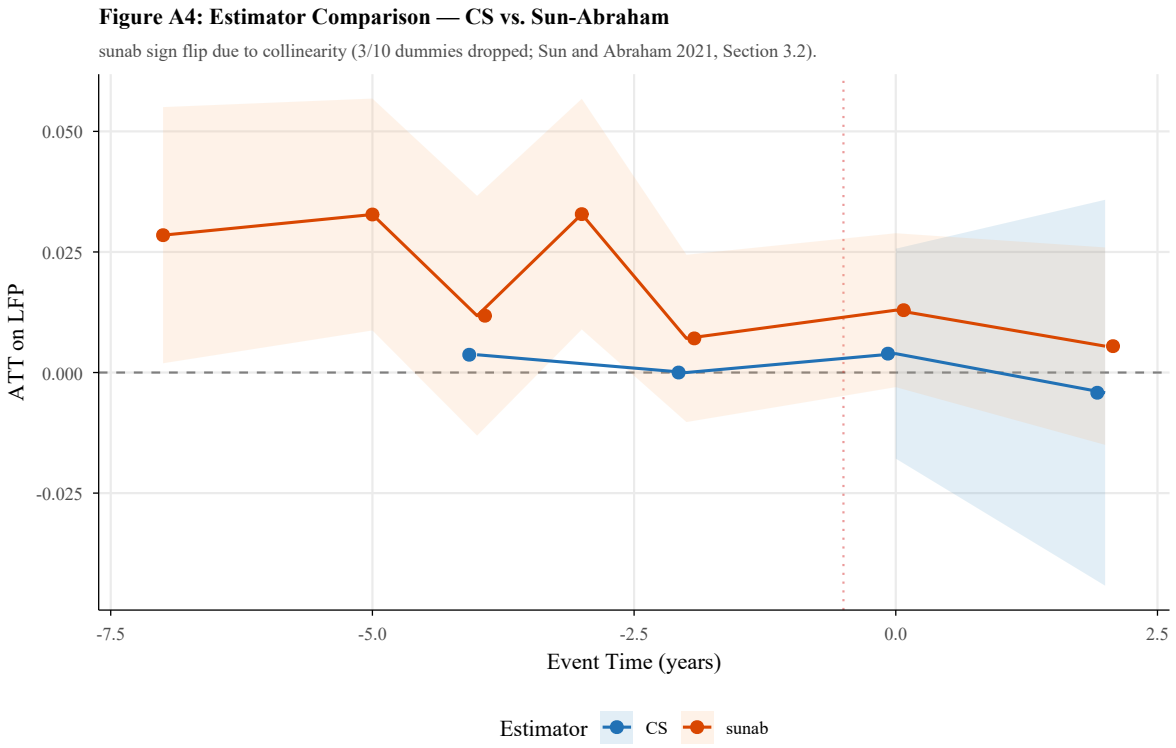
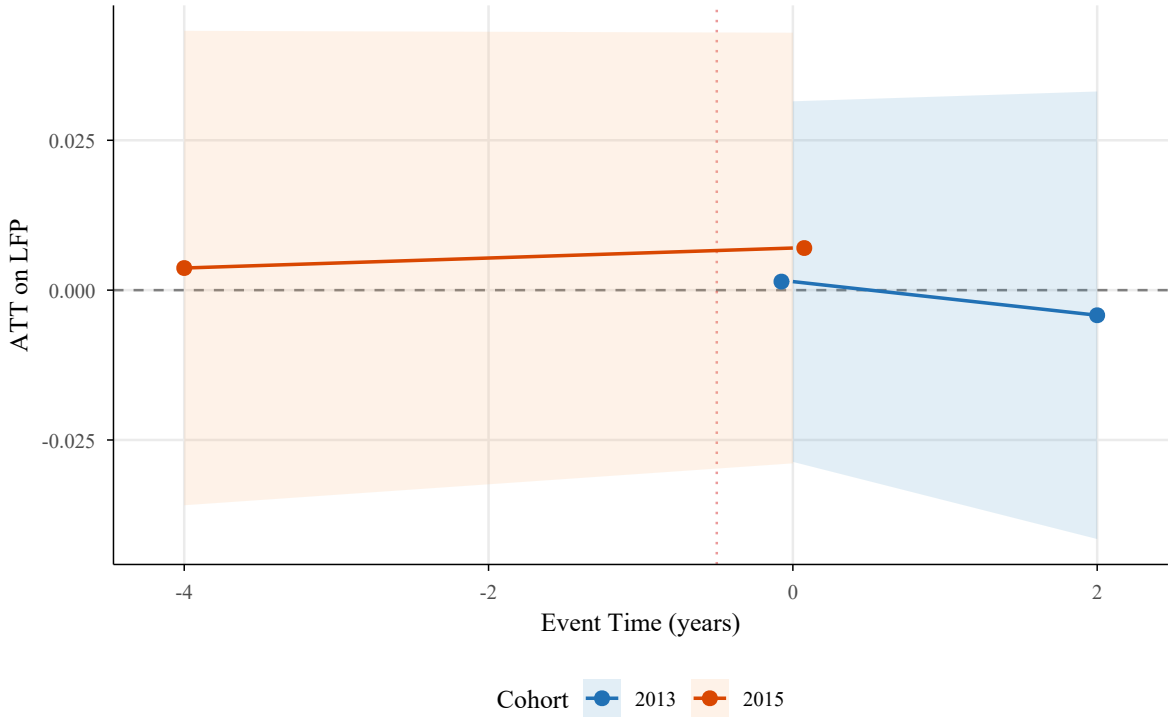


Figure A4: Callaway–Sant'Anna vs. Sun–Abraham Event Studies

Figure A5: Cohort-Specific Event Study

Callaway-Sant'Anna (2021) group-time ATTs by treatment cohort.



2018 cohort omitted: treatment at last observed period leaves zero post-treatment waves for ATT estimation (CHARLS 2011-2018 panel ends at W4).
Pointwise 95% CIs.
SE clustered at community level (450 clusters) in parentheses. No significance asterisks (JHE).
N = 70,263 person-wave obs (21,951 unique individuals). Treated = 6,715, never-treated = 15,236. Cohorts: 2013 = 2,659; 2015 = 2,298; 2018 = 1,758.

