Health Sector Structural Change

By Nick Pretnar and Maria Feldman*

The U.S. health-services sector has grown both in terms of its expenditure share and relative price. Using a two-sector general equilibrium model with monopolistic competition and endogenous population, we find that relative price growth is almost entirely attributable to increasing relative sectoral markups. Demand effects from aging play only a minor role driving up prices and expenditure shares. Controlling for GE effects and rising relative markups, we estimate that health-sector TFP has grown since the 1980s at a rate greater than 1% annually, dampening relative price growth, while leading to higher health-sector output, partially driving up aggregate expenditure shares.

The share of aggregate expenditure devoted to health services has risen by almost 15% since the 1950s in the United States. At the same time relative health-services prices have tripled. While much attention has been given to reasons why the aggregate share of expenditure devoted to health services has risen in the U.S., only recently have researchers considered reasons for the rise in the relative price of health services. Further, the degree to which changing consumption patterns versus production-side factors, like differential rates of technical change and/or markups, are responsible for the structural transformation of the health sector is still unsettled in the literature.

In this paper we explore to what extent both the rise in the relative price of health services and the share of health-services expenditure are driven by three separate factors: 1) rising relative health-sector markups; 2) unbalanced technological change;[1] 3) changes to the efficiency/quality of health-services consumption affecting the composition of demand.[2] While health sector structural change

* Pretnar: Laboratory for Aggregate Economics and Finance, UCSB; Tepper School of Business, CMU (npretnar@ucsb.edu); Feldman: IP Dynamics (mariafeldman93@gmail.com). We are grateful for comments from Arpad Abraham, Daniel Carroll, Karen Kopecky, Dirk Krueger, Alexander Ludwig, Sergio Feijoo Moreira, Victor Ríos-Rull, Ayşegül Şahin, Todd Schoellman, and Hakki Yazici, as well as seminar participants at the University of Bristol, the University of Manchester, Göethe University, the University of Wuerzburg, the Federal Reserve Bank of Cleveland, the 2023 meetings of the Southern Economic Association in New Orleans, and the UCSB, LAEF 2nd Annual Labor Markets and Macroeconomic Outcomes Workshop. We would also like to thank two anonymous referees for very helpful comments. This paper was previously circulated under the titles "The Causal Factors Driving the Rise in Health Services Prices" and "The Causal Factors Driving the Rise in U.S. Health-services Prices."

[1] Unbalanced technical change occurs when total factor productivity (TFP) growth rates differ across different sectors. In models with homothetic preferences and identical, unitary input/output elasticities (i.e., models which do not allow for income effects while also featuring identical Cobb-Douglas production functions across sectors), relative prices between sectors move exactly in inverse proportion to relative sectoral TFP's (Ngai and Pissarides, 2007; Herrendorf, Herrington and Valentinyi, 2015).

[2] It will become clearer what we mean by efficiency/quality in this context. Basically, consumption of health services leads to increased survival probabilities in the sense of Hall and Jones (2007). The degree to which health services consumption can lead to higher survival rates will itself vary across age

is a matter of rising relative prices and increasing sectoral output, to our knowledge nobody has yet explored how different drivers of the evolution of the health sector have affected both prices and output.[3]

Our full general equilibrium (GE) model yields several interesting results. First, when controlling for the re-allocative effects of simultaneously changing relative markups and relative TFP's, our estimates of health-sector TFP growth actually challenge the notion that the health sector is relatively slow growing and subject to cost disease a la Baumol (1967) and Baumol, Blackman and Wolff (1985) as claimed by Triplett and Bosworth (2004), Triplett (2011), and Bates and Santerre (2013). Second, we find that relative markups are almost entirely responsible for rising health-services prices. Indeed, we show that a basic growth accounting exercise would actually underestimate health-sector TFP growth and by doing so under-attribute (even though the effect is already strong) the degree to which markups are responsible for rising prices. This is because, when fully accounting for GE effects that allow for relative markups and relative TFP's to simultaneously impact both input and output prices, we estimate that health-sector TFP has actually grown faster than non-health-sector TFP. We find health-sector TFP to grow faster than the literature would suggest, partly because the literature measuring health-sector TFP has, for the most part, failed to account for both rising markups and re-allocative GE effects.

Further, unbalanced technical change associated with a rising health-sector share of output has actually *bolstered* aggregate economic growth: in an economy without unbalanced technical change, GDP would have grown slower, on average, over the last 70 years or so. Indeed, if our estimates of relative health-sector to non-health-sector markup growth are to be believed, then in order to accurately match the aggregate data, a full GE model requires growth in health-sector productivity to actually *exceed* that of the rest of the economy since the mid 1970's. The rising health-services expenditure share can thus be partially attributed to rising prices due to increasing markups but also due to real gains to health-sector productivity and consumers thus demanding relatively greater quantities of health services.

Meanwhile, we find that changes to the composition of demand for health services, due for example to increasing longevity, have played only a very minor role in driving up prices. In fact, in an economy that experienced no technical change or increase in relative markups while still allowing for the composition of demand to change, relative prices would have risen by less than 7%, expenditure shares would have remained mostly flat, and life expectancy would have still risen,

groups and improve over time, embedding both unmeasured quality improvements to health services and environmental changes that lead to healthier outcomes overall.

[3]Our efforts are partly motivated by the corollary of a question explored in Zhao (2014), as to why health expenditure shares have risen: why have relative prices risen, and how has the rise in prices contributed to structural change? In this sense, our work also complements the analyses in Fonseca et al. (2021), who explore the factors which have led to rising relative health expenditure, and that of Horenstein and Santos (2019); Fonseca et al. (2023), who attribute the increase in U.S. relative prices to markups and price wedges.

though at a slightly slower rate due to slightly slower income growth.

Our results are wrought from two different technical exercises. First, taking the data as given and abstracting from GE effects, we use a multi-sector Dixit and Stiglitz (1977) model to show how relative health-sector to non-health-sector markups can be identified directly from input/output data in a growth accounting exercise. In such an exercise we can then decompose the degree to which the data-implied relative markups versus changing relative sectoral productivities appear to be driving long-run increases in the relative price of aggregate health-services to non-health-services consumption. But, in taking input/output levels and corresponding prices as given, our growth accounting exercise abstracts from the potential for the changing composition of health-services demand to also impact prices. Indeed, as Krueger and Ludwig (2007), Backus, Cooley and Henriksen (2014), and Cooley and Henriksen (2018) all show, in OLG models with population aging, the price of labor and returns on capital investment will depend on the age distribution. If health investment endogenously affects the age distribution too, then its price will also be sensitive to demand changes that result from aging in GE.

Following from this, our second technical endeavor involves introducing a household sector with endogenous survival rates which depend on health investment, a la Hall and Jones (2007). We then calibrate a full general equilibrium (GE) model to match the rise in relative prices, life expectancy, health-services expenditure, and changes to the allocation of labor and capital used to produce health services. Using this model, we counterfactually simulate time series of aggregate data under different assumptions regarding the rates of technical change, relative markup growth, and changes to the composition of health-services demand in order to understand how each channel has contributed to health-sector structural change.

This paper is thus growth and structural change meet health and aging. In this sense our modeling choices and results are in conversation with a recent working paper by Huetsch, Krueger and Ludwig (2023), examining the long-run relationship between income growth and life expectancy via the development of a modern health sector. In focusing on structural change after 1950, we allow for the health sector to use a different technological structure than the broader economy. This is uncommon in the macro-health literature and somewhat uncommon even in the structural change literature, but critical in order to generate testable implications from theory regarding the causes of relative price growth. Indeed, we show that if the changing composition of demand due to aging and increasing efficiency of health investment is contributing to relative price variation in GE, then it *must* be the case that the health sector uses a different production technology than the rest of the economy.

We proceed as follows. We first provide background on stylized facts which summarize health-sector structural change. We then proceed to a growth accounting exercise for a partial equilibrium decomposition of the forces driving up

relative prices. In this exercise, though, we caution about its shortcomings, arguing that a full GE model is needed to decompose the total impact of rising relative markups, unbalanced technical change, and changing demand on aggregate outcomes. In Section II we describe such a stylized model, then calibrate it under various assumptions pertaining to the rate of relative markup growth in Section III. In Section III we also engage in the full counterfactual decompositions of the GE model. In Section IV we conclude by briefly suggesting why policymakers should care about our results.

A. The U.S. Health Services Sector

Figure 1 demonstrates how health-services outlay, prices, and quantities have changed over time. In panel (a) we observe that health's share of domestic personal consumption expenditure (PCE) increased from less than 5% in the late 1940s to 19.6% of all domestic consumption expenditure in 2022. In panel (b) we plot the share of GDP devoted to all health-services outlay, which has also risen since the late 1940s, though not by as much as the health-services share of PCE. In panel (c) we normalize all prices to unity in 1948 and plot the price of health services relative to all non-health-services consumption.[4] The relative price of health services has tripled since 1948. Finally, panel (d) shows relative real quantities of health consumption in units of 1948 real dollars. In 1948 the real value of health consumption was less than 5% of all non-health consumption, though this value rose to greater than 8% by 2022. Examining panels (a), (c), and (d), it should be apparent that the rise in the PCE share is primarily driven by rising prices.

We plot breakdowns of shares and prices for the sub-components that comprise the health-services PCE aggregate in Supplemental Appendix B.3. To summarize, we observe that the relative contribution of prescription drugs and medical appliances to price increases has declined over time, while the relative contribution of hospital services to the aggregated health price has risen. This result is consistent with micro-evidence from Cooper et al. (2019) that hospital systems' local pricing power drives up prices faced by consumers. Horenstein and Santos (2019) also provide aggregate evidence, using data for publicly traded firms from Compustat, that increasing markups have contributed to the rise in the aggregate health-services price level and share of PCE.

Coincidental to rising markups, however, there is also evidence that the health sector has experienced slower total factor productivity (TFP) growth relative to the rest of the economy. In multi-sector models relative prices will be inversely proportional to relative TFP's (Ngai and Pissarides, 2007).[5] In light of this feature of multi-sector models, many have speculated that both rising shares and rising

[4] See Supplemental Appendix B.2 for a list of non-health-services consumption categories and a more detailed discussion about the specific price indices we use.

[5] The correlation is perfectly negative if and only if factor shares are identical for all sectors (Herrendorf, Herrington and Valentinyi, 2015).

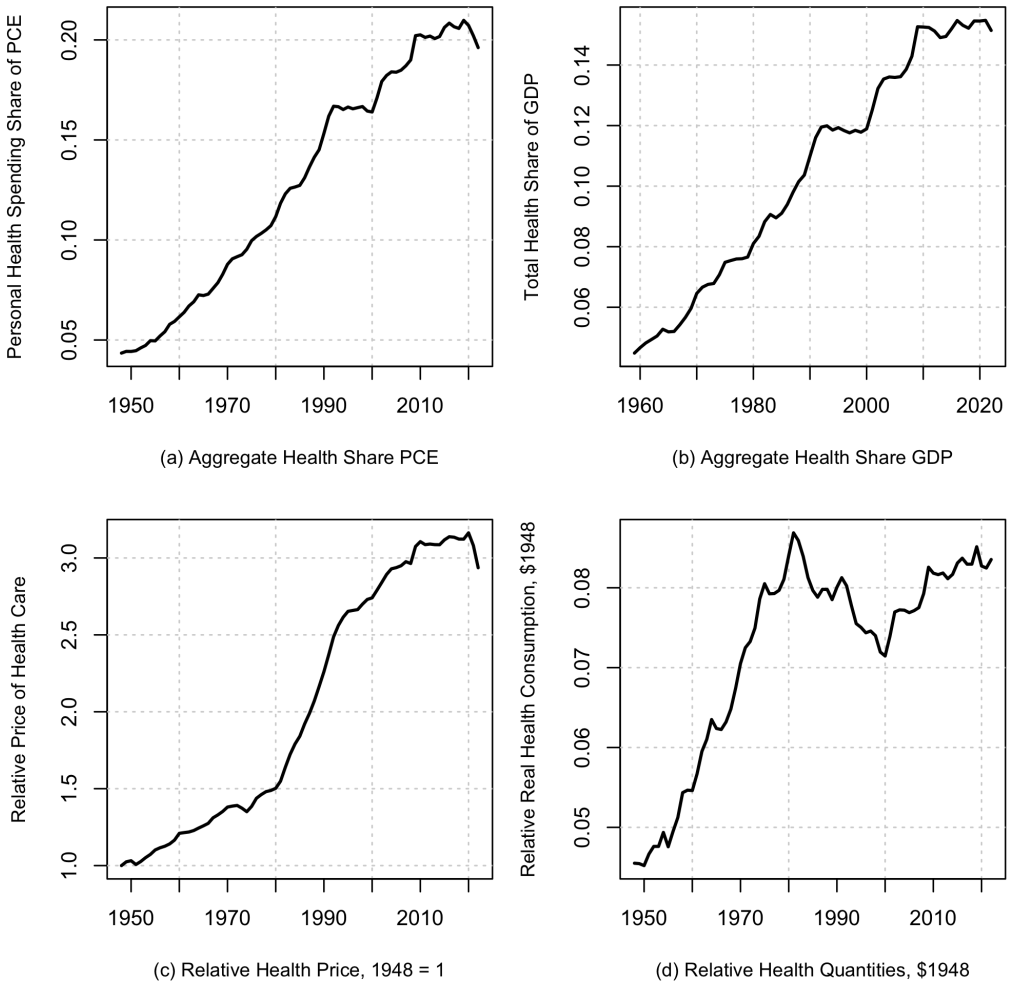

FIGURE 1. PANEL (A) SHOWS THE SHARE OF ALL PERSONAL CONSUMPTION EXPENDITURES DEVOTED TO PERSONAL HEALTH SPENDING WITH DATA TAKEN FROM NIPA TABLE 1.5.5. PANEL (B) SHOWS THE SHARE OF GDP DEVOTED TO ALL HEALTH SERVICES OUTLAY, INCLUDING THE VALUE OF GOVERNMENTAL ADMINISTRATIVE SERVICES NOT INCLUDED IN THE PCE DATA AND WHICH DO NOT DIRECTLY CORRESPOND TO HEALTH-SERVICES RENDERED DIRECTLY TO THE CONSUMER. PANEL (C) SHOWS THE PRICE OF HEALTH SERVICES PCE RELATIVE TO ALL OTHER CONSUMPTION, COMPUTED FROM NIPA PCE DATA (SEE SUPPLEMENTAL APPENDIX B.2). PANEL (D) SHOWS THE REAL QUANTITIES IN $1948 OF HEALTH CONSUMPTION RELATIVE TO ALL OTHER CONSUMPTION.

relative prices may be attributable to unbalanced technical change. Bates and Santerre (2013) conclude that the sector is plagued by cost disease a la Baumol (1967) and Baumol, Blackman and Wolff (1985). Others have also concluded

that long-run health-sector TFP growth is possibly close to zero (Triplett and Bosworth, 2004; Triplett, 2011) or even negative (Cylus and Dickensheets, 2007-2008; Harper et al., 2010; Spitalnic et al., 2022).[6]

B. Population Aging

While there is evidence that slow TFP growth and rising markups have contributed to rising relative health-services prices, the role of general equilibrium (GE) effects on prices, resulting from both population aging and rising incomes, has not been thoroughly explored. As GDP per-capita has risen, the population has also aged, and older individuals require greater amounts of health care and have lower elasticities of health status with respect to health expenditure (Hall and Jones, 2007). Thus, as the population ages we expect that demand for health care will increase and the aggregate elasticity of demand with respect to prices will fall, partially contributing to price increases in GE.

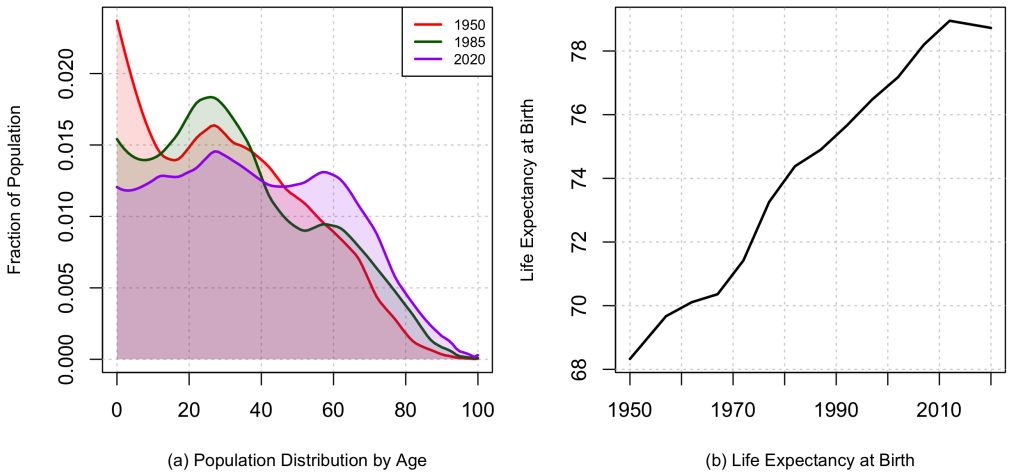

FIGURE 2. PANEL (A) SHOWS THE POPULATION DISTRIBUTION BY AGE, WHERE AGE GROUPS INDEX THE HORIZONTAL AXIS. PANEL (B) SHOWS THE LIFE EXPECTANCY OF NEWBORNS FROM 1950 TO 2020.

