



US Health Care Expenditures, GDP and Health Policy Reforms: Evidence from End-of-Sample Structural Break Tests

Ben Brewer¹ · Karen Smith Conway² · Deniz Ozabaci² · Robert S. Woodward²

Published online: 12 June 2022
© EEA 2022, corrected publication 2022

Abstract

This research investigates the over-time stability of the aggregate US healthcare expenditure (HCE)–GDP relationship, focusing on periods of healthcare reforms. The most consequential reforms—Medicaid/Medicare and the Affordable Care Act (ACA)—are challenging to study because they occur near the ends of the available data. Using annual national- and state-level data and a battery of structural break tests, we find the HCE–GDP relationship to be overwhelmingly stable. An ancillary analysis around the 2006 Massachusetts healthcare reform, which avoids the confounding effects of the Great Recession and the staggered rollout of the ACA, likewise finds no change.

Keywords US health care expenditure · GDP · Income elasticity of health care · Multiple structural changes

JEL Classification I11 (Analysis of Health care Markets) · C22 (Time-Series Models)

Brewer: University of Hartford. Conway: University of New Hampshire. Ozabaci: Department of Economics. Woodward: University of New Hampshire. We thank Andrew Houtenville, Daniel Henderson, an anonymous referee and seminar participants at the Eastern Economic Association Annual Meetings, Western Economic Association Annual Meetings, and the University of Hartford for their comments and suggestions. The authors would like to thank Mohitosh Kejriwal for making his Gauss programs publicly available. All errors are our own. Stata code utilized for the end-of-sample tests is available upon request.

✉ Ben Brewer
bbrewer@hartford.edu

¹ University of Hartford, West Hartford, CT, USA

² University of New Hampshire, Durham, NH, USA



Introduction

As has been well documented, the proportion of gross domestic product (GDP) devoted to healthcare expenditures (HCE) in the USA has more than tripled since 1960 (CMS 2021a). The growth over that time has been far faster than any other country in the world, and the current level of US HCE also represents the largest share spent on health care by a wide margin (OECD 2020). These trends have spurred research into the factors driving healthcare costs and have helped motivate substantial policy changes at the US federal and state levels designed, at least in part, to control costs and ‘bend the curve’ (Cutler 2010). One example of such efforts is the “1 Percent Steps for Health Care Reform Project” (<https://onepercentsteps.com/>), an organization whose goal is to “offer a roadmap to policy makers of tangible steps we as a country can take to lower the cost of health care in the US....(and) to leverage leading scholars’ work to identify discrete problems in the US health system and offer evidence-based steps for reform.” This group recommends a list of possible reforms that, if adopted, are estimated to reduce health care expenditures by 9%.

Our work complements these efforts by examining whether the two most substantial health care reforms to date—the creation of Medicare and Medicaid and enactment of the Affordable Care Act—appear to have affected health care expenditures as a share of US income. Specifically, we investigate whether there has been a structural break in the US HCE–GDP relationship since 1960 using methods designed to deal with the challenges of exploring changes near the ends of the sample. In so doing, we offer an empirical methodology that may prove useful to evaluating future reforms.

Ever since the seminal work by Newhouse (1977) found income to be the primary driving factor of health care expenditures and an income elasticity suggesting health care is a luxury good, many studies have estimated the health expenditure income elasticity using variation across countries/states and/or time (Parkin et al. 1987; Gerdtam et al. 1992; Newhouse 1992; Hitiris 1997; Baltagi and Moscone 2010; Farag et al. 2012; Hartwig and Sturm 2014; Baltagi, Badi, Raffaele Lagravinese, Francesco Moscone, and Elisa Tosetti. 2017).¹ While these results are mixed on whether aggregate health spending is a luxury or a necessity, they generally have concluded that the growth in income per capita is the major factor behind the surge in health care expenditures for developed countries.² This result especially holds when focusing solely on the more uniform US experience (Freeman 2003; Wang and Rettenmaier 2007; Wang 2009; Moscone and Elisa 2010a; Woodward and Wang 2012).