Figure A5: Cohort-Specific Treatment Effects

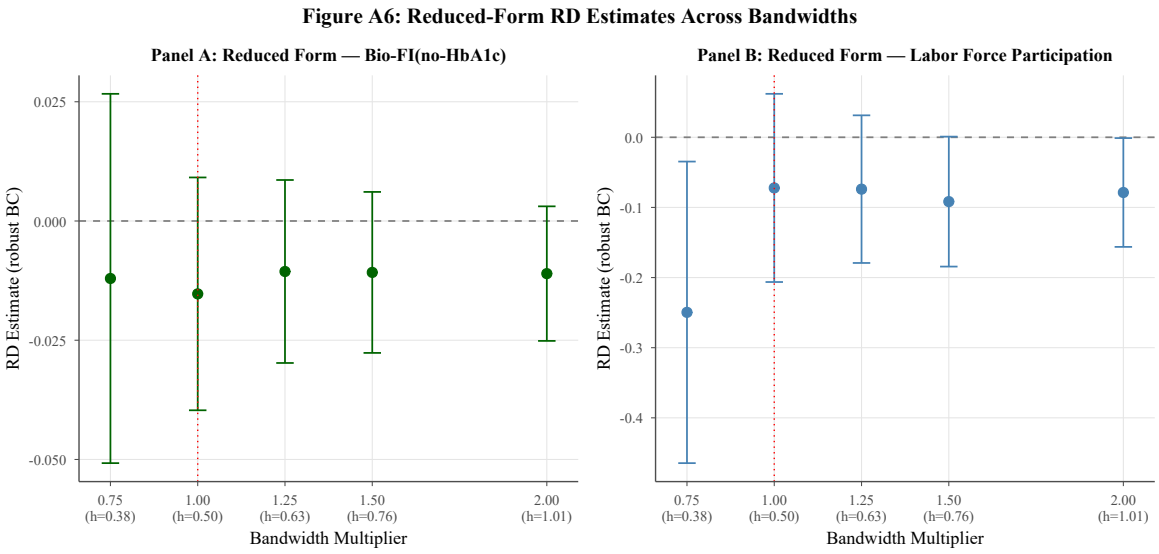


Figure A6: RD Bandwidth Robustness

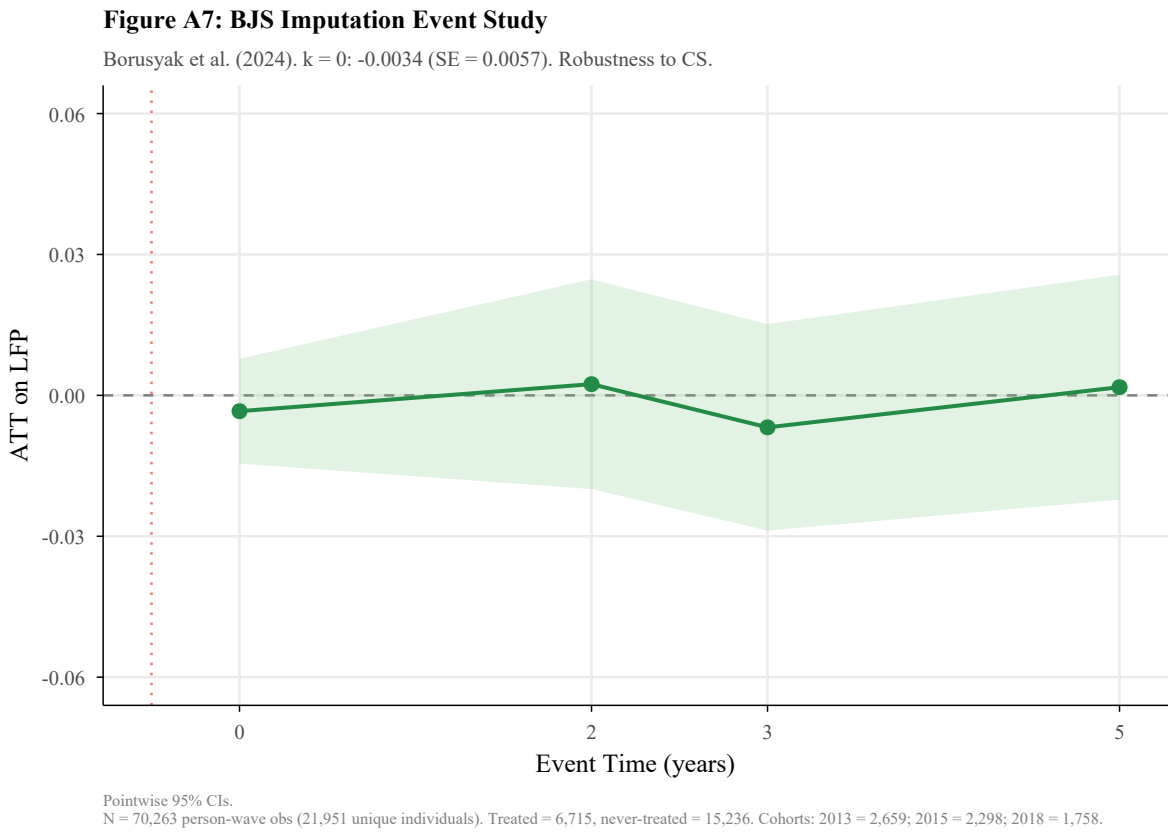
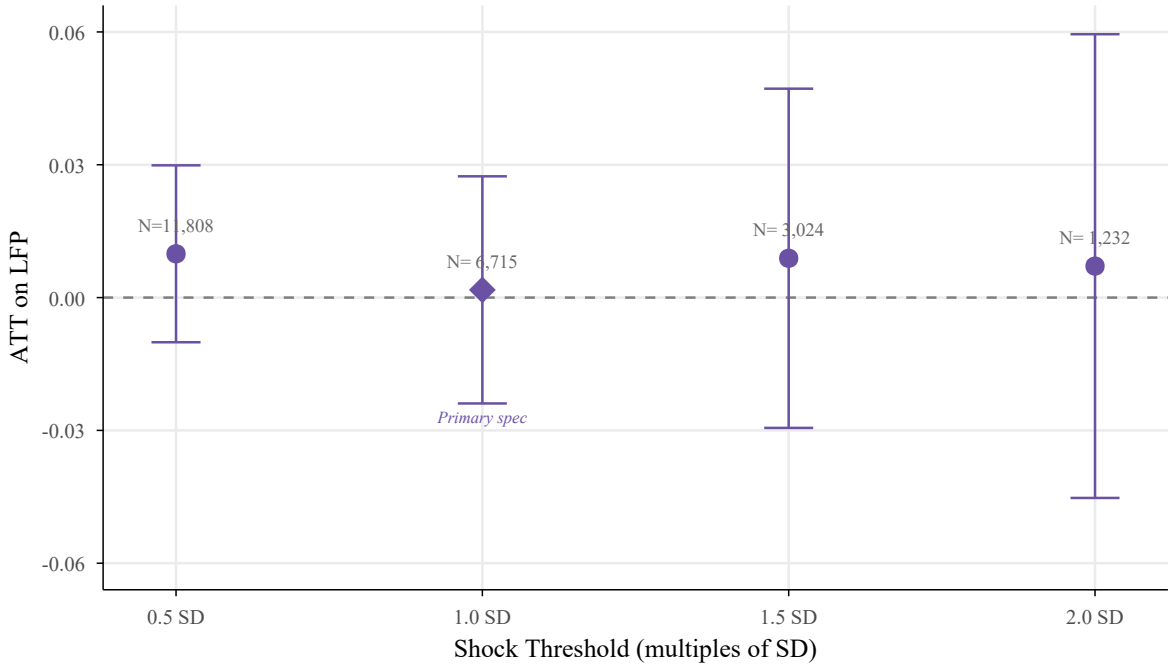