Figure 2 shows how the population distribution by age has skewed further to the right over time (a), and that the life expectancy of newborns has also increased

[6] Additionally, Shatto and Clemens (2022) note that "actual health provider productivity is very unlikely to achieve improvements equal to the economy as a whole over sustained periods" because of "the labor-intensive nature of health care services, and presumed limits on the current excess costs and waste that could be removed from the system." Blumenthal, Stremikis and Cutler (2013) also find that the health-care sector is associated with waste, driving up costs, and thus suppressing TFP.

(b).[7] The median age of the U.S. population rose from 30.2 in 1950 to 38.3 in 2020. Further, the median age is expected to continue to increase as fertility rates continue to decline. Thus, if population aging is an important contributor to health-services structural change, then we would expect future aging to continue driving up prices and outlay shares.[8]

I. Two-sector Monopolistic Competition

In this section we engage in a growth-accounting exercise, where we seek to glean a cursory understanding of the contributors to growth in relative health-services prices given sectoral-aggregate data. The goal is to decompose the degree to which relative markups, relative TFP's (i.e., unbalanced technical change), and GE effects, which embed changes to the composition of demand as well as feedback effects from relative markup and TFP growth, have contributed to the rising relative price. We caution that this exercise should be seen as a first pass: it fails to control for how relative markup and TFP variation will impact the allocation of capital and labor inputs, and thus prices, between sectors, since the allocation of capital and labor is endogenous in a GE setting with endogenous demographic dynamics. This is because the rate of return on capital and the wage depend on the demographic composition of the economy (Krueger and Ludwig, 2007; Backus, Cooley and Henriksen, 2014; Cooley and Henriksen, 2018), the total labor supply will vary in the working-age population level which will depend on health services consumption under an endogenous survival mechanism, and the level of assets (capital) will vary in the retiree to worker ratio, since retirees consume partially by dis-saving (Cooley and Henriksen, 2018). Later, in Sections II and III we will close the two-sector model we present here with a household sector in order to fully simulate GE outcomes under different assumptions regarding markup and TFP growth rates, as well as the composition of demand.

We consider several different exercises in order to understand how different assumptions regarding sectoral factor intensities and markups affect the decomposition. In our primary (baseline) exercise we identify the growth rates of relative

[7] U.S. population data by age are taken from the United Nations Population Division, Department of Economic and Social Affairs, 2019 World Population Prospects — Total population (both sexes combined) by single age, region, sub-region, and country. The median age comes from the median age by region, sub-region, and country table in the same data series. Life expectancy is taken from the age-specific life tables by year.

[8] Of secondary importance, our work also speaks to a broader literature that examines the effects of population aging on long-run aggregate growth rates. Under certain conditions exogenous population aging in general equilibrium overlapping generations models has been shown to contribute to decelerations in long-run GDP growth (Backus, Cooley and Henriksen, 2014; Cooley and Henriksen, 2018; Cooley, Henriksen and Nusbaum, 2019; Kydland and Pretnar, 2019; Maestas, Mullen and Powell, 2023). The source of aging (i.e., increasing longevity versus declining fertility), however, may affect both the magnitude of second-order changes to GDP as well as the sign of such changes (Prettner, 2013). In addition to being able to dissect how different forces are responsible for structural change and rising prices, our framework also allows us to understand how the unique confluence of endogenous aging along with the increasing share of GDP devoted to a low-productivity sector, the outputs of which are demanded at levels which increase in age (i.e., health services), have contributed to declining GDP growth rates.

markups and health-sector TFP directly from sectoral input/output data (i.e., the health-services share of aggregate expenditure and sectoral capital and labor allocations). This primary exercise is performed under the assumption that factor-income shares are constant. In our second exercise we relax the assumption of factor-income-share constancy given evidence that the U.S. labor share has been declining and engage in the same decomposition.[9] In the third exercise we take the rate of relative markup growth as estimated for publicly traded firms in Horenstein and Santos (2019) and back out the growth rate of health-sector TFP as jointly implied by both the Horenstein and Santos (2019) markup estimates and sectoral input/output data. With this measure of TFP in hand, we then repeat the same decomposition as before. Finally, in order to understand how accounting for time-varying relative markups affects estimates for health-sector TFP growth, in our fourth exercise we engage in the decomposition under the assumption that relative markup growth is zero, thereby loading the entirety of the increase in relative health-sector prices on either unbalanced technical change or GE effects.

When estimating relative markup growth rates directly from data, as we do in our baseline exercise and the exercise with a time-varying labor share, we attribute almost the entirety of relative price growth to growth in relative markups. The contributions of unbalanced technical change and GE effects cancel each other out. By contrast in a world in which relative markup growth were to be held fixed at zero, we would attribute relative price increases entirely to unbalanced technical change, not GE effects. The case with Horenstein and Santos (2019) relative markups is more mixed: their relative markup estimates grow slower than ours, thus placing a greater burden on unbalanced technical change (i.e., relatively slow health-sector TFP growth) to drive up relative prices. We discuss in further detail below reasons why their markup estimates, which use data only for publicly traded firms, may not be representative of the true change in industry-wide markups since the 1970s. Still, even with relatively lower estimates of relative markup growth, GE effects play no role in driving up relative prices: in fact, in all four of our exercises, GE effects actually have a *dampening* role on relative price growth. This suggests that relative health-sector prices have risen almost strictly as a result of changes to supply-side factors *not* as a result of the changing composition of demand from population aging.

The growth-accounting exercises are based on a two-sector model, where each sector is monopolistically competitive as in Dixit and Stiglitz (1977). One sector produces non-health consumption and investment, $C_t + I_t$, while the other sector produces health-services, H_t. Production in each sector requires both capital and labor, though sectors differ in their production technologies. Further, all firms in each sector are assumed to be homogeneous. This assumption allows the invocation of a symmetric equilibrium, which can be fully characterized by

[9] For strong evidence that the U.S. labor share has been declining, see Karabarbounis and Neiman (2014), Barkai (2020), and Grossman and Oberfield (2022).

sector-specific, monopolistically competitive representative firms.

As a Dixit and Stiglitz (1977) refresher for the case of homogeneous firms, consider an arbitrary production sector producing sector-aggregate, per-capita output y_t. We abstract from sector-specific market selection and market concentration (an abstraction which is a direct consequence of the assumption of firm homogeneity).[10] Normalize the mass of firms in the sector to unity, so that there is a unit-continuum of operating firms, $i \in [0, 1]$. The object, y_t, is an output index, which is a function of all the imperfectly substitutable varieties within the sector: $y_t = \left(\int_0^1 y_{it}^{1/\mu_t^y} \mathrm{d}i \right)^{\mu_t^y}$. It is well-known that the markup of a profit-maximizing firm is just μ_t^y in this set-up. Thus, variations in the monopolistically competitive industry markup over time entirely result from variations in the relative substitutability of within-industry varieties, where the elasticity of substitution is $\mu_t^y/(\mu_t^y - 1)$.[11]

By symmetry aggregate output per sector is then $Y_t = N_t y_t$, where N_t is the population level. Let the sector-level production technologies be Cobb-Douglas so that

$$C_t + I_t = A_{ct} K_{ct}^{\alpha_c} L_{ct}^{1-\alpha_c} \tag{1}$$

$$\text{and} \qquad H_t = A_{ht} K_{ht}^{\alpha_h} L_{ht}^{1-\alpha_h}. \tag{2}$$

TFP's are A_{yt}. K_{yt} are the sector-specific stocks of productive capital. L_{yt} are sector-specific labor allocations. The parameters α_c and α_h are the capital intensities which may differ across sectors but, for the sake of the baseline exposition here in the main text, are fixed over time. In Supplemental Appendix C.1 we present an analogous identification exercise which allows for α_{ct} to vary over time, while holding α_h fixed, though we still show the results of this decomposition here in the main text.[12]

Continuing with the growth-accounting exposition, assume that the average capital rental rate, r_t, and average wage for quality-adjusted labor, w_t, are constant across sectors.[13] Under Cobb-Douglas production the marginal cost can be written solely as a function of input prices, the sector-specific TFP, and

[10] This two-sector framework is thus a simpler version of the multi-sector framework in Behrens et al. (2020), who study a model with all of endogenous concentration, selection, and markups.

[11] As varieties become more (less) substitutable industry-wide markups fall (rise), since firms have less (more) pricing power.

[12] While the identification procedure is slightly different, the results of the time-varying labor-share decomposition are qualitatively similar to our baseline model. Note that we choose to allow α_{ct} to vary over time while fixing α_h due to data limitations: to our knowledge Donahoe (2000) provides the only high-quality measure of α_h, while there are a number of high-quality measures of α_{ct}. We also contend that it is far less of a leap to assume that the $C_t + I_t$ sector's labor share is equivalent to the economy-wide average labor share than to try to actually compute the health-sector labor share year-by-year, given all of the well-documented measurement difficulties that plague health-sector input/output relationships (Triplett and Bosworth, 2004; Triplett, 2011).

[13] While seemingly (perhaps) a significant assumption on the surface, an investigation of the data suggests it is rather innocuous. Supplemental Appendix B.6 provides strong evidence that average-hourly wages in the health-services sector track exactly with wages for all private-sector workers.

production-function parameters:

$$mc_t^y(w_t, r_t\,;\, A_{yt}) = \left(\frac{r_t}{\alpha_y}\right)^{\alpha_y}\left(\frac{w_t}{1-\alpha_y}\right)^{1-\alpha_y}\left(\frac{1}{A_{yt}}\right). \tag{3}$$

By symmetry sector-specific output prices are $p_t^y = \mu_t^y mc_t^y$. The firm in each sector pays the following profits to share holders: $\Pi_t^y = (\mu_t^y - 1)mc_t^y(w_t, r_t; A_{yt})Y_t$. We will denote total profits across sectors paid to shareholders as $\Pi_t = \Pi_t^h + \Pi_t^c$. Sector-specific profits go to zero as varieties get more substitutable ($\mu_t^y \to_+ 1$).

Let $p_t = p_t^h/p_t^c$ be the relative price, $\mu_t = \mu_t^h/\mu_t^c$ be the relative markup, and $A_t = A_{ct}/A_{ht}$ be relative TFP.[14] Dividing $p_t^h = \mu_t^h mc_t^h$ by the same expression for p_t^c and re-arranging gives an expression for the relative price of health services as a function of relative markups, relative TFP's, input prices, and production parameters:

$$p_t = \mu_t\left(\frac{r_t}{w_t}\right)^{\alpha_h-\alpha_c}\left(\frac{A_{ct}\alpha_c^{\alpha_c}(1-\alpha_c)^{1-\alpha_c}}{A_{ht}\alpha_h^{\alpha_h}(1-\alpha_h)^{1-\alpha_h}}\right). \tag{4}$$

In (4) we write relative TFP's in terms of their sectoral variables, A_{ct} and A_{ht}, not the ratio directly. While it is only necessary to know how relative TFP changes over time in order to decompose the degree to which unbalanced technical change has affected relative prices, our identification procedure, which allows us to estimate relative markups, relies on knowing A_{ct} ex-ante (e.g., from the Penn World Tables a la Feenstra, Inklaar and Timmer (2015)) in order to first identify growth rates in μ_t then A_{ht}.

In (4) all effects on p_t due to changes to the composition of demand from population aging will act via GE effects on r_t/w_t. This is because relative TFP's are exogenous, and in the Dixit and Stiglitz (1977) model so is μ_t. This is *not* to say, however, that variation in the composition of demand is the *only* force acting on r_t/w_t. Rather, GE effects embed three different kinds of effects: 1) changes to demand due to demographic changes which affect prices of savings, r_t, and labor, w_t, a la Krueger and Ludwig (2007), Backus, Cooley and Henriksen (2014), Cooley and Henriksen (2018), and Kydland and Pretnar (2019); 2) changes to demand that affect the price of the health care good, consumption of which we assume will directly impact an agent's survival probability like in Grossman (1972), Hall and Jones (2007), and Fehr and Feldman (2024); 3) feedback effects from variation in μ_t and A_t affecting p_t.[15] Expression (4) clearly shows that if the relative price

[14] Note that p_t and μ_t are such that health-services objects are in the numerator, but A_t is such that the health-services TFP is in the *denominator*. Increases in A_t imply a relative increase in non-health-sector efficiency which leads to higher relative health-sector prices.

[15] A structural model with a well-defined household sector is needed in order to engage in a more complete decomposition. We describe such a model in Section II and engage in a quantitative decomposition of the model in Section III.

of health services varies in response to changes in the composition of household expenditure, then it must be the case that the capital intensity of health-services production does not equal the capital intensity of production for all other consumption (i.e., $\alpha_h \neq \alpha_c$). Taking logs of (4) and differentiating in $\ln(r_t/w_t)$, the elasticity of the relative price of health services with respect to the input-price ratio is $\alpha_h - \alpha_c$.[16,17] The degree to which input factor shares differ across sectors will thus determine the degree to which GE effects have caused relative prices to rise.

Place all variables in units of non-health-services consumption. Let $X_t = p_t H_t + C_t$ be total outlay devoted to consumption. Multiply both sides of (4) by H_t and divide both sides by X_t, to get an expression for the share of total spending that goes to health care, $\sigma_{H,t}$:

$$\sigma_{H,t} \equiv \frac{p_t H_t}{X_t} = \mu_t \left(\frac{r_t}{w_t}\right)^{\alpha_h - \alpha_c} \left(\frac{A_{ct}\alpha_c^{\alpha_c}(1-\alpha_c)^{1-\alpha_c}}{A_{ht}\alpha_h^{\alpha_h}(1-\alpha_h)^{1-\alpha_h}}\right) \frac{H_t}{X_t}. \tag{5}$$

Let $\widetilde{g}_{yt} = \ln(y_t) - \ln(y_{t-1})$ be the logged first difference of arbitrary variable y_t. We can write (5) in terms of growth rates:

$$\widetilde{g}_{\sigma_H,t} = \widetilde{g}_{\mu,t} + (\alpha_h - \alpha_c)\big(\widetilde{g}_{r,t} - \widetilde{g}_{w,t}\big) + \widetilde{g}_{A_c,t} - \widetilde{g}_{A_h,t} + \widetilde{g}_{H,t} - \widetilde{g}_{X,t}. \tag{6}$$

The ratio of input prices can be written $\frac{r_t}{w_t} = \frac{\alpha_h L_{ht}}{(1-\alpha_h)K_{ht}}$, which gives an expression for the difference in growth rates as $\widetilde{g}_{r,t} - \widetilde{g}_{w,t} = \widetilde{g}_{L_h,t} - \widetilde{g}_{K_h,t}$. Further, note that $\widetilde{g}_{H,t} = \widetilde{g}_{A_h,t} + \alpha_h \widetilde{g}_{K_h,t} + (1-\alpha_h)\widetilde{g}_{L_h,t}$. We can then use these objects to eliminate the dependency of $\widetilde{g}_{\mu,t}$ on α_h and $\widetilde{g}_{A_h,t}$ and write (6) as

$$\widetilde{g}_{\sigma_H,t} = \widetilde{g}_{\mu,t} + \widetilde{g}_{A_c,t} + \alpha_c \widetilde{g}_{K_h,t} + (1-\alpha_c)\widetilde{g}_{L_h,t} - \widetilde{g}_{X,t}. \tag{7}$$

Given $\alpha_c = 0.4$ (Horenstein and Santos, 2019) and data for $\widetilde{g}_{A_c,t}$ (Penn World Tables), $\widetilde{g}_{\sigma_H,t}$ (NIPA PCE), $\widetilde{g}_{K_h,t}$ (BEA Fixed Asset Tables 3.1 and 3.2; see Supplemental Appendix B.5), $\widetilde{g}_{L_h,t}$ (NIPA Tables 6.5B, 6.5C, and 6.5D; see Supplemental Appendix B.5), and $\widetilde{g}_{X,t}$ (NIPA PCE), we can back out $\widetilde{g}_{\mu,t}$ from (7). The growth in relative markups is thus identified independent of knowledge of α_h or $\widetilde{g}_{A_h,t}$. This concludes the description of the first stage of our identification procedure, which yields estimates of markup growth rates conditional upon input/output data.