If health policy reforms help control costs, affect access to care, or otherwise alter health care consumer and provider decision making, then they may have an

¹ For a more thorough history of the literature examining the income elasticity of health expenditures, see Baltagi, Badi, Raffaele Lagravinese, Francesco Moscone, and Elisa Tosetti. (2017).

² While reviewing the history of income elasticity of health expenditure studies using cross-country variation, Baltagi, Badi, Raffaele Lagravinese, Francesco Moscone, and Elisa Tosetti. (2017)’s Appendix Table 5 documents 15 studies concluding the income elasticity is a luxury (income elasticity >1) and 21 concluding it is a necessity (income elasticity <1).



impact on the HCE–GDP relationship in general and the income elasticity specifically. Using US quarterly data from 1985Q1 to 2017Q1, Cheng and Nopphol (2019) investigate the impact of health care policy uncertainty (HCPU) and finds that HCPU shocks leads to temporary decreases in both HCE and GDP. While they do not investigate the effect on the income elasticity specifically, health care policy reforms like Medicare/Medicaid or the ACA seem likely to affect HCPU, particularly if they are subject to lengthy and contentious political debate.

In addition to HCPU, the income elasticity may vary due to other factors that are plausibly affected by health reforms and other policy changes. Barati and Fariditavana (2020) posit that the relationship with income could be asymmetric due to the behavioral tendencies of loss aversion (which leads to stronger reactions to income losses than gains) and stockpiling (which predicts the opposite). Their results using US annual data suggest that income gains have bigger impacts. More generally, it suggests the income elasticity may vary depending on the business cycle and overall health of the economy, a possibility we explore here as well. And, if health policy alters the incentives facing consumers and providers, these relationships could change.

Using a 1971–2009 panel of 14 OECD countries, Blazquez-Fernandez et al. (2014) allows for over-time and across-country heterogeneity in the income elasticity. They find it is larger for higher-income countries and has declined over time (if technological progress is controlled for with a temporal time trend). On the other hand, Baltagi et al. (2017) find the opposite effect when using annual data on 167 countries over 18 years. Lee, Oh and Meng (2019, Table 5) estimate the income elasticities for each of 14 OECD countries and find the elasticity for the USA (at 2.34) is well above the rest. As one possible explanation for this high value, we posit that the US health care system depends more on private insurance and private providers, which may have facilitated marketing more expensive luxury health care products and services. Many of the reforms offered by the 1% steps website suggest that regulating or reforming questionable billing practices, concentrated markets and other anti-competitive features could lower health care expenditures substantially and thus could be explanations for why the US is such an outlier among OECD countries. To the extent that reforms to the US system curb (or enhance) those tendencies, the HCE–GDP relationship may change.

To our knowledge, only Woodward and Wang (2012) use formal structural break tests to investigate whether the log–log relationship between US HCE and GDP has changed over time.³ Using US annual data for 1960–2008 and the Kejriwal and Perron (2010) structural break test (henceforth KP test), Woodward and Wang (2012) show that the relationship has been surprisingly stable, suggesting that US policies

³ This exercise is distinct from research that searches for structural breaks in each data series (HCE, GDP) separately, such as explored in Carrion-i-Silvestre (2005). In investigating the heterogeneity of the income elasticity of health care across a panel of OECD countries, Blazquez-Fernandez et al. (2014) do consider sub-time period analyses and dynamically recursive estimations, which can be thought of as informal tests of structural change. Their recursive estimates do not reveal a large change in the elasticity over time, particularly when a temporal time trend is not included. Woodward and Wang (2012) test for structural changes in the HCE–GDP relationship, including the intercept, as we do here. The test is thus broader than examining if the income elasticity—the slope of the log–log relationship—has changed. We therefore use ‘relationship’ rather than the more specific ‘elasticity.’