Figure A7: Borusyak–Jaravel–Spiess Event Study

Figure A8: Robustness to Shock Threshold Definition

CS doubly robust ATT(LFP) at alternative FI shock thresholds (delta-FI > k SD).



Primary specification uses 1.0 SD (diamond marker).
Pointwise 95% CIs.
SE clustered at community level (450 clusters) in parentheses. No significance asterisks (JHE).
N = 70,263 person-wave obs (21,951 unique individuals). Treated = 6,715, never-treated = 15,236. Cohorts: 2013 = 2,659; 2015 = 2,298; 2018 = 1,758.

Figure A8: Multi-Threshold Acute Shock Definitions

Figure A9: Causal Forest Variable Importance and Overlap

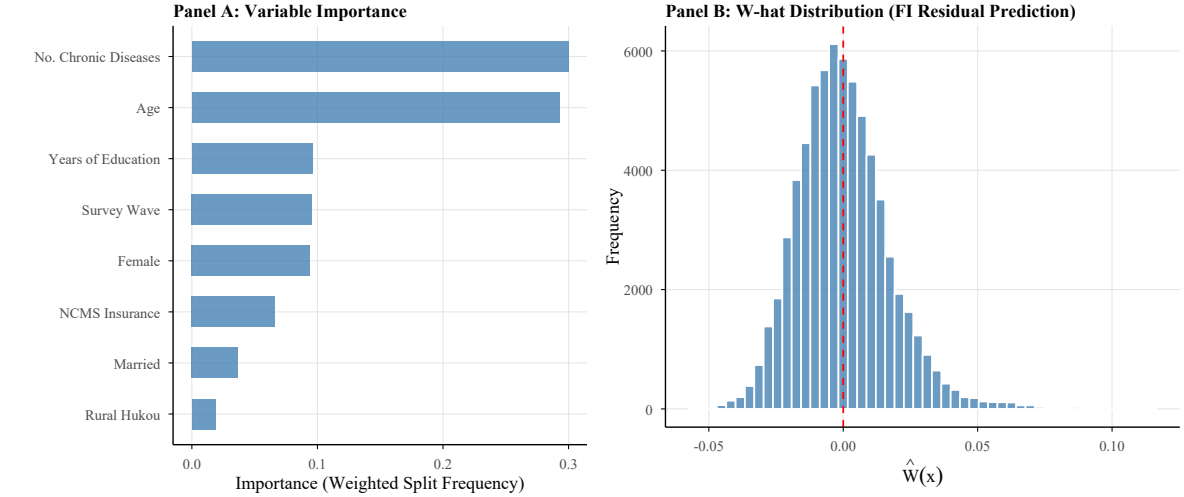


Figure A9: Causal Forest Variable Importance and Overlap

Figure A10: Specification Curve Analysis

96 specifications, 100% negative, median = -0.062

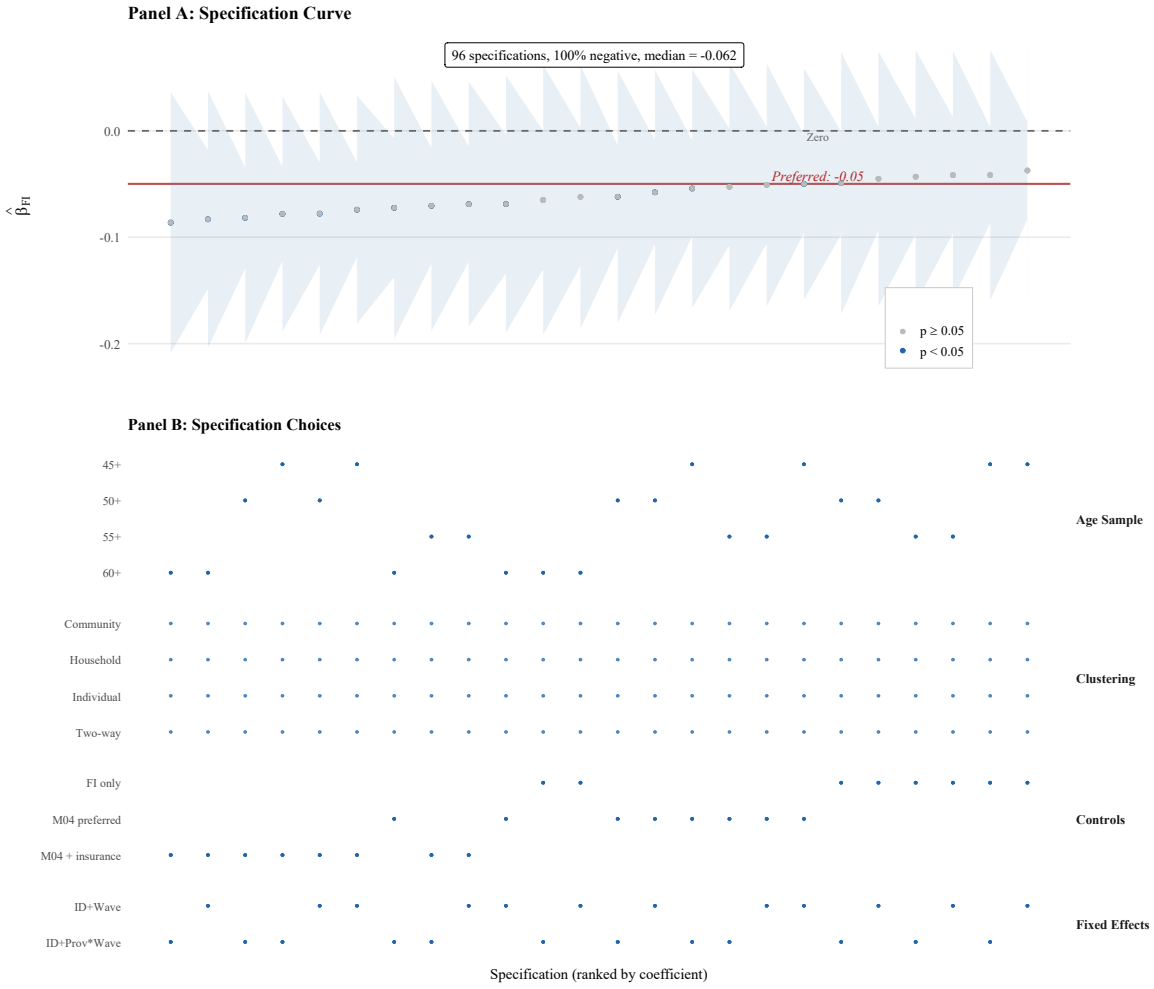


Figure A10: Specification Curve Analysis

Figure A11: Nonlinear FI-LFP Relationship

N = 68,009. Individual + wave FE. SE clustered at community level.