[16] Alonso-Carrera, Caballé and Raurich (2015) show that in the absence of constant sectoral capital intensities between two consumption sectors, constancy of relative consumption prices can still be guaranteed as long as the relative price of labor and capital is itself constant. Thus, $\alpha_h \neq \alpha_c$ is necessary but not sufficient to ensure variation in p_t along the growth path. This is because with Cobb-Douglas sectoral production functions the re-allocation of total expenditure affects prices via the non-proportional re-allocation of capital and labor inputs (Alonso-Carrera, Caballé and Raurich, 2015; Herrendorf, Herrington and Valentinyi, 2015).

[17] In the case of time-varying α_{ct}, the elasticity is also time-varying (i.e., $\alpha_h - \alpha_{ct}$).

Now given $\widetilde{g}_{\mu,t}$, if we also know α_h, we can identify $\widetilde{g}_{A_h,t}$ exactly, year-by-year, from a logged version of (4) after replacing $\widetilde{g}_{r,t} - \widetilde{g}_{w,t} = \widetilde{g}_{L_h,t} - \widetilde{g}_{K_h,t}$. We set $\alpha_h = 0.26$ from Donahoe (2000) and back out health-sector TFP growth rates from

$$\widetilde{g}_{p,t} = \widetilde{g}_{\mu,t} + (\alpha_h - \alpha_c)\big(\widetilde{g}_{L_h,t} - \widetilde{g}_{K_h,t}\big) + \widetilde{g}_{A_c,t} - \widetilde{g}_{A_h,t}. \tag{8}$$

As can be seen in (8), the only additional object needed to complete the second stage of the estimation is $\widetilde{g}_{p,t}$ — the growth rate of relative prices from NIPA PCE data. Expression (8) demonstrates that relative price growth is the sum of three separate objects: 1) a GE effect, given by $(\alpha_h - \alpha_c)\big(\widetilde{g}_{L_h,t} - \widetilde{g}_{K_h,t}\big)$; 2) growth in relative markups, given by $\widetilde{g}_{\mu,t}$; 3) growth in relative TFP, given by $\widetilde{g}_{A,t} = \widetilde{g}_{A_c,t} - \widetilde{g}_{A_h,t}$.[18]

Figure 3 presents the estimated decomposition from (8) for our baseline exercise, as well as the three others. Note that even when α_{ct} is time varying, (8) still describes the decomposition we target, except α_c is replaced with α_{ct}.[19] When we use Horenstein and Santos (2019) markups, we are limited to only analyzing the period 1975-2005, since Table 5 of Horenstein and Santos (2019) only provides 5-year estimates of markups over that time frame. Finally, a word of caution: when we discuss the effects of $\widetilde{g}_{\mu,t}$ and $\widetilde{g}_{A,t}$ on price growth, we are referring simply to the "pure" (or partial) effect which can be directly decomposed from (8). Since $\widetilde{g}_{\mu,t}$ and $\widetilde{g}_{A,t}$ also affect the allocation of L_{ht} and K_{ht}, their changes are embedded in GE effects, along with changes to the composition of demand. In this simple decomposition exercise, though, we cannot separately distinguish between the degree to which $\widetilde{g}_{\mu,t}$ and $\widetilde{g}_{A,t}$ are affecting input prices independent of demand effects. Instead, we close the model in Section II in order to do this in Section III.

In both of our baseline and time-varying α_{ct} decompositions (panels (a) and (b) of Figure 3), relative markup growth (red) tracks closely with relative price growth. In these same decompositions the positive effects of unbalanced technical change (purple) are mostly canceled out by GE effects, though with time-varying α_{ct} GE effects also work to dampen the effects of high relative markup growth. This is because $\widetilde{g}_{A_h,t}$ is slightly weaker (on average) when labor shares vary over time, and thus the effects of unbalanced technical change slightly stronger. Stronger negative GE effects are needed to counter the stronger, pure positive

[18] Note that the sectoral allocation of labor and capital will respond to variation in relative input prices, r_t/w_t, as well as relative markups and productivities. It thus captures both changes to the composition of demand that impact relative input prices a la Krueger and Ludwig (2007), Backus, Cooley and Henriksen (2014), Cooley and Henriksen (2018), and Kydland and Pretnar (2019) *and* feedback effects from varying relative markups and relative productivities. In this sense the object $\widetilde{g}_{L_h,t} - \widetilde{g}_{K_h,t}$ is endogenous to this analysis, and thus decompositions around mere data observations provide only a partial picture of the effect of rising relative markups and relative productivities on relative prices. In Section II we introduce a full GE model to account for such endogeneity.

[19] Allowing for time-varying labor shares only complicates the intermediate process of estimating $\widetilde{g}_{\mu,t}$. See Supplemental Appendix C.1.

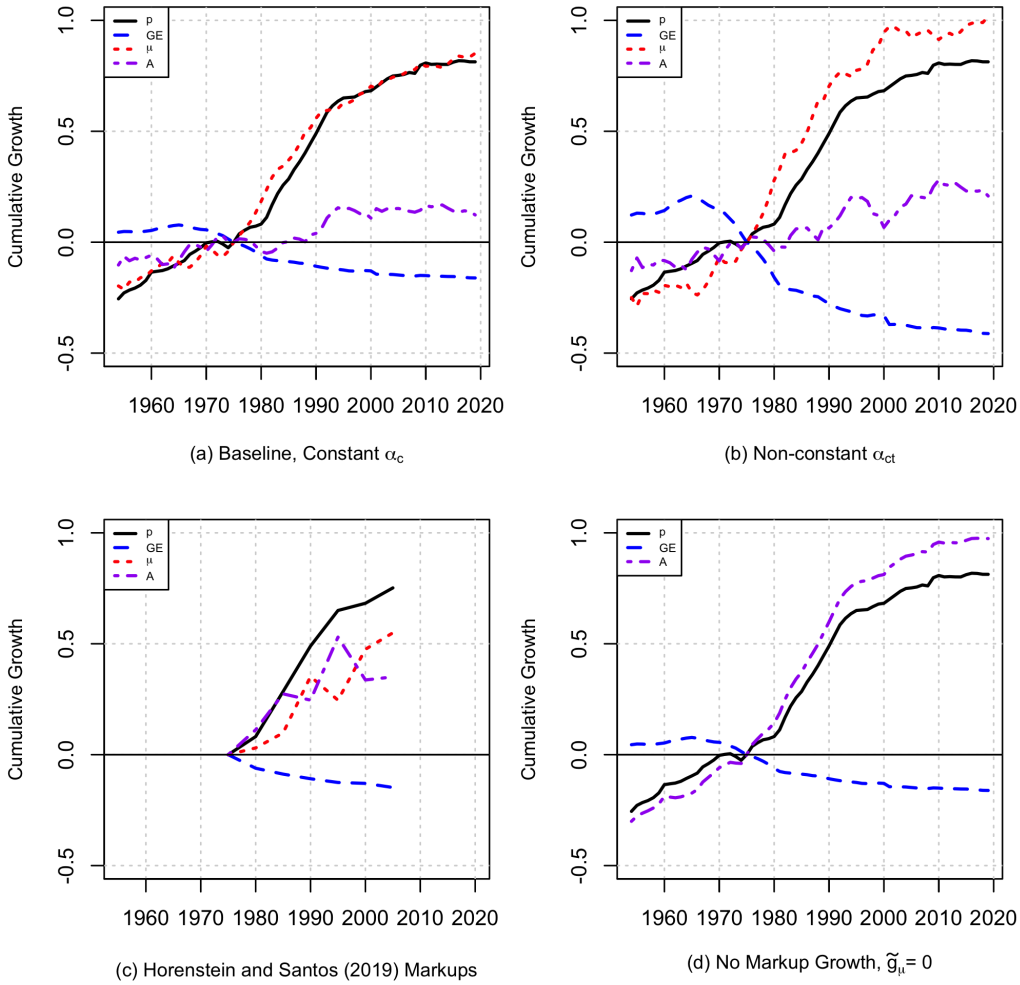

FIGURE 3. PANEL (A) PRESENTS THE GROWTH DECOMPOSITION FROM EXPRESSION (8) FOR OUR BASELINE DECOMPOSITION RANGING FROM 1955 TO 2019. PANEL (B) PRESENTS THE DECOMPOSITION FOR A MODEL WHERE α_{ct} (I.E., THE LABOR SHARE OF NON-HEALTH-SECTOR INCOME) IS ALLOWED TO VARY OVER TIME. PANEL (C) PRESENTS THE DECOMPOSITION FOR A MODEL WITH HORENSTEIN AND SANTOS (2019) MARKUPS FOR PUBLICLY TRADED FIRMS RANGING FROM 1975 TO 2005. PANEL (D) SHOWS THE DECOMPOSITION UNDER THE ASSUMPTION THAT RELATIVE MARKUPS DO NOT GROW.

supply-side effects. Naturally, GE effects are dampening since $\alpha_h - \alpha_c < 0$. When labor shares are time varying, $|\alpha_h - \alpha_{ct}| \to_- 1$ and so GE effects become *more* negative over time, leading comparatively to both higher relative markup growth estimates and slower relative health-sector TFP growth. This can be seen by comparing the series of estimated markups and health-sector TFP's, each

normalized to unity in 1975, in Figure 4 below. From 1954-2019 average annual relative markup growth is estimated at 1.6% in the baseline model and 2% in the time-varying labor-share model. Cumulatively, we estimate that relative markups grew by between 186% (baseline) and 255% (time-varying labor share) over this period. Such estimates suggest that if relative markups continue to grow at their average annual rate from 1955-2015, health-sector markups will be six times those of the non-health-sector by 2050. Average annual health-sector TFP growth, meanwhile, is 0.3% in the baseline model and 0.2% in the time-varying labor-share model, while TFP growth in the non-health sector is approximately 0.7% annually (Feenstra, Inklaar and Timmer, 2015). We estimate that A_{ht} grows by 24.8% from 1954-2019 in the baseline model, but just 11.76% in the time-varying labor-share model, necessitating that model's stronger GE effects. Conditional upon our model structure and aggregate sectoral data, we conclude the following: 1) unbalanced technical change does not appear to drive up p_t; 2) if GE effects do anything, they dampen relative price growth, and they may get stronger over time; 3) almost the entirety of relative price growth can be attributed to rising relative markups.

Panel (c) of Figure 3 shows the decomposition when Horenstein and Santos (2019) markup growth rates are used. Notice that the rising relative markups and unbalanced technical change together exhibit more equal contributions to relative price increases than in our models where we estimate $\widetilde{g}_{\mu,t}$ directly. This is because Horenstein and Santos (2019) estimate that relative markups rose only by an average-annual rate of 1.8% from 1975-2005, while our baseline model predicts an average-annual growth rate of 2.4% over this same period, with the time-varying α_{ct} model coming in at 3%. Indeed, Horenstein and Santos (2019) predict that relative markups increased by 73.2% from 1975-2005, while our baseline and time-varying labor-share models respectively predict increases of 111.3% and 151.5% from 1975-2005. Naturally, given the negative sign on $\alpha_h - \alpha_c$, comparatively slower growth in health-sector TFP (i.e., more unbalanced technical change) is required to reconcile data when relative markup growth is comparatively slower, absent accounting for how changes to A_t also contribute to GE effects. Panel (d) of Figure 3 shows how, in the extreme when there is no markup growth, $\widetilde{g}_{A,t}$ is entirely responsible for rising prices. The Horenstein and Santos (2019) case in panel (c) can simply be thought of as an in-between case.

Note that the period 1975 to 2005 is associated with the fastest rate of relative price growth over the entire time series. Thus, if relative markup growth estimates from this period are for some reason biased downward, this will have significant implications as to the cause of long-run health-sector structural change. There are reasons to believe that because their markup estimates come from Compustat data on publicly traded firms, the results in Horenstein and Santos (2019) may not adequately explain the pricing phenomenon affecting the entirety of the health sector over their sample period. Many hospital systems have only transitioned from not-for-profits to investor-owned since the early-1980s (Gray, 1986). Markup es-

timates inferred from revenue and cost data published by publicly traded hospital systems prior to this time are thus only representative for a small fraction of sectoral output. Further, a large fraction of investor-owned health-service providers are also *not* publicly traded and thus would not be featured in Compustat. For example, in recent years the health services market has seen burgeoning interest by investors with private equity who implement significant cost-control programs which likely lead to markups over marginal costs (Adler, Milhaupt and Valdez, 2023). Along the growth path the health sector has thus become increasingly corporate controlled by both publicly traded corporations and privately held ones (Andreyeva et al., 2024), while markups over marginal costs have risen (Cooper et al., 2019). A model with under-estimated markup growth would mechanically over-attribute the rise in health-services prices (and thus health-sector structural change) to relatively slow TFP growth.

An accurate microeconomic estimate of sectoral markups would need to examine revenues and marginal costs for many different types of firms, from not-for-profits (e.g., religious-affiliated or charitable hospitals) to public hospitals (e.g., county medical facilities) to corporate health systems, both publicly and privately held. Given the transformations that have taken place within the sector, however, such a measurement that relies on publicly available micro data, such as that from Compustat, more accurately reflects aggregate markups today than it did in the past when the market was less concentrated and contained fewer for-profit firms. In their Figure 7 Horenstein and Santos (2019) show that less than 50 health-care firms were present in their Compustat sample in 1970, despite there being approximately 400 firms in the sample by 2006. Estimates of markups from the 1970s using *only* publicly traded firms may bias upward markup *levels* during that period and thus bias *downward* markup growth rates, given strong evidence that the dual forces of concentration and 'corporatization' which have occurred over the last 50 years led to simultaneous cost decreases and price increases (Andreyeva et al., 2024).

It is therefore likely that we estimate that relative markup growth rates are *higher* than Horenstein and Santos (2019) *because* their estimates feature the aforementioned growth-rate bias. From 1975 to 2019, we estimate that relative markups grew by almost 125% in our baseline model. Extrapolating their estimates forward to 2020 would yield 112.5% growth relative to 1975, which is slightly lower than our 125% estimates. Indeed, health services, and particular hospital services, are incredibly concentrated industries. Tawil and DiGiorgio (2022) find Hirschman-Herfindahl Indices (HHI) for managed care $> 2{,}500$ in all of California's counties with 20 of 58 counties having only one, single health-services provider. Nationally, HHI at the MSA level, as measured by the number of hospital beds associated with single firms, increased from a striking average of 5,426 in 2007 to 5,808 in 2017, while 19% of MSAs in 2017 were associated with just one monopolistic provider (Johnson and Frakt, 2020). Going back further in time from 1990 to 2006, MSAs with HHIs $> 2{,}500$ increased from 65% to 77% (Gaynor,

Ho and Town, 2015; Fulton, 2017). The increase in local HHI can be attributed to local mergers, which have additionally been shown to drive up prices in local markets by more than 20% in some cases (Gaynor and Town, 2012; Gaynor, Ho and Town, 2015; Fulton, 2017). Local market concentration is a significant predictor of cross-sectional variation in local prices and thus local markups (Cooper et al., 2019).[20] Given micro evidence that local market concentration is presently strong, has been increasing, and that it contributes to increased markups, we should not be surprised that aggregate estimates of relative markup growth rates are also rather strong.

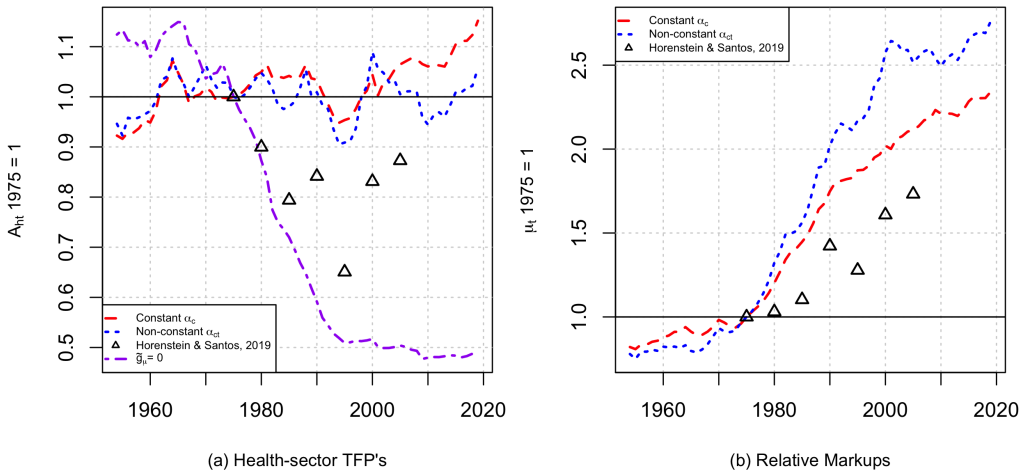

FIGURE 4. PANEL (A) SHOWS ESTIMATES OF RELATIVE MARKUPS FROM THE BASELINE AND TIME-VARYING α_{ct} EXERCISE AGAINST 5-YEAR ESTIMATES FROM TABLE 5 OF HORENSTEIN AND SANTOS (2019). PANEL (B) COMPARES ESTIMATES OF HEALTH-SECTOR TFP'S FOR ALL FOUR OF OUR EXERCISES. ALL VALUES ARE NORMALIZED TO UNITY IN 1975, THE FIRST YEAR OF DATA FROM TABLE 5 OF HORENSTEIN AND SANTOS (2019).