have not changed the income elasticity nor ‘bent’ the curve. However, their data end before the Affordable Care Act (ACA), and their structural break methods preclude testing the effects of Medicare/Medicaid because that policy occurred too early in the sample.⁴

As shown by the Kaiser Family Foundations compilations (Kaiser Family Foundation 2011, 2013), the USA has experienced a large number of health policy reforms since 1960, with Medicare/Medicaid and the ACA being the most ambitious.⁵ As such, empirically analyzing whether they have affected the HCE–GDP relationship is important as both were designed to improve access and/or control costs. While the ACA’s first objective was to expand access, stemming the rise in health care expenditures was a second and critical goal. As President Barack Obama noted in an address to the House Democratic Caucus, “Every single good idea to bend the cost curve and start actually reducing health care costs [is] in this bill” (Obama 2010). Subsequent research, however, provides mixed evidence that the ACA has delivered on this goal. Focusing specifically on hospital utilization, Gaffney et al. (2019) find no evidence that either the ACA or Medicare/Medicaid impacted days spent in the hospital or hospital discharges. Chandra et al. (2013) and Weiner et al. (2017) suggest that the Great Recession was more likely to be responsible for the immediate slowdown in costs and the delayed expansion of coverage.

In this study, we focus on whether the HCE–GDP relationship in the USA changed during times of health care policy changes and reforms over the past 60 years. We investigate the possibility of these changes with 1960–2020 national data and 1963–2020 state-level data using structural break tests. As also pointed out by Piehl et al. (2003), structural break tests can be helpful in a policy evaluation setting, especially when the timings of the event and the effect of intervention are not clear. We employ both conventional KP tests and tests that permit end-of-sample (EOS) testing (Andrews and Kim 2006). To our knowledge, ours is the first health economics study to use these EOS structural break tests. These EOS tests are critical for studying the years during which the most substantial reforms took place. The ACA contains several provisions that did not become effective until as late as 2014 (Kaiser Family Foundation 2013) and some, such as the Medicaid expansion, continue to be debated and adopted by the states. The end-of-sample tests we use permit testing for structural breaks as late as 2018 and so should catch breaks associated with the ACA’s later changes. Similarly, these tests can search for breaks at the beginning of the sample, when Medicaid/Medicare was enacted (1965) and implemented (1966).

In addition to analyzing the nation as a whole, we examine each state independently and conduct more extensive analyses for Massachusetts, focusing on the

⁴ Specifically, the authors use the standard 15% trimming method, though other trimming levels are possible. As noted by Andrews (1993), “trimming” means that a regression is estimated over the full sample but the iterative testing for structural breaks is only performed over a predetermined inner portion of the sample. Done to allow an initial regime to form and to ensure there are enough degrees of freedom to conduct a structural break test, the standard practice is to test only the inner 70% of the sample for structural change.

⁵ For instance, Skinner and Chandra (2016) p. 497, state that the ACA is “the most comprehensive health care reform since Medicare.”



period around its 2006 health care reform law. Using two different data sources for Massachusetts health expenditures, we are able to perform both traditional and EOS structural break tests that include 2006. The first source and the one used above in the separate analysis for each state is from the Bureau of Economic Analysis (BEA). It spans 1963 to 2020, thereby allowing 2006 to be subjected to both types of tests. The second is the more commonly used data from the Center for Medicare and Medicaid Services' (CMS) National Health Expenditure Accounts (NHEA) (CMS 2019). Because it only includes 1980 through 2014, we cannot use it to test Medicaid/Medicare or the ACA but it can test the 2006 Massachusetts policy reform. Our Massachusetts analyses provide complementary evidence and robustness checks to our national- and individual state-level findings because (1) it does not rely solely on EOS structural break tests; (2) it uses two different data sources; and (3) the Massachusetts health care reform occurred before the Great Recession and the state was not as hard hit as most.⁶ While the 2006 Massachusetts reform differed in key aspects, like the generosity of low-income subsidies and the level of employer responsibility, "the basic structure of the ACA was pioneered in the Bay State," and it was "the explicit model for the coverage and insurance market reform portions of the ACA" (Gruber 2013). The similar results from the two state-level datasets lend support for using the much longer and continually updated BEA data in future state-level research.