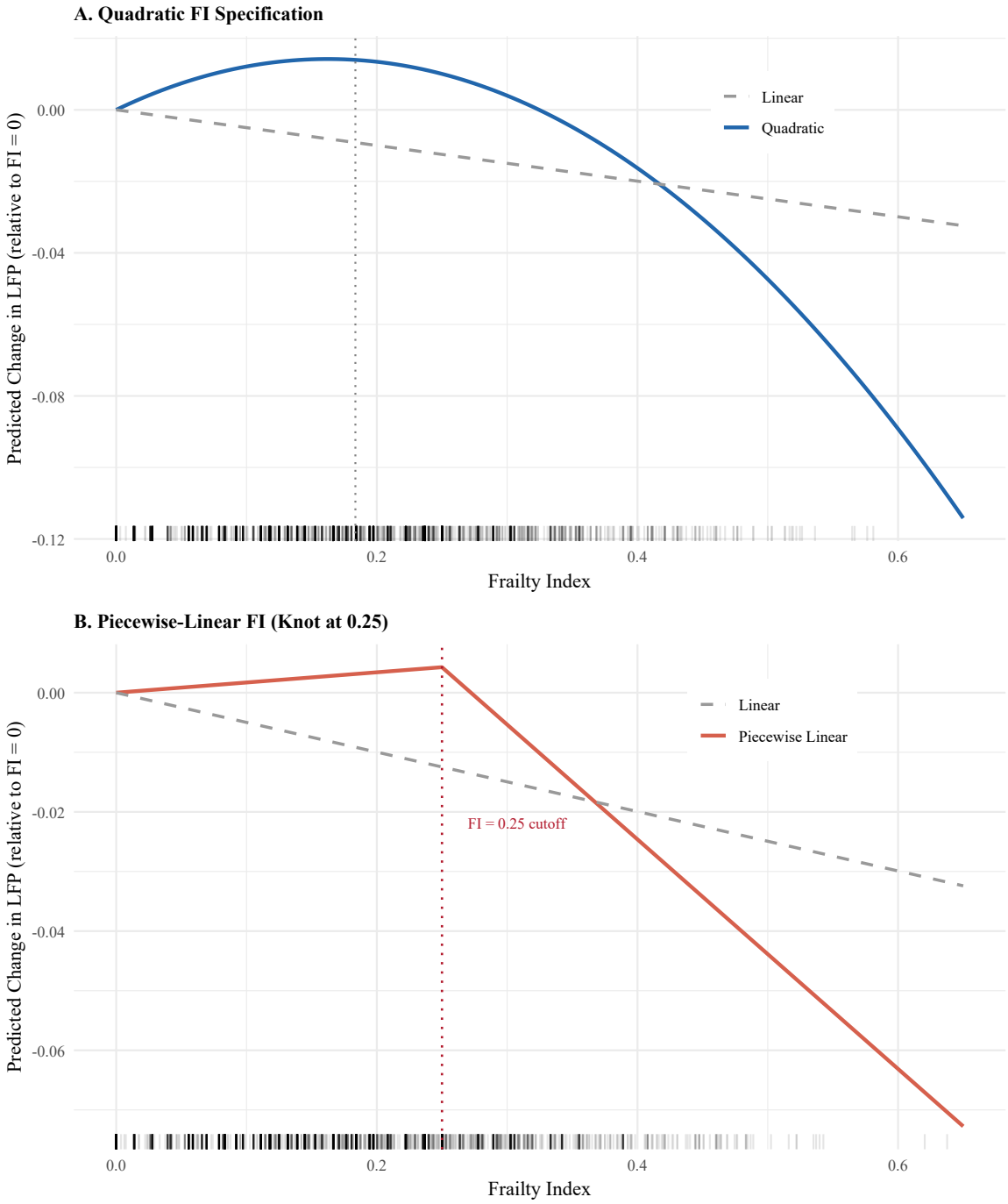


Figure A11: Nonlinear Frailty–LFP Relationship

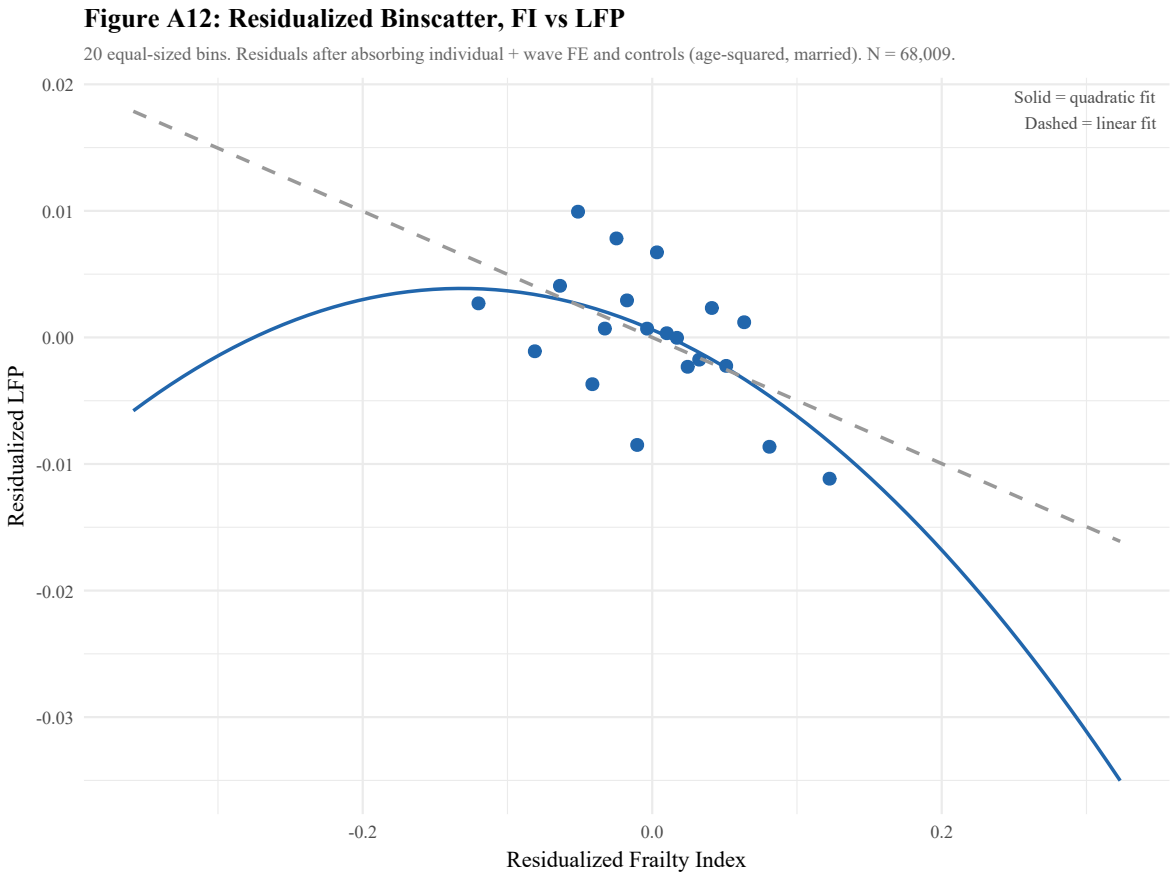