Panel (a) of Figure 4 shows the estimation of A_{ht} for the different models, and panel (b) of Figure 4 presents estimated relative markups alongside those from Horenstein and Santos (2019). Because our markup estimates are more extreme than Horenstein and Santos (2019), our health-sector TFP growth rates are also on the high side relative to both theirs and the rest of the literature. This makes sense: if markups are a bigger driver of relative price growth, then we would expect the impacts of unbalanced technical change to be more muted (i.e., stronger A_{ht}

[20] There is also evidence that physicians tend to call for a greater number of discretionary procedures and encourage patients to take up more elective surgeries in regions both where health-services prices are effectively higher and physicians earn a relatively greater marginal return on each service provided (Clemens and Gottlieb, 2014).

growth relative to the literature). Our average-annual growth estimates for A_{ht} between 0.2% (time-varying α_{ct}) and 0.3% (baseline) from 1955 to 2019 are close to TFP estimates from Spitalnic et al. (2022) for the hospital sector, which are at the high end of estimates from the literature. Meanwhile, using Horenstein and Santos (2019) relative markups and aggregate input/output data, we get -0.4% average annual TFP growth.

Our exercise shows that accounting for relative markup growth is important, otherwise we would attribute the entirety of rising prices to slow sectoral TFP growth (see panel (a) of Figure 4). When $\widetilde{g}_{\mu,t} = 0$, $\forall t$ (by assumption), average annual growth in A_{ht} is estimated to be -1.3% (see the purple line in panel (a) of Figure 4). Indeed, by ignoring relative markup growth we would predict a whopping 56.3% decline in A_{ht} from 1954 to 2019. Given actual advances in health-care technology over this period, this seems fundamentally implausible. True, the literature suggests that health-sector TFP growth has been low and possibly is slightly negative but *not* to the tune of -1.3%. But, aside from Horenstein and Santos (2019), to our knowledge few researchers measuring health-sector TFP have controlled for the possibility that markups are rising. Estimates in the literature thus tend to be lower than what our baseline estimate average-annual growth rate for A_{ht} suggests, though it is still within range in the current exercise. However, we are not controlling for the endogeneity of the GE effects, which is important in any model featuring endogenous aging. As will be seen, when we calibrate a full general equilibrium model in Section III, growth in A_{ht} is likely positive, as our estimates in this section predict, though a stronger growth rate for A_{ht} is needed when accounting for GE effects.

The literature posits estimates of average-annual health-sector growth between -0.6% and 1%, annually. At the high end, such estimates would exceed economy-wide TFP growth, which, recall, Feenstra, Inklaar and Timmer (2015) suggests is approximately 0.7% annually. Triplett and Bosworth (2004) and Triplett (2011) contend that health-sector TFP growth is zero. Estimates from both the ambulatory services (-0.3%) and nursing and residential care (-0.6%) sub-sectors support this contention (Shatto and Clemens, 2022). Cylus and Dickensheets (2007-2008) estimate hospital-sector TFP growth of approximately 0.1% annually, consistent with the "low" estimate for 1987-2018 from Shatto and Clemens (2022). When only labor productivity is considered, this estimate rises to 1% annually (Cylus and Dickensheets, 2007-2008). Spitalnic et al. (2022), meanwhile, estimate that hospital-sector TFP growth is approximately 0.4%, which is higher than the "low" estimate from Shatto and Clemens (2022), but still below aggregate TFP growth rates.

Thus, we conclude that health-sector productivity growth rates implied by input/output data in a version of our model *without* markup growth are far too low relative to the literature (purple line in panel (a) of Figure 4). Still, health-sector TFP growth estimates are likely biased downward if markup growth is not accounted for. Attributing all of relative price growth to unbalanced tech-

nical change would run counter to micro evidence from two separate literatures: 1) health-sector TFP growth rates are between -0.6% and 1% annually; 2) the health-sector is highly concentrated, and concentration has been increasing over time, suggesting markups have likely been rising too.

Our estimates of health-sector TFP growth rates also depend on an assumption regarding the value of α_h. To understand how much assumptions on α_h are affecting estimates of $\widetilde{g}_{A_h,t}$, in Supplemental Appendix C.2 we ex-ante set $\widetilde{g}_{A_h}$ according to different values taken from the literature and estimate α_h conditional upon these values and our first-stage estimates of relative markup growth rates from the baseline model with constant α_c. This robustness check yields reasonable results: values of α_h are between 0.25 and 0.38 for values of $\widetilde{g}_{A_h}$ between -0.6% and 0.4%, annually. Because identification of $\widetilde{g}_{A_h,t}$ in the baseline exercise and α_h in the secondary robustness check both depend on the initial identification of $\widetilde{g}_{\mu,t}$, that reasonable values of these variables can be extracted from our model also lends credibility to our initial estimate of relative markup variation. We conclude that our estimates of $\widetilde{g}_{\mu,t}$ and $\widetilde{g}_{A_h,t}$ are robust, and we are thus satisfied with the validity and power of the growth decomposition exercise in panel (a) of Figure 3.

While it is clear that rising relative markups are primarily responsible for rising relative health-services prices, the growth accounting exercise leaves us with still deeper, mechanical questions. How do structural change and aging interact in this unique environment? We have a slow-growth sector associated with the production of a service, the consumption of which endogenously impacts the population distribution and whose expenditure share is growing over time. As aging has been shown to be exogenously associated with declining aggregate output growth rates (Cooley and Henriksen, 2018; Kydland and Pretnar, 2019), what happens to inferences when we endogenize the aging mechanism inside of this multi-sector model, while also explicitly allowing for changes to the supply-side variables to affect input prices (i.e., controlling for GE effects)?

To explore these and other questions, in the next section we introduce consumer behavior to the model along with an explicit process for endogenous aging. We will then calibrate and simulate a full quantitative-macro GE model, which will allow us to explore deeper GE questions. Our qualitative conclusions as to which factors are primarily responsible for structural change and rising relative prices do not change when controlling for GE effects: rising markups are still the primary culprit. Yet, by endogenizing population aging as a function of health investment, we can also understand how both health-sector market power (as captured by rising markups) and technological change may each be impacting life expectancy via the relative affordability of survival-increasing health-services consumption. Finally, we can also provide estimates of A_{ht} which fully control for *both* relative markup variation and the endogeneity of input prices, adding to the debate as to whether the health sector has experienced so-called cost disease.

II. Demand for Health and Endogenous Population Dynamics

In this section we close the model by adding a household sector and a government that provides social insurance. The production sectors are those described in Section I, and we will not revisit their structures here. For the model exposition we will assume that α_c is constant. We relegate market-clearing conditions and the equilibrium definition to Appendix A.

Consumers/households are heterogeneous only by age, so that each age group is characterized by a single, representative household of age $j \in \{1, \ldots, J\}$.[21] Households can be either of working-age, $j < J_R$, or retired, $j \geq J_R$. They die with certainty after age J, though some fraction of them die accidentally each period. They may save by investing in productive capital, and they supply labor inelastically and automatically retire at age J_R. Finally, households receive dividend payments, π_{jt}, from firms engaged in monopolistic competition.

A government exists solely to tax labor and rebate it to retirees using a pay-as-you-go (PAYGO) pension system. We assume the government's budget clears every period, so that receipts from labor taxes are fully rebated to retired households via transfer payments.

A. Health

We model the period-t health of an age-j agent as the inverse of the non-accidental mortality rate. Denote the mortality rate (i.e., the probability an age-j agent dies in period t before becoming an age-$j+1$ agent in period $t+1$) by m_{jt}. Mortality is the sum of accidental mortality (exogenous) and non-accidental mortality (endogenous): $m_{jt} = m_{jt}^{acc} + m_{jt}^{non}$. The survival rate (i.e., the probability an age-j agent in period t becomes an age-$j+1$ agent in period $t+1$) is $s_{jt} = 1 - m_{jt}$.

Denote health status by x_{jt} and health expenditure by h_{jt}. Health status is a function of health expenditure and determines the inverse of the non-accidental mortality rate: $x_{jt}(h_{jt}) := 1/m_{jt}^{non}$. We let the function $x_{jt}(h_{jt})$ be as follows:

$$x_{jt}(h_{jt}) = z_t \, \phi_j (\zeta_{jt} h_{jt})^{\theta_j}. \tag{9}$$

As in Hall and Jones (2007), z_t is an exogenous aggregate productivity term which explains the efficiency by which health investment can be converted to health outcomes. This object is assumed to grow at a constant rate, g_z, so that $z_{t+1} = (1 + g_z) z_t$. Meanwhile, ζ_{jt} is an exogenous age-specific productivity term. The time-independent output intensity, ϕ_j, and elasticity, θ_j, also vary by age.

[21] We use the terms "household" and "consumer" synonymously.

Given (9) survival rates are

$$s_{jt}(h_{jt}) = 1 - m_{jt}^{acc} - \frac{1}{x_{jt}(h_{jt})}. \tag{10}$$

What does the age-specific productivity term, $z_t\zeta_{jt}^{\theta_j}$, capture? It describes the efficiency by which health investment, as measured in units of outlay or real consumption dollars, actually scales health outcomes. Any technological improvements to the delivery of health services by medical professionals or health-care institutions will be embedded in the real value of h_{jt}, which is increasing in sectoral TFP. As Hall and Jones (2007) point out, $z_t\zeta_{jt}^{\theta_j}$ captures variation orthogonal to variation in TFP.[22] For example, the efficiency of health investment at generating better health outcomes may increase because of policies that reduce pollution, mandate safer food preparation practices, and enforce water-quality standards. Additionally, along the development path people have learned how to live healthier lives (e.g., smoking less, reducing intake of certain kinds of processed foods, exercising regularly, taking vitamins, having safe sex, etc.) so that every dollar invested in health services goes further toward generating healthy outcomes and reducing non-accidental mortality rates. One can also think of z_t as the aggregate quality of health services and ζ_{jt} as an age-specific quality residual. The total quality of health services consumed by an age-j agent in period t is $z_t\zeta_{jt}^{\theta_j}$. Both z_t and ζ_{jt} thus capture how productive health-services quantities are at generating increased survival probabilities.[23]

It is also not difficult to think of reasons as to why this term is age specific. Think, for example, of changes to knowledge with regards to how we engage in infant care: babies are not lain prostrate anymore due to risk of sudden infant death syndrome. While such knowledge was indeed acquired via medical research, this knowledge does not directly affect the *quantity* of health-services consumption (i.e., health investment, h_{jt}) that the consumer purchases, but rather affects the *efficiency* by which such investments may be converted into higher health and lower risk of mortality for humans of a specific age. What about an example of technological improvements to health investment that may affect agents at different ages differently? Well, take the smoking example again: a sudden change to regulations on indoor smoking will affect z_t, yes, but the effect may not be uniform across generations. After all, a generation of middle-aged adults who

[22] Palangkaraya and Yong (2009) also suggest a role for exogenous changes to health-consumption efficiency: if younger people engage in healthier lifestyles, disease rates like heart disease and cancer fall leading to relatively less demand for expensive, acute care at or near the terminal stages of life. Thus, by prolonging the terminal stages of life, demand for acute care can also be prolonged until, eventually, a single individual's demand for such care falls to zero, assuming they live long enough.

[23] Additionally, there is a sense in which the quality of health services is actually poorly reflected in health-services prices and thus *measured* quantities (Lawver, 2010). Mis-measurement of the *quality* of health services will thus bias price measurements. One can think of both z_t and ζ_{jt} as residuals that could be used to assess quality mis-measurements a la Lawver (2010) based on variation in actual health outcomes (i.e., survival rates in our context).

spent 20-30 years inhaling second-hand smoke in indoor spaces will still face significant risks of chronic respiratory illnesses, so that for these adults ζ_{jt} may reduce the impact of the increase in z_t on health outcomes, while for younger people ζ_{jt} may actually scale the impact positively.

Finally, note the distinction between TFP associated with the *production* of health services via the combination of capital and labor versus the age-weighted productivity $z_t\zeta_{jt}^{\theta_j}$. Some, such as Romley, Goldman and Sood (2015), have argued that estimates which suggest the health-services sector is associated with relatively low TFP growth must be wrong, given gains to health outcomes.[24] But the logic of such an argument is entirely incorrect, for it supposes that scaling the combination of capital and labor inputs at the production level (i.e., increasing A_{ht}) should lead to better health *outcomes*, rather than just leading to more health services H_t at every possible combination of capital and labor, K_{ht} and L_{ht}. Truly, healthier outcomes result from re-scaling health *investment* in some type of health production function, and this is what $z_t\zeta_{jt}^{\theta_j}$ captures — the unmeasured efficiency/quality of *using* health-care services.

B. Demographics

Let N_{jt} denote the total number of agents of age j that are alive in period t. This object evolves endogenously as follows: $N_{j+1,t+1} = s_{jt}N_{jt}$, $\forall j > 0$. We assume that $s_{Jt} = 0$, forcing $N_{J+1,t+1} = 0$, for all t. The population level at which a cohort enters the economy is denoted by N_{1t} and exogenously grows at net rate g_N. We assume that all migration for each cohort happens at the beginning of working-age life.

C. Preferences

Households have preferences over their non-health-services consumption as follows:

$$u(c_{jt}) = \chi + \xi \frac{c_{jt}^{1-\gamma}}{1-\gamma}. \tag{11}$$

$\chi > 0$ is an intercept that forces utility to be positive so that consumers have an incentive to live. ξ scales the contribution of non-health flow consumption to utility and is needed in order to match the health-services share in the model calibration.

[24]There are methodological issues with the approach in Romley, Goldman and Sood (2015). The authors cannot account for the kinds of near-perfectly discriminatory pricing schema hospital systems engage in, as they index all costs to the administrative prices used by Medicare. This biases their productivity estimates upward, considering that hospitals can markup costs at higher rates for patients of private insurers that have less monopsonistic pricing power than Medicare, as suggested by the findings in Cooper et al. (2019).

D. Wealth

Households may save by investing ι_{jt} toward the accumulation of capital assets a_{jt}. Further, given some fraction of each cohort's members perish each period, we must re-distribute the assets bequeathed by those who die. Let b_t denote the bequests received by any living agent at the beginning of period t. These bequests were bequeathed accidentally by those who died in period $t-1$. We assume all living agents receive the same bequests, so we do not index these objects by j. Personal assets evolve according to the law of motion: $a_{j+1,t+1} = (1-\delta)(a_{jt}+b_t)+\iota_{jt}$. The gross rate of capital depreciation is $(1-\delta)$.[25] Finally, we assume that $a_{J+1,t+1}=0$ for all t, so that everyone wishes to consume all of their assets prior to entering the terminal phase of life.

E. Income & Budget Constraint

Working-age households are endowed with one unit of labor, which they supply inelastically and earn after-tax labor income $w_t\eta_j(1-\tau_t)$, where w_t is the economy-wide average wage, η_j captures the hump-shaped life-cycle wage profile following Hansen (1993), and τ_t is the net tax on labor earnings, which the government requires to fund the pensions and health spending of retirees. Households also earn income on capital holdings, $r_t(a_{jt}+b_t)$, as well as dividends, π_{jt}, from the economic profits of firms operating under monopolistic competition. Agents receive dividends in proportion to their period-t asset holdings: $\pi_{jt} = (a_{jt}+b_t)\Pi_t/K_t$. Note that $K_t = K_{ht}+K_{ct}$ is the aggregate capital level. This ensures that dividends enhance chosen investment and that the wealth distribution over the life cycle is realistic.