All our findings support the hypothesis that the log-log relationship between aggregate US HCE and GDP has been remarkably stable. Even the periods during the most substantial reforms like Medicare/Medicaid and the ACA yield no evidence of a change in the relationship. Across a range of tests, specifications, and samples, we find no consistent evidence of a structural break and instead find estimated income elasticities that barely budge over our data sample. We also find that all income elasticities well exceed 1.0, a result that is consistent with the hypothesis that health care is a luxury good in the USA. These two findings are also in line with Hall and Jones (2007) whose model based on standard assumptions predicts an income elasticity of health expenditures far greater than one. This finding suggests that the steady rise in health expenditures may in fact be a rational response to rising income in part because the marginal utility of extending life never decreases.

The rest of the paper is organized as follows. The next section describes the data used in the analyses and also provides details of the techniques designed to mitigate the problems encountered in estimating this long-run relationship in general and the income elasticity in particular and credibly testing for structural breaks that occur close to the ends of the sample. Section "Results" reports the results of our estimates and tests, and Section "Concluding Remarks" concludes.

⁶ See, for example, <https://www.thefiscaltimes.com/Articles/2012/06/22/Great-Recession-Hurt-Some-States-More-Than-Others>, and http://archive.boston.com/business/articles/2010/09/29/mass_weathered_recession_better_than_others_data_indicate/.



Data and Empirical Strategy

To investigate the stability of the HCE–GDP relationship and income elasticity over time, we estimate time-series models using annual data on HCE and GDP, first at the national level and then for each state individually, with an in depth focus on Massachusetts. As is common in this literature, we estimate the log of per capita HCE as a function of the log of per capita GDP (Carrion-i-Silvestre 2005; Wang and Rettenmaier 2007; Baltagi and Moscone 2010; Moscone and Tosetti 2010a; Woodward and Wang 2012). We also follow most research in using real values (inflation-adjusted via the Consumer Price Index (CPI) for all goods and services), although in robustness checks we explore using nominal values instead as in Woodward and Wang (2012) and Hartwig (2011). As summarized briefly below, we follow recent research that tests the time-series properties of each data series and adjust our estimation accordingly.

The basic model is

$$h_t = \alpha + \beta y_t + \varepsilon_t \text{ for } t = 1, \dots, T \quad (1)$$

where h_t and y_t are the natural logs of health care expenditures per capita (*HCEPC*) and gross domestic product per capita (*GDPPC*) at year t , β represents the income elasticity of health expenditures and T is the sample size. In a model without structural breaks, α and β are constants over time, but with M structural breaks, there could be an α and β for each separate regime (for a total of $M+1$ regimes).⁷ Our empirical approach is to test for such regime changes and, for those that emerge, see if they coincide with the timing of major health care reforms. To test for changes near the beginning or end of the sample, we use EOS tests. We also explore the sensitivity of income elasticity estimates as more years of data are excluded at the ends of our sample to search for evidence of less abrupt structural changes and to verify that our findings for the overall relationship extend to the income elasticity in particular (i.e., are not driven by the intercept).

Data Description

National-level health expenditure data comes from the CMS's NHEA (CMS 2021b) and is available from 1960 through 2020. Similar to the production-based framework used to measure GDP, the NHEA's national health expenditure series represents the total annual amount spent on final health care consumption in the USA, as well as final spending on administration, public health activities and investment in structures, equipment and non-commercial research in the medical sector.⁸

⁷ For example, in the case of a single structural break ($M = 1$) at time $t = t_i$, there are $M+1=2$ different regimes or $h_t = \begin{cases} \alpha_1 + \beta_1 y_t + \varepsilon_t & \text{for } t = 1, \dots, t_i \\ \alpha_2 + \beta_2 y_t + \varepsilon_t & \text{for } t = t_i + 1, \dots, T \end{cases}$

⁸ For a more detailed account on the NHEA's methodology and how that relates to GDP, see Hartman et al. (2010).