Figure A12: Nonlinear Binscatter

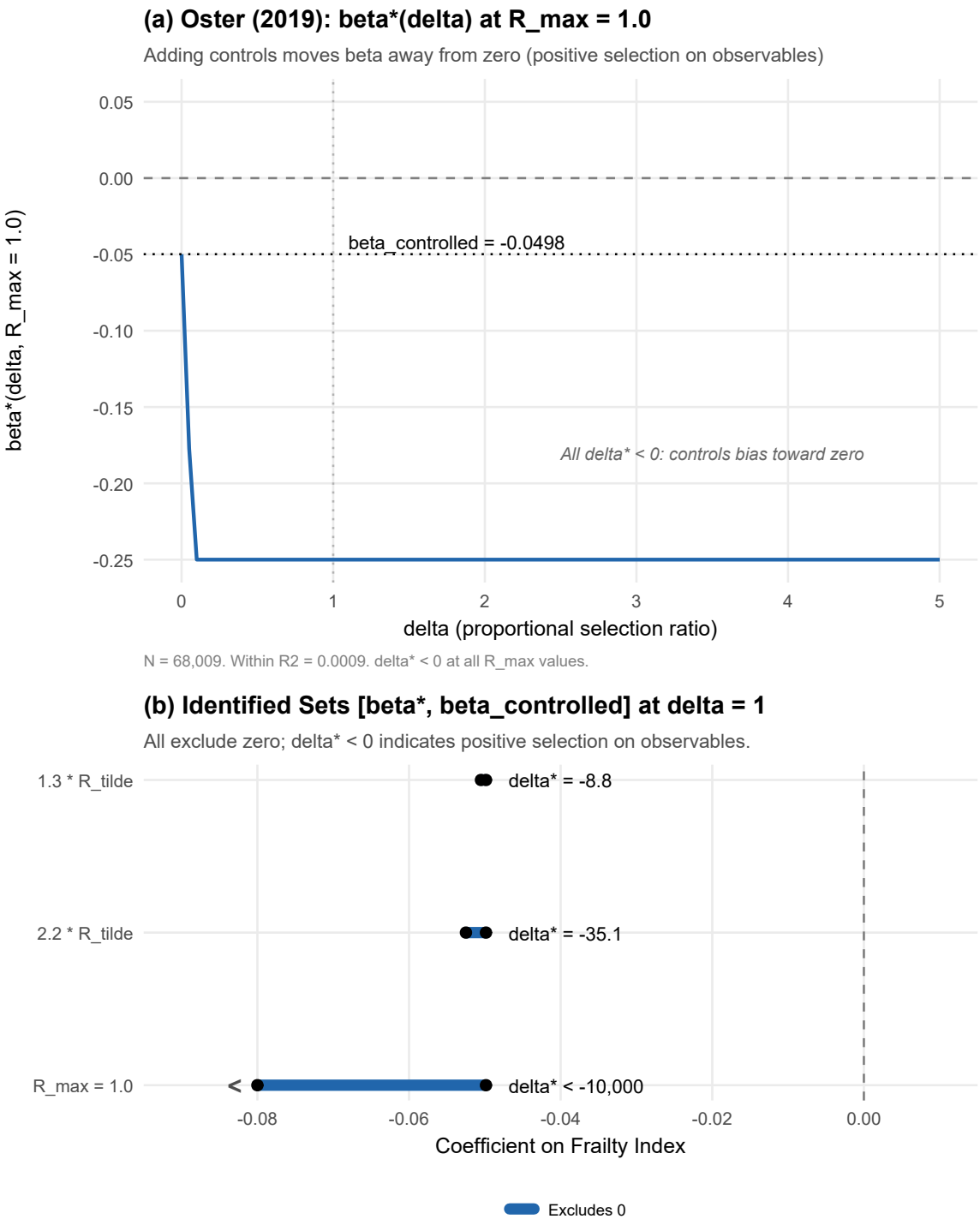


Figure A13: Oster (2019) Coefficient Stability

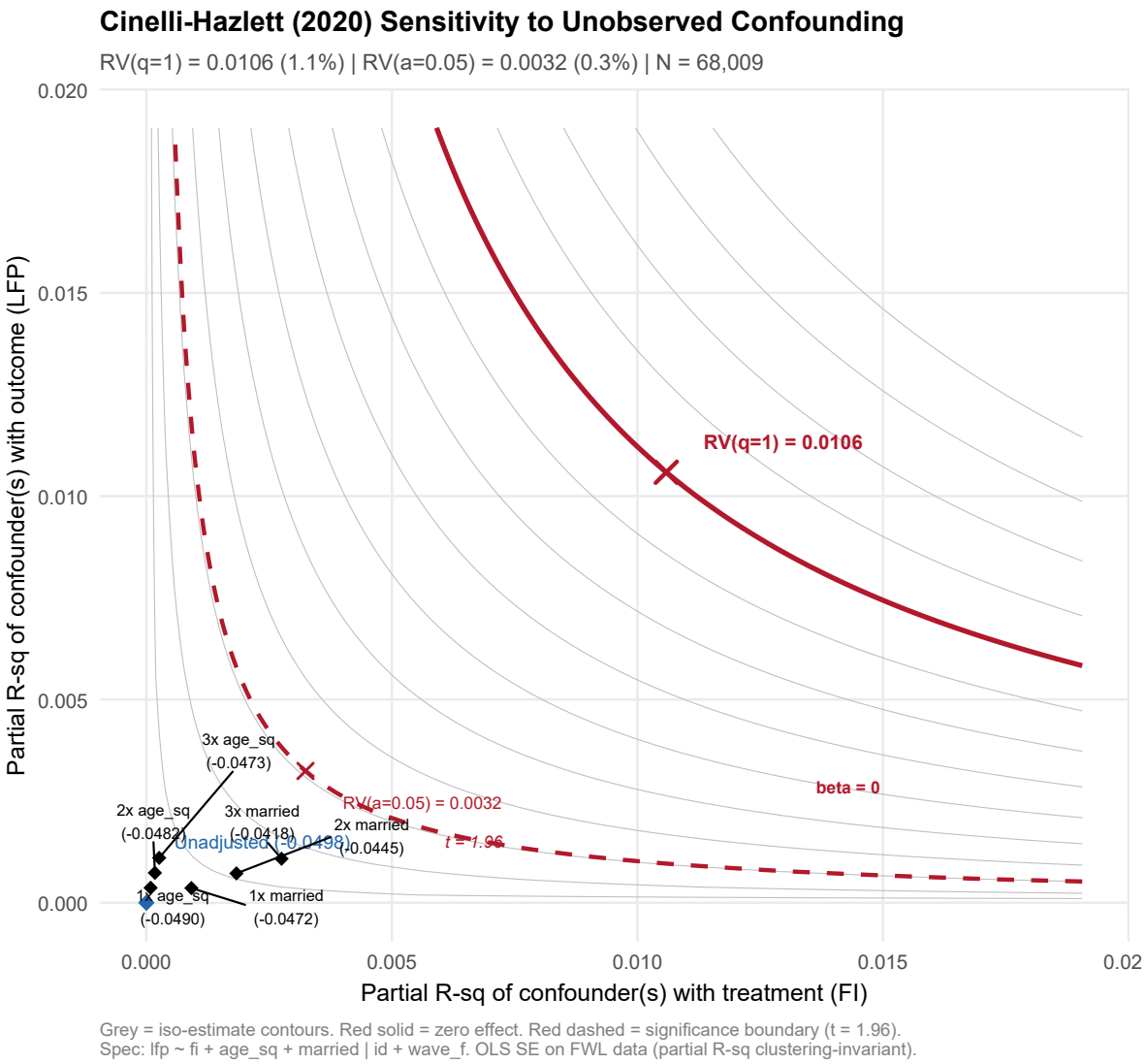
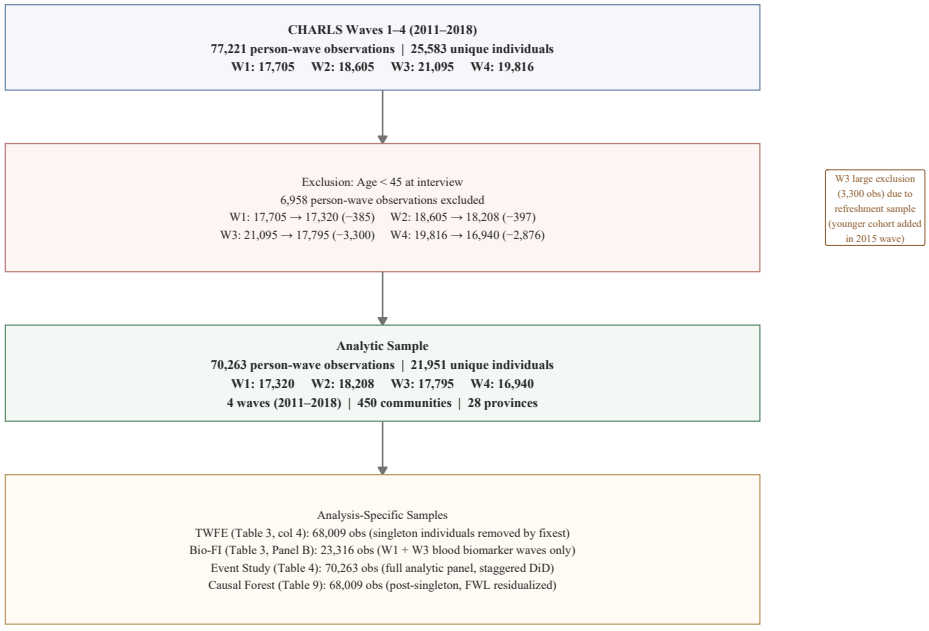