Assume the price of non-health-services consumption is the numeraire each period. For working-age adults the budget constraint is

$$c_{jt} + p_t\, h_{jt} + a_{j+1,t+1} \leq w_t\eta_j(1-\tau_t) + R_t(a_{jt}+b_t) + \pi_{jt}, \qquad j < J_R. \tag{12}$$

Gross returns net of depreciation are $R_t = r_t + 1 - \delta$. Upon retirement households no longer work. Their labor income is then supplanted by transfers, T_t. All individual elderly agents receive the same transfer in a given period. The budget constraint of retirees is

$$c_{jt} + p_t\, h_{jt} + a_{j+1,t+1} \leq T_t + R_t(a_{jt}+b_t) + \pi_{jt}, \qquad j \geq J_R. \tag{13}$$

F. Choices & Optimization

Households face only idiosyncratic uncertainty over survival. Sectoral TFP's driving the evolution of prices are assumed to grow deterministically. Further,

[25] Note that bequests are assumed to be received at the beginning of the period prior to depreciation occurring.

N_{1t} and z_t grow deterministically at constant rates, while ζ_{jt} is assumed to grow deterministically as well but at rates that may or may not be constant. Finally, because households know how the population distribution will evolve, they also know how b_t will evolve under the assumption they know all age-specific policy functions. The only endogenous state variable is a_{jt}.

Households choose c_{jt}, h_{jt}, and $a_{j+1,t+1}$ to solve the following recursive optimization problem subject to their age-specific budget constraint:

$$\mathcal{V}_{jt}(a_{jt}) = \max_{c_{jt},h_{jt},a_{j+1,t+1}} \Big\{ u(c_{jt}) + \beta\, s_{jt}(h_{jt})\, \mathcal{V}_{j+1,t+1}(a_{j+1,t+1}) \Big\}, \tag{14}$$

where the survival rate, s_{jt}, is a function of health expenditure, h_{jt}, from (9).

The consumption Euler equation is the usual one, except the discount rate is proportional to the consumer's idiosyncratic survival risk:

$$u'(c_{jt}) = \beta s_{jt}(h_{jt}) R_{t+1} u'(c_{j+1,t+1}), \qquad j < J. \tag{15}$$

Additionally, consumers tradeoff today's non-health consumption against health-services consumption according to

$$p_t u'(c_{jt}) = \beta \frac{\partial s_{jt}}{\partial h_{jt}} \mathcal{V}_{j+1,t+1}(a_{j+1,t+1}), \qquad j < J. \tag{16}$$

The left-hand side of (16) is the marginal utility of non-health consumption in units of health-services, while the right-hand side is the consumer's value of health-services consumption. The right-hand side can be interpreted as the welfare value of future consumption (conditional upon survival) weighted by the marginal increase to survival that results from an additional unit of health-services consumption. Alternatively, we can write an equivalent expression to (16) in terms of the marginal utility of age $j+1$ consumption:

$$\frac{\partial s_{jt}}{\partial h_{jt}} \mathcal{V}_{j+1,t+1}(a_{j+1,t+1}) = p_t s_{jt}(h_{jt}) R_{t+1} u'(c_{j+1,t+1}), \qquad j < J. \tag{17}$$

Expression (17) says that the annuity value of consumption tomorrow conditional upon survival must be equal to the lifetime value of all future consumption weighted by the marginal gain to survival due to additional health-services consumption. Health spending today thus trades off both present and future non-health consumption. Finally, note that in their last period of life ($j = J$), consumers set $h_{Jt} = 0$, $a_{J+1,t+1} = 0$, and eat all of their wealth as non-health consumption.

G. Governmental Transfers

The government taxes labor income and rebates it to retirees in the form of lump-sum transfers designed to cover both medical and non-medical expenditure. In each period we re-distribute labor-income tax receipts evenly to all individual retirees: $T_t = (w_t \tau_t \sum_{j=1}^{J_R-1} N_{jt}\eta_j)/\sum_{j=J_R}^{J} N_{jt}$. The government budget constraint then clears by construction.

To close out the model description, note that the aggregation and market clearing conditions, along with a definition of the equilibrium, are featured in Appendix A.

III. Simulation Exercises

We calibrate the model's transition path to an observable time series of data following methods deployed in Krueger and Ludwig (2007). Our primary calibration goal is to match data series over the period 1960-2015 for 1) the relative price of health services, 2) life expectancy, 3) the health share of total expenditure, 4) the share of aggregate capital devoted to the production of health services, and 5) the share of aggregate labor devoted to its production.[26,27] We fully calibrate the model separately under five different assumptions designed to reflect the four different decomposition exercises from Section I. Our first two calibrations each use estimated markup growth rates from the baseline decomposition of Section I while holding α_c fixed over time. In calibration (1), which is our "baseline" calibration, we take estimated markup growth rates from Section I and allow them to grow out to 2100 at the average annual rate from 1955-2015 of 1.7%. In calibration (2) we also use the estimated markup growth rates from Section I, but we assume that relative markups stop growing after 2015. Calibration (3) is an analogous calibration to (2), except we use the markup growth rates from the time-varying α_{ct} decomposition, while also (naturally) allowing for α_{ct} to time vary. In calibration (4) we use relative markups from Horenstein and Santos (2019) and constant α_c. Calibration (5) assumes relative markups do not grow but are fixed over time at the level we estimate for the year 1955, while also (again) holding α_c fixed.

For most of this section, we focus our results on the baseline calibration fixed α_c and baseline markup estimates from Section I growing out to 2100, though we do engage in a few model comparisons along the way. Following calibration, we engage in several counterfactual exercises, fixing various variables of interest in order to understand how sensitive model outputs are to the forces driving the rise

[26]Note that we abstract from comparing our simulations against data for 2020, intentionally. Our model (and this paper) is about what forces are driving long-run changes to the composition of the economy as it pertains to the role of health services. Short-term fluctuations are not the key focus. To this point the 2020 recession, however brief it was, significantly affected households' access to certain essential services, health care being one of them, thus impacting its aggregate share of expenditure.

[27]We discuss the data we use to get shares of capital and labor in health production in Supplemental Appendix B.5.

in the health services share.[28] For robustness we also examine the relationship between the rising health share and aggregate output (GDP) growth, which is not directly targeted in the calibration.

A. Calibration

The economy begins with an artificial steady state in year $t = 1$ which is set to 1950. There are $J = 16$ age groups with agents working until age 65 ($J_R = 10$) when they automatically retire. Agents enter the economy at age 20 ($j = 1$) and die automatically at the end of age 99. Agents are binned into 5-year age groups (i.e., 20-24, 25-29, 30-34, ..., 95-99). In all of our quantitative exercises, we simulate the economy's transition path forward from an artificial, initial steady state for $\mathcal{T} = 31$ (1950-2100) periods, assuming market concentration, technologies, and population grow over this period at rates we calibrate. After $\mathcal{T} = 31$ we assume that all time-varying (i.e., growing) exogenous variables stop growing, and we simulate the model for another 15 periods so that a stationary, terminal steady state is reached. We calibrate the model's parameters so that the targeted data moments match moments along the transition path, where our data observations begin in period $t = 3$ (1960) and extend out in 5-year intervals to period $t = 14$ (2015). Our solution algorithm and the calibrated age-specific and time-specific parameter sets are presented in detail in Supplemental Appendix D.

Objects Driving Growth. — Since the model is calibrated to a transition path, we must take stands on how all technologies, including health-investment efficiencies, as well as the initial population level, grow. We assume that the following objects grow at constant rates along the transition path: N_{1t}, z_t, ζ_{jt}, and A_{ct}, where we let g denote net growth rates.[29] In order to best match prices and input/output shares by sector, we let the growth rate of A_{ht} be time-varying, and we internally calibrate its values over four different eras (i.e., 1950-1970, 1975-1980, 1985-1995, and post-1995). μ_t is also time-varying but we calibrate this series directly to data using our estimates for growth rates from Section I and level estimates from Horenstein and Santos (2019). Specifically, since we cannot estimate the level of μ_t, only $\widetilde{g}_{\mu,t}$, we use the level estimate from Figure 7.b. of Horenstein and Santos (2019) for 1975 and extrapolate backwards and forwards using the growth rates from Section I. Note that Horenstein and Santos (2019) estimate that relative health-sector markups were approximately 1.5 in 1975. Thus, in all of our different calibrations, $\mu_{1975} = 1.5$ though estimates for μ_t before and after 1975 will differ depending on the growth rates we estimate, or in the case

[28] Our counterfactuals focus on what forces have primary driven health-sector structural change. In Supplemental Appendix D.7 we include an additional assessment as to what our various models predict for the evolution of health-services expenditure shares over the twenty-first century in conversation with results in Online Appendix A.3.3 of Fonseca et al. (2023).

[29] For example, let g_y denote the net growth rate of arbitrary variable y, so that $y_{t+1} = (1 + g_y)y_t$. This value is distinct from $\widetilde{g}_y$, which is the log of the gross growth rate.

of $\widetilde{g}_{\mu,t} = 0$, relative markups will always be 1.5. In our baseline calibration we assume that after 2015 μ_t grows at the average annual rate of 1.7% as estimated in the "Constant α_c" model of Section I. Further, we only have estimates of relative growth rates going back to 1955, so we assume that μ_t for 1950 is identical to that in 1955.

Calibrated markups for all five of our calibrations are presented below in panel (a) of Figure 6. Via this calibration procedure, in our baseline estimates health-sector markups are approximately 1.18 times higher than non-health-sector markups in 1955. By 2010 health-sector markups are 3.2 times higher than non-health-sector markups, with most of this increase occurring between 1980 and 1995, as can be seen by inspecting the red line in panel(a) of Figure 3. Calibrated μ_t grows the fastest between 1955-2015 for the time-varying α_{ct} model, where health-sector markups are 1.09 times non-health-sector markups in 1955 before ballooning to 3.9 times non-health-sector markups by 2010. The Horenstein and Santos (2019) relative markups (black line in panel (a) of Figure 6) grow the slowest, going from 1.10 in 1955 to 3.04 in 2010, though this is still a 176% increase over a 55-year period. Finally, in our baseline calibration (green line in panel (a) of Figure 6) where we project relative markups to continue growing at 1.7% annually, health-sector markups would be approximately 6 times non-health-sector markups by 2050.

In benchmarking tests we find that model-implied time series are sensitive to starting levels for z_1 and $A_{h,1}$. We thus also calibrate these internally, while fixing initial values for the population and consumption-sector TFP such that $N_{1,1} = A_{c,1} = 1$. In order to match our targeted time series, we have found that allowing A_{ht} to grow at different rates over different periods provides the best model fit. We thus internally calibrate both the starting level $A_{h,1}$ and different growth rates $g_{A_h,t}$ which are period dependent.

Health & Survival. — The accidental mortality rates come from data associated with the normalized rates in Figure B.3 of Supplemental Appendix B.4, but they are adjusted to accommodate our 5-year age bins. We take $\{\phi_j, \theta_j\}_j$ and $\{\zeta_{jt}\}_{j,t}$ directly from Hall and Jones (2007). We assume the time-varying, age-specific productivity parameters follow their estimated trends using their provided estimation code from the paper's supplementary files. We plot these age-specific parameter values in Supplemental Appendix D.3, though we also include the age-specific growth rates of health-investment efficiencies net of aggregate health-investment efficiency growth in panel (c) of Figure 6, below.

Calibrated Parameters & Model Fitness. — Table 1 presents the exogenously calibrated parameters, their sources, and descriptions as to what they represent. Table 2 presents the internally calibrated parameters across all five calibration procedures. Calibration (1) is that which we refer to as our "baseline" calibration, which includes constant α_c and relative markups as estimated in the corre-

sponding estimation in Section I while assumed to continue growing out to 2100. Calibration (2) is analogous to calibration (1), except we assume relative markups stop growing in 2015. Calibrations (3) through (5) are those with time-varying α_{ct}, Horenstein and Santos (2019) markups, and no relative markup growth since 1955. While the values of exogenously fixed parameters, ϕ_j, θ_j, η_j, and g_{ζ_j} are presented in Supplemental Appendix D, we include calibrated values for μ_t, A_{ht}, and composite health-consumption productivity growth, $(1 + g_z)(1 + g_{\zeta_j})^{\theta_j}$, in Figure 6 here in the main text in Section III.A below.

Our calibration targets five time series in 5-year intervals for periods ranging from $3 \leq t \leq 14$ (1960-2015): 1) relative health-services prices, p_t; 2) life expectancy, LE_t; 3) health share of aggregate consumption expenditure, $p_t H_t/(p_t H_t + C_t)$; 4) the share of aggregate capital in health-services production, $K_{ht}/(K_{ht} + K_{ct})$; 5) the share of aggregate labor in health-services production, $L_{ht}/(L_{ht} + L_{ct})$. In Table 1 externally calibrated parameters are either taken directly from sources or computed directly from the associated data series. Internally calibrated parameters in Table 2 are designed to primarily match the five time series along the growth path. As a robustness check we compare model-predicted GDP growth rates to data, as well. Note, however, that we do not directly target GDP growth in our calibration procedure (i.e., our RMSE estimates for our different calibrations' loss functions do not account for fitting the GDP growth rate), but we can still successfully match average-annual GDP growth over the 5-year intervals of our target sample for our baseline calibration.

Table 1—Externally Calibrated Parameters & Sources/Targeted Moments

Parameter	Value	What	Source
$\{\phi_j\}_j$	Supplemental Appendix D.3	Health Intercept	Hall and Jones (2007)
$\{\theta_j\}_j$	Supplemental Appendix D.3	Health Elasticity	Hall and Jones (2007)
$\{\eta_j\}_j$	Supplemental Appendix D.3	Labor Productivity	Hansen (1993)
$\{\zeta_{jt}\}_{j,t}$	Supplemental Appendix D.3	Health Productivity by Age	Hall and Jones (2007)
$\{m_{jt}^{acc}\}_{j,t}$	Supplemental Appendix B.4	Accidental Mortality	CDC, *Health, United States 2017*
$\{\mu_t\}_t$	Panel (a) of Figure 6	Markups	Horenstein and Santos (2019) & Section I
γ	2.000	Intertemporal Elasticity	Hall and Jones (2007)
α_c	0.400*	Capital Intensity, $C_t + I_t$	Horenstein and Santos (2019)
α_h	0.260	Capital Intensity, H_t	Donahoe (2000)
g_{A_c}	0.034	A_{ct} 5-yr Avg. Growth	Feenstra, Inklaar and Timmer (2015)
χ	191.999	Utility Intercept	Fixed after Benchmarking
δ	0.185	4% annual depreciation (Krueger and Ludwig, 2007)	
g_N	0.032	5-yr. growth newborns 1990-2019	
β	0.918	Annual Discount Rate of 0.983 (Convention)	
τ	0.086	1990-2000 S.S. + Medicare Average	

Note:
* When $1 - \alpha_{ct}$ is time-varying, we take estimates directly from Feenstra, Inklaar and Timmer (2015) and plot these in Supplemental Appendix B.5.

While our externally calibrated parameters are naturally picked to readily

Table 2—Internally Calibrated Parameters (All Models)

Model	(1)	(2)	(3)	(4)	(5)	
$A_{h,1}$	0.099	0.099	0.085	0.097	0.099	Health-sector TFP
$g_{A_h,1950-1970}$	0.017	0.017	0.037	0.041	0.023	5-yr Growth 1950-1970
$g_{A_h,1975-1980}$	0.042	0.042	0.099	0.109	-0.004	5-yr Growth 1975-1980
$g_{A_h,1985-1995}$	0.068	0.068	0.105	-0.013	-0.005	5-yr Growth 1985-1995
$g_{A_h,>1995}$	0.076	0.076	0.061	0.117	-0.012	5-yr Growth After 1995
ξ	42.977	42.977	58.758	37.118	35.870	Utility Scalar
g_z	0.003	0.003	0.004	0.001	0.006	5-yr Growth z_t
z_1	16.456	16.467	16.455	16.467	16.315	Health-demand Efficiency
RMSE	0.039	0.039	0.047	0.042	0.061	Loss Function

match data and/or what is typically done in the literature, our internally calibrated parameters warrant more discussion. Across models we observe little variation in estimates of health-sector TFP starting levels (1950) relative to that of the non-health-sector. We normalize $A_{c,1} = 1$, so that we estimate health-sector TFP is always approximately $8 - 10\%$ of non-health-sector TFP in 1950. Growth rates, $g_{A_h,t}$, however, vary in ways that make sense given the markups we use in our calibration. For example, calibration (3) with time-varying α_{ct} has the fastest markup growth. In calibrations (1), (2) and (3) $g_{A_h,t}$ estimates are on average greater than for models with slower growth (i.e., calibration (4) with Horenstein and Santos (2019) markups) or no growth at all (i.e., calibration (5) with $\widetilde{g}_\mu = 0$). When relative markup growth is fast, a fast growing health sector actually acts to dampen the impact of relative markup growth (this will become more apparent below), while when relative markup growth is slow, a slow-growing health sector is needed in order to match the evolution of prices. Indeed, if health-sector TFP grew more slowly in models with relatively high markup growth, such models would over-predict relative price growth.