The NHEA data is also the most commonly used in state-level analyses (e.g., Moscone and Tosetti (2010a), Moscone and Tosetti (2010b) and Panopoulou and Pantelidis (2013)). The NHEA state-level data does not begin until 1980 and has not been updated since 2014, which precludes investigating the adoption of either Medicaid/Medicare or the ACA. The NHEA state-level data can be used, however, in our investigation of the 2006 Massachusetts health policy reform, a precursor to the ACA. In our NHEA-MA analyses, we use the State of Provider data which provides estimates of health care spending based on where the provider of care is located. The Massachusetts data differs slightly from the national level in that it does not include spending on administrative, public health or investment endeavors.

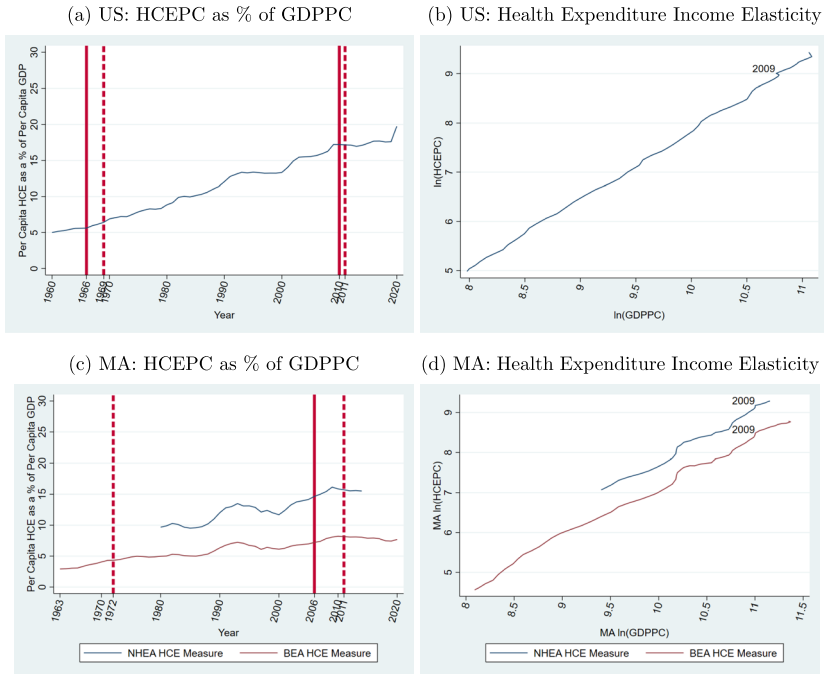
For all other state-level analyses, we use health expenditures from the BEA. This data spans the time period 1963–2020 and therefore allows us to test almost the same period as we do at the national level. Because the BEA changed its industry classifications in 1997, we create this aggregate measure using spending on “health services” from 1963 to 1997 and then sum spending on health and personal care stores, ambulatory health care services, hospitals and nursing/residential care services together from 1998–2020.⁹ We perform the 2006 Massachusetts analyses using both data sources, and the similar findings lends support to our use of the BEA data for the other states and reforms. The other measures required for equation 1 come from the BEA (national- and state-level GDP) and the Census Bureau (population).