Figure A14: Cinelli–Hazlett (2020) Sensitivity Contour Plot

Online Supplement

Figure S1: Sample Construction Flow Diagram
China Health and Retirement Longitudinal Study (CHARLS), 2011–2018



Notes: CHARLS is a nationally representative longitudinal survey of Chinese adults aged 45+. Exclusion criterion: age < 45 at interview (CHARLS eligibility threshold). W3 (2015) shows the largest exclusion due to a refreshment sample that added younger respondents. TWFE = two-way fixed effects (individual + wave). Singleton individuals (appearing in only one wave) removed by fixest. Bio-FI = Frailty Index augmented with blood biomarkers (collected in W1 and W3 only). FI = 19-item Frailty Index (Searle et al. 2008). W5 (2020) excluded from main analysis due to COVID-19 contamination of labor supply outcomes; used in robustness appendix only. CHARLS approved by Peking University IRB (No. IRB00001052-11015).

Figure S1: STROBE Flow Diagram

Figure S2: Event Study — Spousal Dementia Diagnosis and Own Labor Supply

Callaway-Sant'Anna (2021) DR estimator, not-yet-treated controls. N(treated) = 828; N(control) = 16797. 449 community clusters. Effective balanced N ~ 7,100 (40.6% of panel); onset-2011 cohort (207) dropped.

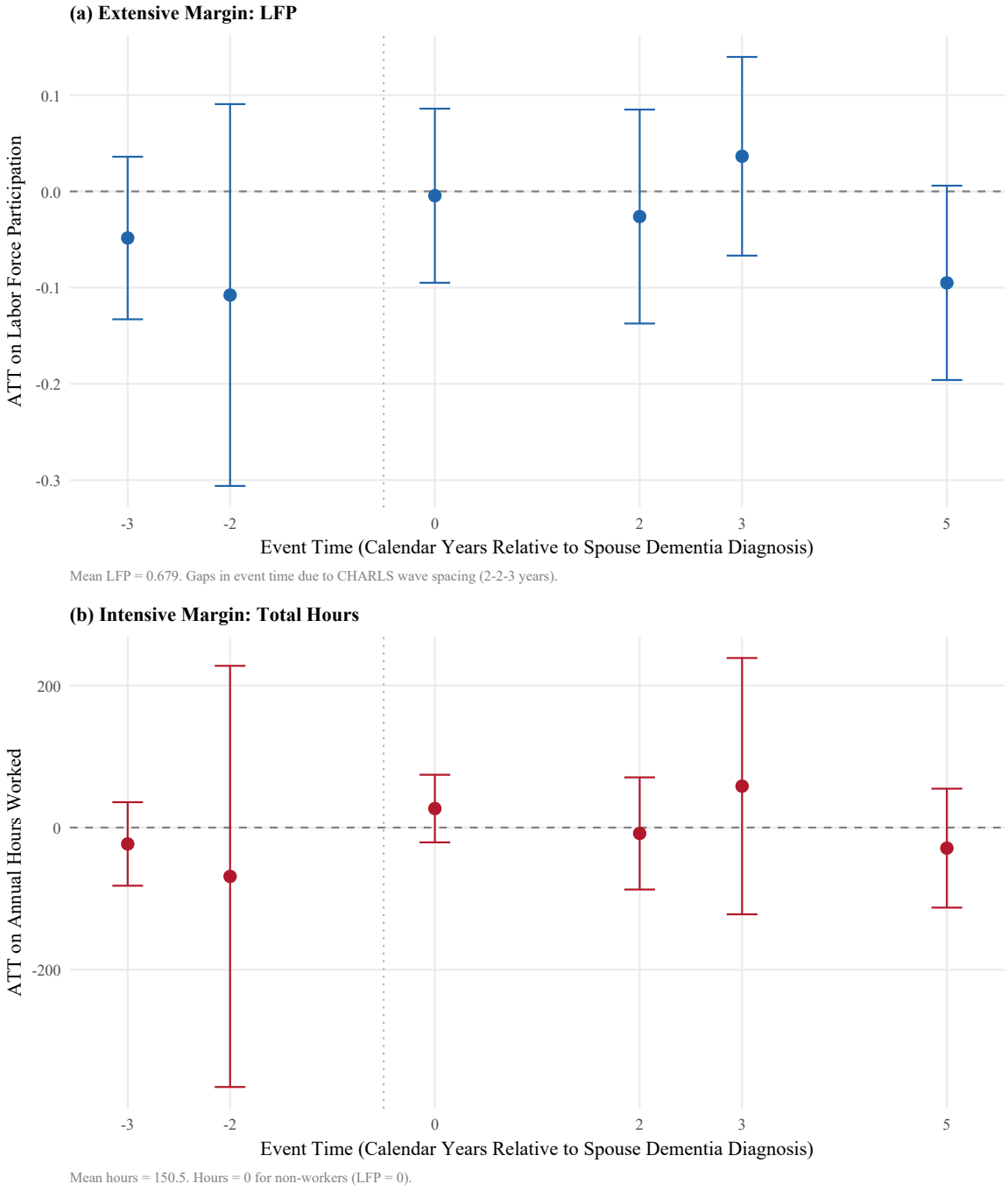
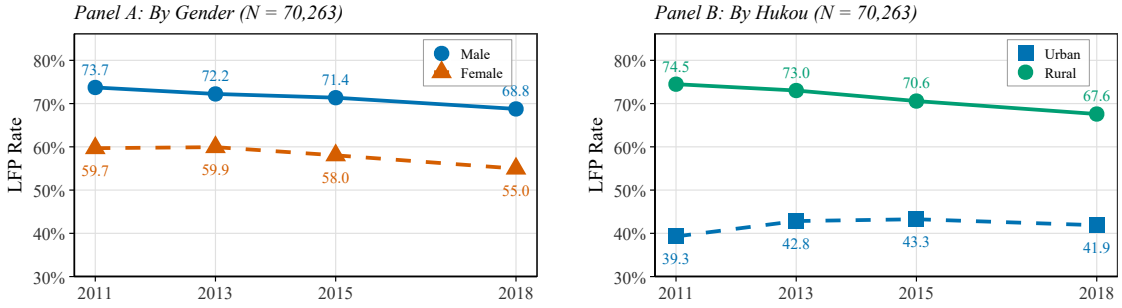