Health-services demand parameters, ξ, g_z, and z_1 are also internally calibrated. The parameter ξ controls demand for the quantity of health services relative to non-health-services consumption. This parameter helps match the health-services share of consumption expenditure. z_1, meanwhile, scales the effect of h_{jt} on health, x_{jt}, thus directly impacting survival rates. z_1 is an important parameter to match life expectancy in the initial year of the sample and varies little across the four different calibrations. Similarly, while z_1 matches the baseline level of life expectancy, g_z helps exogenously determine how life expectancy grows. Because h_{jt} is also endogenously growing and because its rate of endogenous growth will vary across agents of different ages within a particular calibration but also across calibrations depending on what is driving relative price growth, different models will require relatively higher or lower g_z to best match the data. Still, across all models average-annual growth in g_z ranges from 0.02% to 0.12% (note that 5-year growth rates are reported in Table 2, *not* annual ones). Finally, extending the

growth of relative markups out to 2100 does very little to change the calibration compared to assuming relative markup growth stops in 2015: this can be seen by comparing calibration (1) with calibration (2), where only the parameter governing the initial aggregate health-services efficiency, z_1, significantly changes by reportable magnitudes (three significant digits). We thus estimate that health-consumption efficiency growth is very slow relative to health-sector productivity growth.

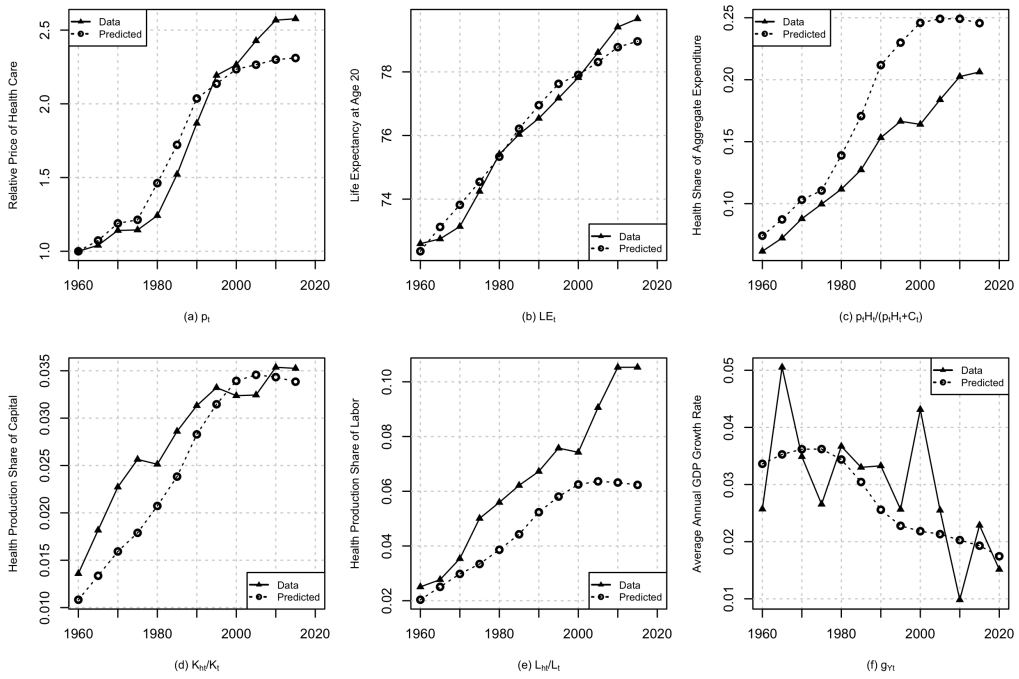

FIGURE 5. WE PRESENT PREDICTED VALUES (ROUND DOTS AND DASHED LINE) FROM THE CALIBRATED BASELINE MODEL WITH CONSTANT α_c AND RELATIVE MARKUPS GROWING OUT TO 2100 AGAINST DATA (SOLID TRIANGLES AND SOLID LINE). WE CALIBRATE THE MODEL TO THE ECONOMY'S TRANSITION PATH FOR 5-YEAR PERIODS RANGING FROM 1960 TO 2015, DIRECTLY TARGETING FIVE TIME SERIES DISCUSSED IN THE MAIN TEXT AND FEATURED IN PANELS (A) THROUGH (E). IN PANEL (F) WE SHOW HOW WELL THE MODEL PREDICTS THE NON-TARGETED DECLINE IN AVERAGE GDP GROWTH RATES. IN SUPPLEMENTAL APPENDIX D.4 WE INCLUDE ANALOGOUS PLOTS FOR CALIBRATIONS (2) THROUGH (5).

To assess model fitness we present plots of predicted time series against data in Figure 5 for the baseline calibration (1) which fits the data best as measured by root mean squared error (RMSE). Note that panels (a) through (e) feature fits for time series that are directly targeted in our internal calibration. Panel (f), meanwhile, shows how the model performs in fitting non-targeted, 5-year

GDP growth averages: it does very well. We include analogous plots for all other calibrations in Supplemental Appendix D.4, while discussing the subtle differences between the baseline model fit and the other model fits here in the main text.

Our calibration sufficiently captures the qualitative increase in relative prices, life expectancy, and expenditure shares over time, as well as the increase in the shares of inputs (capital and labor) devoted to the production of health services. In the spirit of quantitative macroeconomics, our goal is to find a set of parameters that can yield model outcomes which reasonably adhere to the general patterns observed in the data. We then seek to simulate the model, using it as a laboratory to both understand which of the growth factors have primarily contributed to rising prices and spending shares *and* the model's joint implications for structural change, aging, and GDP growth. We are thus satisfied with the baseline calibration, especially given the difficulty in matching the increase in life expectancy along the growth path in the Hall and Jones (2007) setting.

Denote aggregate GDP by Y_t, which is $Y_t = p_t H_t + C_t + I_t$. Average annual GDP growth within 5-year intervals is then $g_Y = (Y_t/Y_{t-1})^{1/5} - 1$. The calibrated model predicts the average decline in growth rates we observe from mid-century to 2020 well, as can be seen in panel(f) of Figure 5. The quality of fit of the predicted model lends credibility to our quantitative results, particularly our counterfactuals, which isolate the different growth channels to understand how their predictions deviate from the predicted baseline.

At the bottom of Table 2 we compare RMSE across the different models. Both the baseline calibration and the calibration with constant α_c and relative markups not growing after 2015 fit the data best, followed by a calibration with Horenstein and Santos (2019) markups. In Supplemental Appendix D.4 we show the fitness plots for the four alternative calibrations. Calibration (2) has an almost identical fit to the baseline calibration. The model with Horenstein and Santos (2019) markups (calibration (4)) has a qualitatively similar fit as the baseline model but performs slightly worse by under-fitting relative price growth and less accurately predicting growth in factor-input shares. Meanwhile, when there is no relative markup growth (calibration (5) with $\widetilde{g}_\mu = 0$), we fit the life-expectancy and relative expenditure series well, but fitted relative prices grow at a much slower pace than data, and factor-input shares are off. The model with time-varying α_{ct} (calibration (3)) is a more mixed bag: it most accurately fits relative price growth but misses factor-input shares badly. Further, it vastly under-predicts GDP growth in our validity check against non-targeted data.

Finally, in order to understand how much our markup growth estimates from Section I drive model outcomes, in Supplemental Appendix D.5 we re-simulate calibration (2) using all of the parameters from Tables 1 and 2 except that we replace μ_t with relative markups taken directly from Horenstein and Santos (2019). Not surprisingly, relative prices are most affected by this change, though we are also less able to accurately match GDP growth rates. We conclude that the combination of our more swiftly growing relative markups, our estimates for health-sector

TFP growth, and demand effects are all needed to best explain the evolution of health-sector structural change.

Discussion of Key Parameters. — Figure 6 plots the calibrated time series of μ_t and $A_t = A_{ct}/A_{ht}$, as well as age-specific, health-consumption technology growth rates for all five calibrations. $\{\mu_t\}_t$ is exogenously calibrated by first taking the estimated relative markup level of 1.5 from Horenstein and Santos (2019) then using growth rates for μ_t as estimated in Section I and iterating backwards to 1955 and forwards to 2015 to get the calibrated time series. After 2015 we take the average annual growth rate from 1955-2015 of 1.7% and simulate forward to 2100. Relative TFP, A_t, is a combination of externally calibrated estimates for A_{ct} from Feenstra, Inklaar and Timmer (2015) and internally calibrated estimates of A_{ht}, where the series is normalized such that $A_1 = 1/A_{h,1}$ in 1950. Age-specific health-efficiency growth rates are also a combination of externally and internally calibrated parameters. We take age-specific values of ζ_{jt} from Hall and Jones (2007) who assume constant g_{ζ_j} and calibrate g_z and z_1 to match life expectancy. The age-specific values in panel (c) of Figure 6 thus differ across model calibrations solely because estimates of g_z differ. Notice that in Figure 6, time series estimates for calibrations (1) and (2) are almost identical, except for relative markups after 2015, reflecting the fact that all of the other estimated parameters in calibrations (1) and (2) are very similar.

Since calibration of μ_t follows directly from Section I and since calibration of $z_t\zeta_{jt}^{\theta_j}$ almost exclusively affects life expectancy (and there are few differences in the calibration profile across models), we focus our attention in this section on the calibration of A_t. From Table 1 one can see that non-health-sector TFP grows at a 5-year average rate of 3.4%, which amounts to an annual average of 0.7%. Looking at Table 2, in our baseline calibration (green line in panel (b) of Figure 6) A_{ht} grows slower than A_{ct} until 1975, then faster thereafter. We estimate that annual health-sector TFP growth is 0.3% from 1950-1970, 0.8% from 1975-1980, 1.3% from 1985-1995, and 1.5% thereafter. These estimates are higher than those from our decomposition exercise in Section I and at the high end of estimates from the literature, but close to health-sector labor productivity estimates from Cylus and Dickensheets (2007-2008). Model calibrations using markups inferred from time-varying labor shares and Horenstein and Santos (2019) yield similar results: notice that both the blue and black lines in panel (b) of Figure 6 fall at approximately the same long-run pace as in the baseline calibration.

Why do our calibrated estimates of A_{ht} in the full GE model differ from those in the growth-accounting decomposition? Recall that in Section I we do not directly control for the changing composition of demand, but rather all GE effects operate through variation in the observed (from data) allocation of health-sector inputs. However, in a full GE setting the input levels, K_{ht} and L_{ht}, are endogenous, but our growth accounting exercises in Section I fail to account for such endogeneity, simply taking their values as given in the data. Clearly, values of A_{ht} and A_{ct} (as

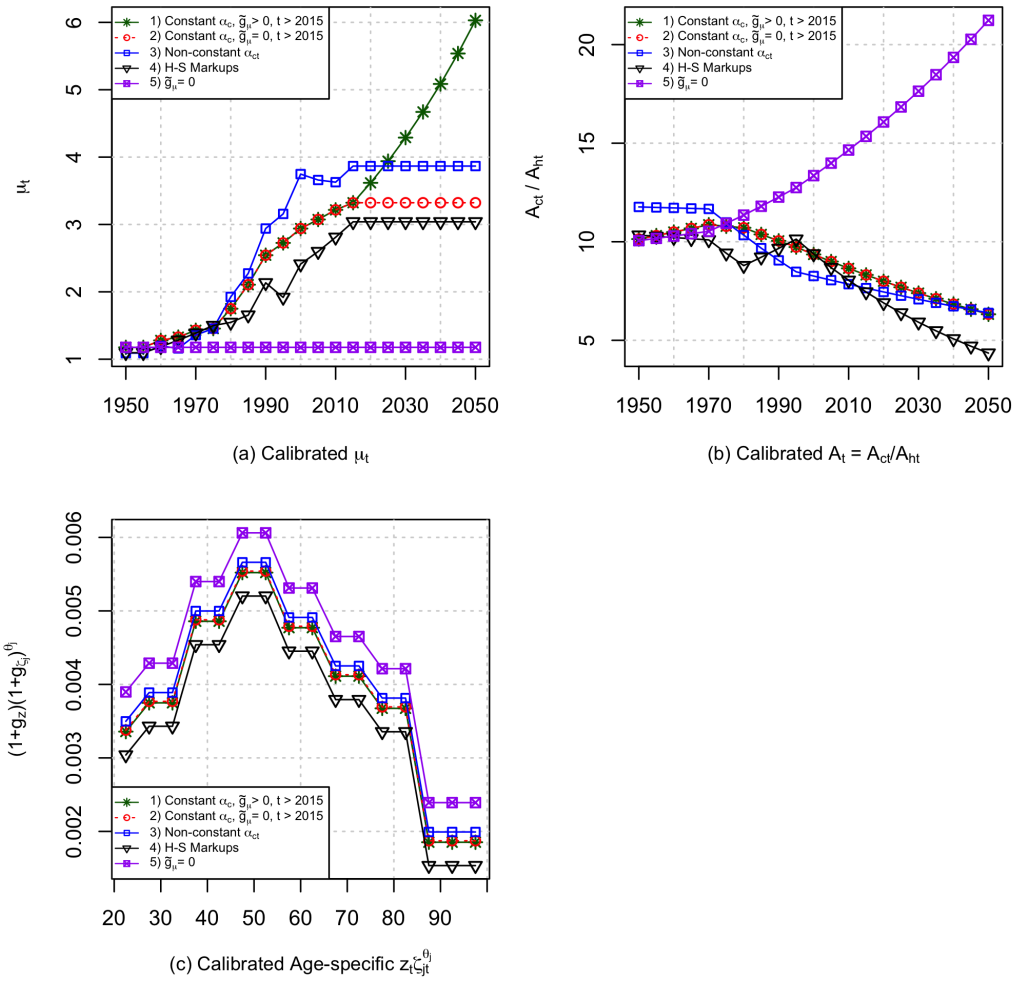

FIGURE 6. PANEL (A) SHOWS THE CALIBRATED VALUES OF μ_t. PANEL (B) PRESENTS A_{ct}/A_{ht}. PANEL (C) SHOWS THE COMBINED, AGE-SPECIFIC 5-YEAR GROWTH RATES FOR THE HEALTH-CONSUMPTION EFFICIENCY COMPOSITE, $z_t\zeta_{jt}^{\theta_j}$. IN ALL PANELS THE BASELINE CALIBRATION IS IN GREEN, THE CALIBRATION WITH CONSTANT α_c BUT RELATIVE MARKUP GROWTH CEASING IN 2015 IS IN RED, TIME-VARYING α_{ct} IN BLUE, HORENSTEIN AND SANTOS (2019) MARKUPS IN BLACK, AND $\widetilde{g}_\mu = 0$ IN PURPLE.

well as μ_t) will also affect the allocation of factor inputs across sectors, so that our GE summary statistics used in Section I actually embed variation in relative markups and TFP's. The effects of unbalanced technical change described in Section I are thus only partial in nature and fail to account for GE feedback effects via the re-allocation of factor inputs. Estimating relative productivities in a full GE model which allows for factor-input re-allocation leads to results which show

that relative productivities (i.e., A_t) backed out of a simple growth-accounting exercise are biased upward. Since we are the first paper to estimate a full GE model of health-sector structural change which accounts for rising relative markups and the re-allocative effects of changes to both the composition of demand and relative sectoral TFP's, it should not be controversial that our TFP estimates depart from those in other studies which have failed to account for markups and GE effects.

Still, to more precisely illustrate how the GE effects operate through the allocation of relative inputs, consider a version of (8) where $\widetilde{g}_{L_h,t} - \widetilde{g}_{K_h,t}$ is replaced by the difference in growth rates of relative prices, $\widetilde{g}_{r,t} - \widetilde{g}_{w,t}$. Further, suppose just for simplicity that $\widetilde{g}_{r,t} = 0$, so that the long-run capital rate of return is constant. Then, (8) becomes

$$\widetilde{g}_{p,t} = \widetilde{g}_{\mu,t} - (\alpha_h - \alpha_c)\widetilde{g}_{w,t} + \widetilde{g}_{A,t}, \tag{18}$$

where we use relative TFP, $\widetilde{g}_{A,t} = \widetilde{g}_{A_c,t} - \widetilde{g}_{A_h,t}$. From a pure accounting perspective if $\widetilde{g}_{w,t} > 0$ (real wages are growing), $\alpha_h - \alpha_c < 0$, and further $\widetilde{g}_{p,t} < \widetilde{g}_{\mu,t}$ (as we estimate), then it must be that health-sector productivity is growing faster than non-health-sector productivity (i.e., $\widetilde{g}_{A,t} < 0$). Clearly, in this simple environment we would need $\widetilde{g}_{A,t} < 0$ and thus $\widetilde{g}_{A_h,t} > \widetilde{g}_{A_c,t}$ to counter the positive effects of wage growth and relative markup growth. We should caution, though, that like the exercises in Section I, this thought experiment fails to capture how variation in TFP's will *feedback* into the model to affect endogenous factor input prices. There are any number of ways such GE feedback effects could work. For example, increasing health-sector productivity would impact the efficiency by which health-services firms can produce output and thus the prices they would charge to be profitable. The degree to which they would tradeoff increasing quantities of production versus increasing prices would depend on the elasticity of demand for health services, which is an age-dependent object that depends on the continuation value of consumption and the elasticity of survival with respect to health-services consumption, as can be seen by inspecting the health Euler equation in (16). Thus, while this simple exercise helps clarify how GE effects may operate in this model, in order to truly understand how the different growth channels affect prices, we must simulate GE outcomes under different counterfactual growth regimes along the entire growth path.