We begin our empirical investigation with a descriptive look at the relationship between HCE and GDP over our sample periods. Figure 1a plots the US HCE as a percentage of GDP and confirms that HCE has steadily grown as a percentage of GDP during this time period. This trend is consistent with an income elasticity that is greater than 1.0, the key parameter of interest in equation (1). The uptick at the end of Fig. 1a also highlights the impact of COVID-19 in 2020, which led to both an increase in health expenditures and a stark fall in GDP. Though we cannot test 2020 using our formal structural break tests, we do find that the estimated income elasticity is very robust to excluding 2020. The figure also denotes with dashed lines the 15% at both ends of the sample trimmed for the standard structural break tests and shows how this trimming excludes from the search most of the years likely affected by the most substantial US health care reforms.

Figure 1b provides a direct look at the relationship estimated in equation 1, the relationship between log HCE per capita and log GDP per capita. The slope of the plotted line ($\frac{\partial \ln[HCE]}{\partial \ln[Y]}$), which is the HCE income elasticity, is strikingly constant over time and greater than 1.0. A closer look at both figures does reveal a temporary

⁹ The BEA changed its industry classifications in 1997 from the Standard Industrial Classification (SIC) to the North American Industry Classification System (NAICS). NAICS accounting uses some different source data and estimation methodologies to achieve their final measurements, though they are not drastically different than the SIC numbers. A careful inspection of the growth of GDP or HCE suggests combining the two classifications to achieve a continuous measure is not introducing a break in the relationship and none of our analyses at the national or state level find a break around this time.





Notes: United States estimation sample is annual from 1960-2020. Massachusetts estimation sample is from 1980-2014 when using the National Health Expenditure Account’s measure and 1963-2020 when using the Bureau of Economic Analysis’s measure. Health expenditures (HCE) and GDP are both in per capita terms in (a) and (c) and are in logged per capita terms in (b) and (d). Solid vertical lines represent major health care reforms at the national level (Medicare/Medicaid in 1966 and the Affordable Care Act in 2010) and for Massachusetts (Massachusetts health reform in 2006). Years in between the dashed lines represent testable years once the standard 15% trimming is applied for the Kejriwal & Perron (2010) tests while years outside the dashed lines are untestable under this trimming percentage. For the Massachusetts data in Figure 1(c), these dashed lines apply only to the BEA data.

Fig. 1 Plots of health care expenditures and GDP per capita (HCEPC and GDPPC) over time

leveling off in 2009 in the share of GDP made up by HCE in Fig. 1a and a corresponding blip around the same time in the income elasticity in Fig. 1b, suggesting the possibility of a breakpoint at the start of the ACA or the Great Recession. Otherwise, this simple descriptive look at the data thus provides evidence that the relationship has remained stable and suggests that income elasticity is greater than 1.0.

Figure 1c and 1d repeats these plots for the two Massachusetts samples. In contrast to the US sample, Fig. 1c shows how the period spanning the 2006 reform remains in the BEA data even after trimming, which allows us to perform both standard and EOS structural break tests.¹⁰ This figure also shows that the two data

¹⁰ While we can perform both sets of structural break tests on the BEA data, we are only able to perform the end-of-sample tests for the NHEA sample due to its relatively short length.



sources for Massachusetts track each other reasonably well and thus lend support to our use of the BEA data; their correlation coefficient is 0.99. While both figures suggest less stability in the relationship than for the USA as a whole, they do not reveal an obvious difference before and after the 2006 Massachusetts reform and the slope of both series in Fig. 1d is also fairly constant. These figures thus provide the first evidence that the HCE–GDP relationship is quite stable over time and echo the findings of Woodward and Wang (2012), who find a similar stability using nominal data and a sample that ends in 2008. However, these figures provide us with only descriptive information, and as such, we move on to our formal statistical analyses.

Time Series Properties

Before formally investigating the stability of the log–log health expenditure–GDP relationship elasticity, we also determine the time-series properties of our different variables so that we can properly model the relationship between them over time.¹¹ Table 1 summarizes the key findings from these exercises and the different structural break tests we are able to perform on our national-level data and two Massachusetts datasets while Appendix Figure A1 illustrates those time series. For the sake of brevity, we do not report this information for the other states and DC, but they are in general similar to what we find for the national dataset and are available upon request.