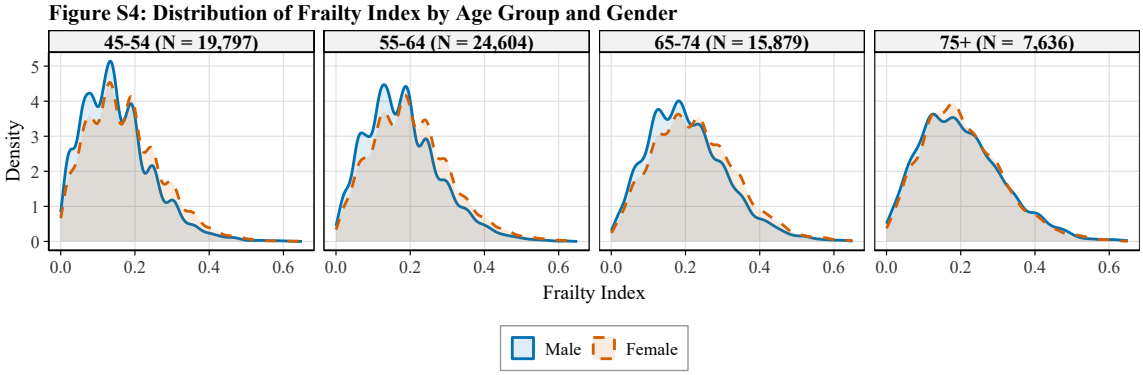
Figure S2: Spousal Dementia Event Study

Figure S3: Labor Force Participation Trends, CHARLS 2011–2018



Notes: LFP = share reporting any work. Labels show percentage points. Unified y-axis (30–82%) for cross-panel comparison. Panel A uses gender imputed via LOCF for 1,238 originally missing obs. Panel B excludes 0 hukou-missing obs (0 if LOCF-imputed). Urban LFP uptick in W2 may partly reflect the inclusion of W2 refreshment-sample respondents, who tend to be younger, inflating urban LFP; interpret cross-wave urban trends with caution. CHARLS Waves 1–4.

Figure S3: LFP Trends by Wave

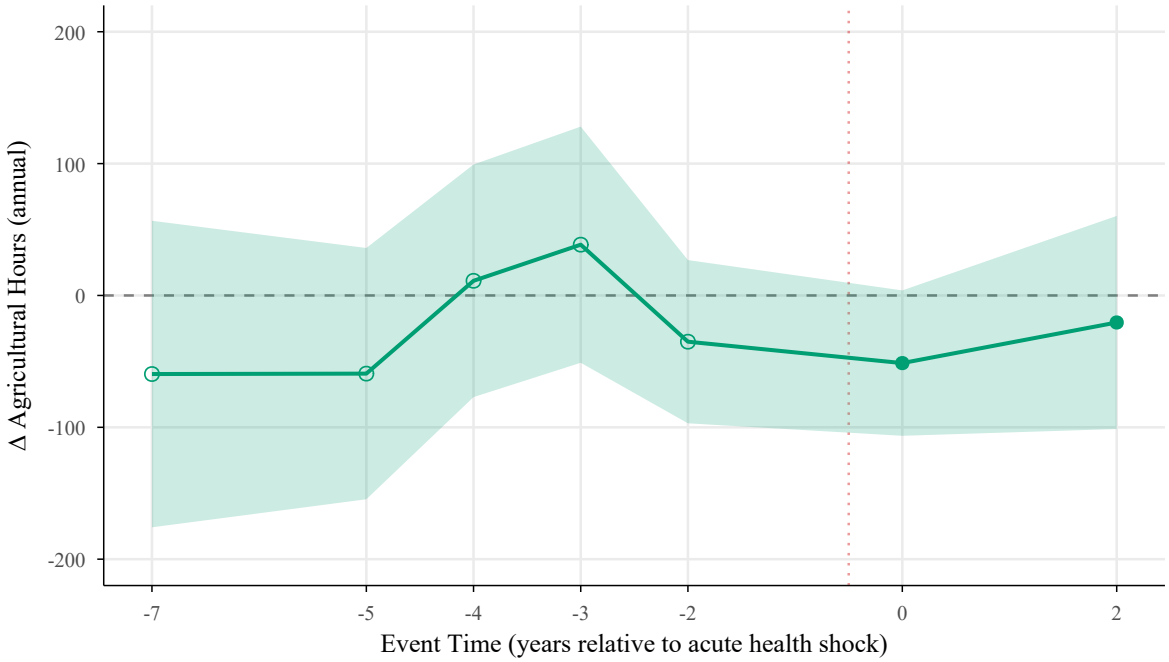


Notes: Kernel density estimates. Male = solid blue; Female = dashed orange. FI > 0 only. Gender-missing excluded. Per-panel N shown in facet headers. CHARLS 2011–2018, all waves pooled.

Figure S4: FI Distribution by Age and Sex

Figure S5: Effect of Acute Health Shock on Agricultural Hours

Sun and Abraham (2021). ATT = -41.2 (SE = 28.5). Interaction-weighted.



Ag hours = FC009 × FC010 × 4.33 × FC011.
sunab ATT = -41.2 vs. CS ATT = -154.7: difference reflects collinearity-induced
sign instability in the sunab estimator with few periods (3/10 dummies dropped);
the larger CS magnitude may also reflect weighting differences across cohorts.
sunab ATT = -41.2 (p = 0.148, insignificant); CS ATT = -154.7 (p = 0.017, significant at 5%).
Unadjusted p = 0.017; Romano-Wolf adjusted p = 0.421 (Table 10).
Pointwise 95% CIs.
SE clustered at community level (450 clusters) in parentheses. No significance asterisks (JHE).
N = 70,263 person-wave obs (21,951 unique individuals). Treated = 6,715, never-treated = 15,236. Cohorts: 2013 = 2,659; 2015 = 2,298; 2018 = 1,758.

Figure S5: Agricultural Hours Event Study