In light of falling (not rising) A_t, we can indirectly glean the validity of our model estimates and, by extension, the validity of the GE effects we will discuss in the next section by assessing whether the equilibrium net capital-rental rate, $r_t = R_t - 1$, generated by the model is realistic. Our baseline model yields annualized capital rental rates between 8% and 10%, close to post-World War II pre-tax ranges estimated by Gomme, Ravikumar and Rupert (2011), Caballero, Farhi and Gourinchas (2017), and Jordà et al. (2019) for public equities and private wealth. This is important because, as we have discussed extensively,

GE effects can be summarized by input prices. But, when examining the GE effects in Section I, we are actually looking at the simultaneous effects that both re-allocation due to variation in supply-side objects (i.e., μ_t and A_t) and the changing composition of demand have on prices. To understand the pure (total) effects of relative markups, unbalanced technical change, and the composition of demand on aggregate health-sector outcomes, we must simulate the full GE model while separately fixing the various channels of exogenous growth and structural change. We now do just that.

B. Counterfactuals

We consider two different constellations of counterfactual exercises, each designed to help better understand how the various drivers of growth and structural change have contributed to the health sector's economic transformation. Here in the main text we only show counterfactual simulations from calibration (1). Supplemental Appendix D.6 contains results from counterfactuals run on the alternative calibrations. In one set of counterfactuals, we fix the growth drivers one at a time, negating the effect of one particular channel while allowing the others to still determine outcomes. In the other set of counterfactuals, we fix *all but one* driver of growth, showing how much that one particular channel itself (alone) affects outcomes. The three main drivers of growth we consider are as follows: 1) changes to the composition of demand, summarized by exogenous changes to the population, g_N, and the age-specific health-consumption efficiencies, $z_t\zeta_{jt}^{\theta_j}$;[30] 2) changes to relative markups, μ_t; 3) unbalanced TFP variation, A_t.

Figures 7 and 8 present the two different sets of counterfactual simulations. Figure 7 shows the various series when we fix one (and only one) channel at a time, while Figure 8 shows the same plot, fixing *all but one* channel at a time. Meanwhile, Table 3 shows the total change in the targeted moments from 1960-2015 in the different baseline counterfactuals relative to the baseline model's predictions. In Table 3 simulations (3) through (5) are those featured in Figure 7, while simulations (6) through (8) are those in Figure 8.

Several conclusions jump out at first glance. Growth in the population and health-care consumption efficiency (i.e., demand effects) mostly impact life expectancy and GDP growth, with only minor impacts for health-services inputs and outputs (see the blue lines in Figure 7). Without increasing health-consumption efficiency, demand for h_{jt} is relatively high, but GDP growth being relatively low shows that consumers are sacrificing non-health-care consumption and investment in order to devote a greater share of wallet to health-services consumption, each unit of which does not go as far as it would had $z_t\zeta_{jt}^{\theta_j}$ been growing at the cal-

[30] Note that variation in $z_t\zeta_{jt}^{\theta_j}$ will directly impact demand for health services, as it changes the consumer's Euler equation in (16). Thus, variation in $z_t\zeta_{jt}^{\theta_j}$ will lead to variation in the consumer's price elasticity of demand for health services.

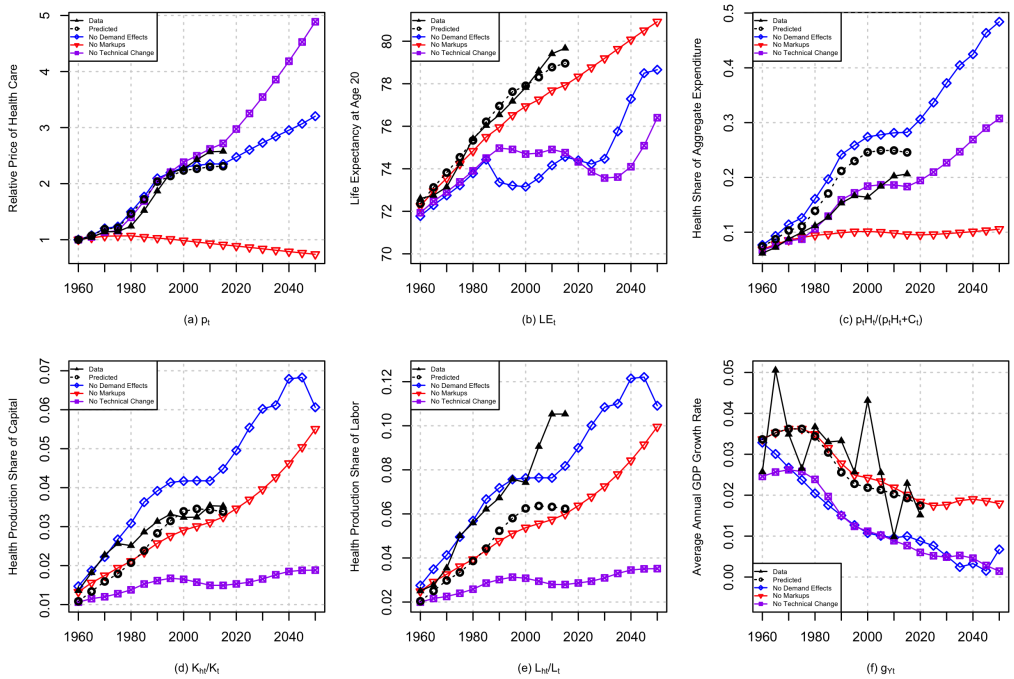

FIGURE 7. THIS FIGURE IS ANALOGOUS TO FIGURE 5, EXCEPT HERE WE INCLUDE THE COUNTERFACTUAL PREDICTIONS AFTER WE TURN OFF VARIOUS GROWTH CHANNELS. BASELINE PREDICTED VALUES ARE ROUND DOTS, DATA ARE TRIANGLES, AND THE COUNTERFACTUAL PREDICTIONS ARE LINES. WE SHOW ALL SERIES OUT TO 2050, THOUGH VARIABLES ARE ALLOWED TO GROW (FOR COMPUTATIONAL REASONS) UNTIL 2100. BLUE VALUES ARE TIME SERIES WHEN $g_N = g_z = g_{\zeta_j} = 0$. RED VALUES ARE TIME SERIES WHEN $g_{\mu,t} = 0$. PURPLE VALUES ARE TIME SERIES WHEN $g_{A_t} = 0$.

ibrated rate. Meanwhile, markup growth is entirely responsible for rising prices and, to a lesser extent, rising health-services expenditure shares. Indeed, looking at panels (a) and (c) of Figure 8, freezing both demand changes and markup growth (brown lines) leads to a *decline* in relative prices and a slightly rising share of expenditure on H_t, but this is entirely due to consumers wanting to buy greater *quantities* of h_{jt}, since each dollar of health investment does not go as far as it otherwise would if $z_t\zeta_{jt}^{\theta_j}$ were, again, growing at the calibrated rate. Variation in relative TFP thus helps drive the re-allocation of resources toward the production of health services, yet absent TFP changes, relative health prices would still rise because of markup growth, and this would be almost entirely the driver of rising health-services expenditure shares, not increasing quantities demanded (see the purple lines in panels (a) and (c) of Figure 7). Unbalanced TFP growth, as estimated in our calibration, thus ensures that some of the rising health-services share is not just due to prices but also due to increases in

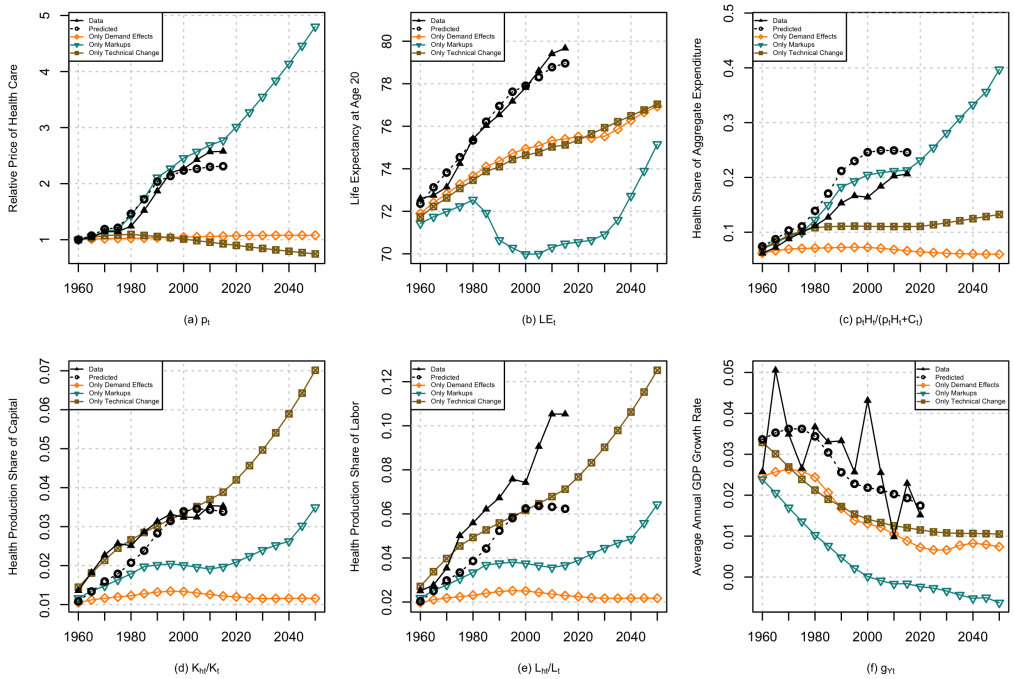

FIGURE 8. THIS FIGURE IS ANALOGOUS TO BOTH FIGURES 5 AND 7, EXCEPT NOW WE INCLUDE THE COUNTERFACTUAL PREDICTIONS AFTER WE LEAVE *on* ONLY A SINGLE GROWTH CHANNEL. BASELINE PREDICTED VALUES ARE ROUND DOTS, DATA ARE TRIANGLES, AND THE COUNTERFACTUAL PREDICTIONS ARE LINES. ORANGE VALUES ARE TIME SERIES WHEN $g_{\mu,t} = g_{A_t} = 0$. TEAL VALUES ARE TIME SERIES WHEN $g_N = g_z = g_{\zeta_j} = g_{A_t} = 0$. BROWN VALUES ARE TIME SERIES WHEN $g_N = g_z = g_{\zeta_j} = g_{\mu t} = 0$.

aggregate health-services output, H_t. Indeed, in Supplemental Appendix D.7 we show that predicted future increases to the health-services share of expenditure, while partially attributable to rising relative prices from rising relative markups, can also be attributed to increasing H_t due to continuing unbalanced technical change that favors the health sector. Absent relatively faster health-sector TFP growth, life-expectancy also fails to significantly improve as total consumption of health services remains flat despite consumers spending a greater share of wallet on health services due to markup-driven price increases.

It should not be surprising that relative price growth is most sensitive to markups, given the results from Section I. Meanwhile, in simulation (8) (refer to Table 3 and brown lines in Figure 8), when we only allow for TFP's to change, relative prices actually *decline*. Whereas in Section I we found that TFP changes were at most offset by GE effects, if not at least partially dampening (due to being embedded in the GE effect), now controlling for the impact of TFP changes on those very re-allocative GE effects, we find that TFP changes do ac-

tually dampen relative price growth. Note, though, the brown lines in panels (d) and (e) of Figure 8: when the economy is only evolving due to productivity growth, the health sector itself is growing in real terms, and this is what drives up (though only slightly) health-sector expenditure shares in panel (c). TFP changes also have an impact on life expectancy, which should not be surprising since when we shut off growth in A_t, we are also shutting down growth in the levels of A_{ct} and A_{ht} and thus growth in aggregate income. Looking at the purple line in panel (b) of Figure 7, life expectancy rises until the 1980s and then stagnates, absent only TFP growth. Further, looking at panel (f) of Figure 7, notice that GDP growth is slower without unbalanced variation in TFP. Without changing TFP's the main driver of income growth is exogenous population growth, g_N, but income effects coupled with a growing efficiency of health investment are not enough to bolster life expectancy: consumers also need the health-services sector to become relatively more productive (see orange line in panel (b) of Figure 8). Expenditure share growth as evidenced by the purple line in panel (c) of Figure 7 is almost entirely due to rising prices, which is apparent by examining the purple lines in panels (d) and (e) of Figure 7 that show that the health sector is not growing in terms of its share of inputs. Thus, unbalanced TFP variation that is skewed toward relatively high health-sector TFP growth has contributed positively to both GDP growth and life expectancy increases, while also ensuring that part of the reason the health-sector share of expenditure is rising is due to real, productive gains to health-sector output, not just nominal price increases.

Table 3—Counterfactual Simulations Relative to Model Predictions

		Targeted Moments, Change from 1960-2015				
	Counterfactual	p_t Growth*	LE Lost$^{\%}$	Health Share$^{\$}$	$K_{ht}/K_t{}^{\$}$	$L_{ht}/L_t{}^{\$}$
(1)	Data	157.687	-0.711	0.145	0.022	0.080
(2)	Predicted Baseline	131.050	0.000	0.172	0.023	0.042
(3)	$g_{\zeta_j} = g_z = g_N = 0$	135.047	4.401	0.206	0.030	0.054
(4)	$\mu_t = \mu_1, \forall t$	-9.049	1.030	0.028	0.019	0.035
(5)	$g_{A_c} = g_{A_h} = 0$	171.834	4.199	0.116	0.004	0.008
(6)	$g_{A_c} = g_{A_h} = 0$, $\mu_t = \mu_1, \forall t$	6.420	3.557	0.004	0.002	0.003
(7)	$g_{A_c} = g_{A_h} = g_{\zeta_j} = g_z = g_N = 0$	177.190	8.481	0.144	0.008	0.015
(8)	$g_{\zeta_j} = g_z = g_N = 0$,$\mu_t = \mu_1, \forall t$	-7.341	3.835	0.040	0.024	0.044

Note:
* Percent change from 1960-2015.
% Life expectancy years lost relative to 2015 baseline prediction. Negative values imply life-years gained.
$ Percentage point difference from 1960-2015.

The intersection of life-expectancy gains and output growth has been of interest to macroeconomists for decades.[31] In our model these two forces are endogenously

[31]See, for example, the literature on aging and GDP growth (Krueger and Ludwig, 2007; Prettner, 2013; Backus, Cooley and Henriksen, 2014; Cooley and Henriksen, 2018; Cooley, Henriksen and Nusbaum,

linked and depend, in complicated ways, on how the structure of the multi-sector economy has evolved. What can our model tell us about the forces that drive life-expectancy gains and how these forces also affect both sectoral and aggregate output growth rates? To get life-expectancy gains that match the data requires that all of the growth drivers be turned on: this is evident by inspecting the counterfactual plots in panel (b) of both Figures 7 and 8. Life-expectancy gains are mostly driven by gains to income and improvements in health-consumption efficiency, as evidenced by the fact that when we shut down both of these channels in panel (b) of Figure 8 (teal line) life expectancy actually declines. The teal line in panel (b) of Figure 8 demonstrates what would have happened in a world in which markups grew but nothing else changed, illustrating how unbalanced technical change that favors the health-care sector has helped offset the negative effects of markup growth. GDP grows slowest in this simulation, as well, slowing health investment relative to the baseline prediction. This signals the status of health services as a luxury which, in turn, speaks to the importance of income growth for driving up life expectancy: as consumers get richer, they are willing to invest greater fractions of resources to health investment, driving up longevity gains.

A similarly important driver of gains to life expectancy appears to be increases to the efficiency of health investment, $z_t \zeta_{jt}^{\theta_j}$, which can be seen by looking at the orange line in panel (b) of Figure 8. In this simulation, when all but changes to the population level and health-investment efficiency are shut down, gains to longevity are almost entirely exogenous, since health-services expenditure shares and inputs are flat. Still, without the care-efficiency and demographic effects as in panel (b) of Figure 7, life expectancy eventually rises, though more slowly. The increase in life expectancy in this simulation, though, is entirely endogenous: consumers are getting richer and want to live longer to enjoy more consumption, but they must actually purchase longevity gains through increased health investment rather than rely on increases to such investments' efficiency improving their survival abilities. The blue lines in Figure 7 thus represent how the economy would have evolved had supply-side factors changed but society had failed to make the consumption environment healthier (e.g., smoking bans, pollution mitigation, advances to preventative care and health knowledge independent of A_{ht}, etc.). When shutting down population growth and $z_t \zeta_{jt}^{\theta_j}$, life expectancy grows more slowly despite health spending rising slightly more quickly because every dollar of new health spending does not go as far as in the baseline model.

The structural transformation of the health sector, as represented by rising relative prices and health expenditure shares, is almost entirely driven by rising relative markups and unbalanced TFP changes. The orange simulations in Figure 8 affirm this: when the model economy is driven by only population growth and improvements to health efficiency, the relative price grows by only 6.4%,

2019; Kydland and Pretnar, 2019; Maestas, Mullen and Powell, 2023).

compared to 131% in the predicted baseline, the health share of expenditure increases by 0.004 compared to 0.171 in the predicted baseline, and the shares of labor and capital inputs devoted to health-services production also barely increase (see simulation (6) in Table 3). Meanwhile, with only markup growth (teal simulations in Figure 8) we get mostly structural change driven by price growth; with only unbalanced TFP changes we get entirely structural change driven by changes to the composition of output. Further, turning off the primary driver of price growth (red line in panel (f) of Figure 7) does nothing to GDP growth and in fact increases growth rates ever-so-slightly relative to the predicted baseline. When markups are turned off, structural change becomes totally driven by the rise in aggregate health quantities, H_t, as can be seen by inspecting the red lines in Figure 7, noting that expenditure shares rise but prices fall. Note that the red lines in panel (f) of Figure 7 which show how the economy would evolve when just markup growth is shutdown mirror the brown lines in panel (f) of Figure 8 when *only* TFP's are allowed to vary, further highlighting the minimal role that changes to the composition of demand play in driving structural transformation. GDP growth is slightly lower in panel (f) of Figure 7 when only TFP's are allowed to vary simply because the population is not also growing, and so the supply of labor is not also increasing.

We draw several big conclusions from our decompositions via simulation. First, the health sector is getting relatively more productive and this has contributed to an increase in expenditure shares as a result of increasing output. Second, increasing health-sector market concentration has led to increasing relative markups which is entirely responsible for rising relative prices. Third, there is only a very minor role for changes to the composition of demand driving health-sector structural change.

IV. Conclusion

We have shown why relative health-services prices and the health share of spending have risen in the U.S. since the mid-twentieth century. We have demonstrated that rising relative markups have only marginally impacted life expectancy. We have shown that health-sector structural change is driven by two disparate factors: 1) nominal changes to relative prices from rising relative markups; 2) real output increases due to improving health-sector productivity. Further, we have found only a minor role for demand effects to contribute to the structural transformation of the health sector.

Our findings have broader implications for policy. While increasing market concentration is primarily to blame for rising relative prices, unbalanced technical change acts as a countervailing force from which consumers benefit from increased health-sector output leading to health and life-expectancy improvements. Policymakers could use many tools at their disposal to curb price growth, namely anti-trust regulations and the continuing encouragement of technology adoption in the health sector to improve productive efficiency and leverage the sector's

relatively fast-growing TFP to benefit consumers. We should caution, though, that our findings result from a definition of the health-services sector that is considerably broad, including both service providers, pharmaceutical companies, and other producers of technologies and equipment. The welfare implications of policies targeting market power in the sector will be different whether such policies attempt to curb the market power of care providers versus the market power of producers of equipment and new technologies who invest in research and development. To this extent, our exercise also abstracts from *where* sectoral productivity improvements are coming from (e.g., hospitals or pharmaceuticals). Future research is needed to more precisely understand the drivers of aggregate health-sector productivity growth, which could help policymakers faced with trading off how to curb firms' pricing power without adversely affecting innovation incentives.

REFERENCES

Adler, Loren, Conrad Milhaupt, and Samuel Valdez. 2023. "Measuring private equity penetration and consolidation in emergency medicine and anesthesiology." *1*, 1.

Alonso-Carrera, Jaime, Jordi Caballé, and Xavier Raurich. 2015. "Consumption composition and macroeconomic dynamics." *The B.E. Journal of Macroeconomics*, 15.

Andreyeva, Elena, Atul Gupta, Catherine Ishitani, Malgorzata Sylwestrzak, and Benjamin Ukert. 2024. "The Corporatization of Independent Hospitals." *Journal of Political Economy Microeconomics*, 2(3).

Backus, David, Thomas Cooley, and Espen Henriksen. 2014. "Demography and low-frequency capital flows." *Journal of International Economics*, 92.

Barkai, Simcha. 2020. "Declining labor and capital shares." *The Journal of Finance*, 75(5): 2421–2463.

Bates, Laurie, and Rexford Santerre. 2013. "Does the U.S. health care sector suffer from Baumol's cost disease? Evidence from the 50 states." *Journal of Health Economics*, 32: 386–391.

Baumol, William. 1967. "Macroeconomics of Unbalanced Growth: The Anatomy of Urban Crisis." *The American Economic Review*, 57(3).

Baumol, William, Sue Anne Batey Blackman, and Edward Wolff. 1985. "Unbalance growth revisited: asymptotic stagnancy and new evidence." *The American Economic Review*, 75(4): 806–817.

Behrens, Kristian, Giordano Mion, Yasusada Murata, and Jens Suedekum. 2020. "Quantifying the Gap Between Equilibrium and Optimum under Monopolistic Competition." *The Quarterly Journal of Economics*, 135(4): 2299–2360.

Blumenthal, David, Kristof Stremikis, and David Cutler. 2013. "Health Care Spending — A Giant Slain or Sleeping?" *The New England Journal of Medicine*, 369(26).

Caballero, Ricardo J, Emmanuel Farhi, and Pierre-Olivier Gourinchas. 2017. "Rents, technical change, and risk premia accounting for secular trends in interest rates, returns on capital, earning yields, and factor shares." *American Economic Review*, 107(5): 614–620.

Clemens, Jeffrey, and Joshua D. Gottlieb. 2014. "Do Physicians' Financial Incentives Affect Medical Treatment and Patient Health?" *American Economic Review*, 104(4): 1320–1349.

Cooley, Thomas, and Espen Henriksen. 2018. "The demographic deficit." *Journal of Monetary Economics*, 93: 45–62.

Cooley, Thomas, Espen Henriksen, and Charlie Nusbaum. 2019. "Demographic Obstacles to European Growth." *NBER Working Paper #26503.*

Cooper, Zack, Stuart V. Craig, Martin Gaynor, and John Van Reenen. 2019. "The Price Ain't Right? Hospital Prices and Health Spending on the Privately Insured." *The Quarterly Journal of Economics*, 51–107.

Cylus, Jonathan D., and Bridget A. Dickensheets. 2007-2008. "Hospital Multifactor Productivity: A Presentation and Analysis of Two Methodologies." *Health Care Financing Review*, 29(2).

Dixit, Avinash K., and Joseph E. Stiglitz. 1977. "Monopolistic competition and optimum product diversity." *The American Economic Review*, 67(3): 297–308.

Donahoe, Gerald F. 2000. "Capital in the national health accounts." *Health Care Financing Administration.*

Feenstra, Robert, Robert Inklaar, and Marcel Timmer. 2015. "The Next Generation of the Penn World Table." *American Economic Review*, 105(10): 3150–3182. Available for download at www.ggdc.net/pwt.

Fehr, Hans, and Maria Feldman. 2024. "Financing universal health care: Premiums or payroll taxes?" *European Economic Review*, 166: 104755.

Fonseca, Raquel, Francois Langot, Pierre-Carl Michaud, and Thepthida Sopraseuth. 2023. "Understanding Cross-country Differences in Health

Status and Expenditures: Health Prices Matter." *Journal of Political Economy*, 131(8): 1949–1993.

Fonseca, Raquel, Titus Galama, Pierre-Carl Michaud, and Arie Kapteyn. 2021. "Accounting for the Rise of Health Spending and Longevity." *Journal of the European Economic Association*, 19(1): 536–579.

Fulton, Brent D. 2017. "Health care market concentration trends in the United States: evidence and policy responses." *Health Affairs*, 36(9): 1530–1538.

Gaynor, Martin, and Robert J. Town. 2012. "Competition in Health Care Markets." *Working Paper No. 12/282, Centre for Market and Public Organisation, Bristol Institute of Public Affairs, University of Bristol.*

Gaynor, Martin, Kate Ho, and Robert J. Town. 2015. "The Industrial Organization of Health-Care Markets." *Journal of Economic Literature*, 53(2): 235–284.

Gomme, Paul, B. Ravikumar, and Peter Rupert. 2011. "The return to capital and the business cycle." *Review of Economic Dynamics*, 14(2): 262–278.

Gray, Bradford H., ed. 1986. *For-Profit Enterprise in Health Care.* National Academies Press, Committee on Implications of For-Profit Enterprise in Health Care; Institute of Medicine; Washington D.C.

Grossman, Gene M, and Ezra Oberfield. 2022. "The elusive explanation for the declining labor share." *Annual Review of Economics*, 14(1): 93–124.

Grossman, Michael. 1972. "On the Concept of Health Capital and the Demand for Health." *Journal of Political Economy*, 80(2).

Hall, Robert, and Charles Jones. 2007. "The Value of Life and the Rise in Health Spending." *The Quarterly Journal of Economics.*

Hansen, Gary. 1993. "The Cyclical and Secular Behaviour of the Labour Input: Comparing Efficiency Units and Hours Worked." *Journal of Applied Econometrics*, 8(1): 71–80.

Harper, Michael J., Bhavani Khandrika, Randal Kinoshita, and Steven Rosenthal. 2010. "Nonmanufacturing industry contributions to multifactor productivity, 1987–2006." *Monthly Labor Review, U.S. Bureau of Labor Statistics.*

Herrendorf, Berthold, Christopher Herrington, and Ákos Valentinyi. 2015. "Sectoral Technology and Structural Transformation." *American Economic Journal: Macroeconomics*, 7(4): 104–133.

Horenstein, Alex R., and Manuel S. Santos. 2019. "Understanding Growth Patterns in US Health Care Expenditures." *Journal of the European Economic Association*, 17(1): 284–326.

Huetsch, Leon, Dirk Krueger, and Alexander Ludwig. 2023. "The Medical Expansion, Life-Expectancy and Endogenous Directed Technical Change." *Working Paper*.

Johnson, Garret, and Austin Frakt. 2020. "Hospital markets in the United States, 2007–2017." Vol. 8, 100445, Elsevier.

Jordà, Òscar, Katharina Knoll, Dmitry Kuvshinov, Moritz Schularick, and Alan M Taylor. 2019. "The Rate of Return on Everything, 1870–2015*." *The Quarterly Journal of Economics*, 134(3): 1225–1298.

Karabarbounis, Loukas, and Brent Neiman. 2014. "The Global Decline of the Labor Share." *The Quarterly Journal of Economics*, 61–103.

Krueger, Dirk, and Alexander Ludwig. 2007. "On the consequences of demographic change for rates of returns to capital, and the distribution of wealth and welfare." *Journal of Monetary Economics*, 54: 49–87.

Kydland, Finn, and Nick Pretnar. 2019. "The Costs and Benefits of Caring: Aggregate Burdens of an Aging Population." *NBER Working Paper #25498*.

Lawver, Daniel. 2010. "Measuring Quality Increases in the Medical Sector."

Maestas, Nicole, Kathleen J. Mullen, and David Powell. 2023. "The Effect of Population Aging on Economic Growth, the Labor Force, and Productivity." *American Economic Journal: Macroeconomics*, 15(2): 306–322.

Ngai, Rachel, and Christopher Pissarides. 2007. "Structural Change in a Multisector Model of Growth." *American Economic Review*, 97(1).

Palangkaraya, Alfons, and Jongsay Yong. 2009. "Population ageing and its implications on aggregate health care demand: empirical evidence from 22 OECD countries." *International Journal of Health Care Finance and Economics*, 9: 391–402.

Prettner, Klaus. 2013. "Population Aging and Endogenous Economic Growth." *Journal of Population Economics*, 26(2).

Romley, John A., Dana P. Goldman, and Neeraj Sood. 2015. "US hospitals experienced substantial productivity growth during 2002–11." *Health Affairs*, 34(3): 511–518.

Shatto, John D., and M. Kent Clemens. 2022. "Projected Medicare Expenditures under an Illustrative Scenario with Alternative Payment Updates to Medicare Providers." Centers for Medicare & Medicaid Services.

Spitalnic, Paul, Stephen Heffler, Bridget A. Dickensheets, and Mollie Knight. 2022. "Hospital Multifactor Productivity: An Updated Presentation of Two Methodologies Using Data through 2019." Centers for Medicare & Medicaid Services.

Tawil, Michael, and Anthony M DiGiorgio. 2022. "Competition in California's Medi-Cal Managed Care Market Assessed by Herfindahl-Hirschman Index." *INQUIRY: The Journal of Health Care Organization, Provision, and Financing*, 59: 00469580221127063.

Triplett, Jack. 2011. *Health System Productivity.*

Triplett, Jack, and Barry Bosworth. 2004. *Productivity in the U.S. Services Sector: New Sources of Economic Growth.* Brookings Institution Press.

Zhao, Kai. 2014. "Social security and the rise in health spending." *Journal of Monetary Economics*, 64: 21–37.

Model Appendix

A1. Aggregation and Market Clearing

Markets for consumption, health services, and investment satisfy:

$$\sum_{j=1}^{J} N_{jt} c_{jt} = C_t \tag{A1}$$

$$\sum_{j=1}^{J} N_{jt} h_{jt} = H_t \tag{A2}$$

$$\sum_{j=1}^{J} N_{jt} \iota_{jt} = I_t \tag{A3}$$

Capital markets must satisfy:

$$\sum_{j=1}^{J} N_{jt}(a_{jt} + b_t) = K_{ct} + K_{ht} = K_t \tag{A4}$$

Total bequests given must satisfy total bequests received:

$$\sum_{j=1}^{J} N_{j,t-1}(1 - s_{j,t-1}) a_{jt} = \sum_{j=1}^{J} N_{jt} b_t \tag{A5}$$

Labor markets must satisfy:

$$\sum_{j=1}^{J} N_{jt}\eta_j = L_{ct} + L_{ht} = L_t \tag{A6}$$

Finally, profits must satisfy:

$$\sum_{j=1}^{J} N_{jt}\pi_{jt} = \Pi_t^c + \Pi_t^h = \Pi_t \tag{A7}$$

A2. Multi-sector, Monopolistically Competitive Overlapping Generations Equilibrium with Transfers

Given deterministic sequences of Social Security and Medicare tax rates $\{\tau_t\}_t$, deterministic sequences of total-factor productivities $\{A_{ct}, A_{ht}\}_t$, deterministic sequences of newborns $\{N_{1t}\}_t$, deterministic sequences of accidental mortality rates $\{m_{jt}^{acc}\}_{j,t}$, known exogenous sequences of health productivity rates $\left\{z_t, \{\zeta_{jt}\}_j\right\}_t$, and sequences of relative markups $\{\mu_t\}_t$, a monopolistically competitive equilibrium with transfers consists of:

i. Sequences of household policies $\{c_{jt}, h_{jt}, a_{j+1,t+1}\}_{j,t}$.

ii. Sequences of producers' policies $\{K_{ct}, L_{ct}, K_{ht}, L_{ht}\}_t$.

iii. Sequences of prices $\{p_t, r_t, w_t\}_t$.

iv. Sequences of population distributions $\{N_{jt}\}_{j,t}$.

v. Sequences of bequests $\{b_t\}_t$.

vi. Sequences of dividends $\{\pi_{jt}\}_{j,t}$.

vii. Sequences of transfers $\{T_t\}_t$.

such that

a. Household policies solve the optimization problems for both workers and retirees.

b. Producers maximize profits.

c. Relative prices satisfy the markup condition.

d. Population is exactly the number of survivors from the previous period plus the number of newborns.

e. Bequests distributed are exactly equal to the leftover assets of the recently deceased.

f. Dividends are distributed in proportion to asset holdings.

g. Transfers equate with total governmental revenues.

h. Markets for consumption, investment, health services, capital, and labor clear.