To assess the order of integration for our series, we use both the standard augmented Dickey–Fuller (ADF) test and the modified GLS de-trended test. We adopt the generalized least squares (GLS) de-trended test due to the cited power issues of the standard Dickey–fuller test in the case of negative serial correlation (Perron and Ng 1996; Elliot et al. 1996; Ng and Perron 2001).

It is now well recognized that a structural break in a series could nullify the validity of the ADF and DF-GLS tests. In particular, these tests will tend to fail to reject the null hypothesis of a unit root when a break is present because the test is not able to reliably distinguish non-stationarity from a shift in the series (Perron, 1990). To account for this possibility, we implement the test proposed by Clemente et al. (1998) (CMR). We prefer the CMR approach since it allows for up to two breaks, i.e., to ensure that the stationarity tests of each series are robust if there are multiple breaks.

Once the stationarity properties of the health expenditure and income series are established, it is important to then determine whether or not the two series are cointegrated. To test for cointegration, we use the Engle–Granger (EG) two-stage procedure.¹²

¹¹ In all stability tests, we test both the intercept and the slope of the log–log HCE–GDP relationship for structural breaks. For simplicity, we will refer to the health expenditure income elasticity when discussing the stability of the relationship between HCE and GDP.

¹² As robustness, we also use the Johansen trace test. The trace test allows for the testing of multiple cointegrating relationships if more than two variables are being tested. However, since we can have at most one cointegrating relationship between health spending and income and the trace test is based only on asymptotic properties, we prefer the EG test in this context.



Table 1 Time-series properties of the USA and Massachusetts $\ln(\text{NHEPC}) (h_t)$ and $\ln(\text{GDPPC}) (y_t)$

	USA (1960-2020)				Massachusetts NHEA Data (1980-2014)				Massachusetts BEA Data (1963-2020)			
	(1) h_t	(2) y_t	(3) Δh_t	(4) Δy_t	(5) h_t	(6) y_t	(7) Δh_t	(8) Δy_t	(9) h_t	(10) y_t	(11) Δh_t	(12) Δy_t
Augmented	-2.313	-2.878	-3.463***	-4.381***	-1.419	-3.929**	-2.172	-2.679*	-2.030	-2.486	-3.564***	-4.669***
Dickey-Fuller	(-3.492)	(-3.492)	(-2.924)	(-2.924)	(-3.572)	(-3.572)	(-2.980)	(-2.980)	(-3.497)	(-3.497)	(-2.928)	(-2.928)
Test												
GLS	-1.538	-0.960	-3.937***	-5.259***	-1.177	-2.220	-3.803***	-3.481**	-1.056	-2.848	-3.803***	-4.252***
de-trended												
Dickey-Fuller	(-2.987)	(-3.129)	(-3.171)	(-3.171)	(-3.286)	(-3.386)	(-2.811)	(-3.195)	(-3.045)	(-2.873)	(-2.811)	(-3.148)
Test												
Clemente	-2.332	-2.682	-4.594**	-6.234**	-2.730	-2.601	-3.954**	-4.423**	-2.587	-2.821	-4.708**	-5.070**
Montanes												
Reyes AO	(-3.560)	(-3.560)	(-3.560)	(-3.560)	(-3.560)	(-3.560)	(-3.560)	(-3.560)	(-3.560)	(-3.560)	(-3.560)	(-3.560)
Test												
Cointegration tests ^b												
USA												
Engle-Granger Test	-3.742**		26.339***				26.207***				33.78***	
	(-3.440)		(15.410)				(15.410)				(15.410)	
Massachusetts BEA Data												
Engle-Granger Test												
Massachusetts NHEA Data												
Engle-Granger Test												
Model Information												
USA												
# of DOLS	2				2				4			
lags needed												
# of DOLS	2				0				4			
leads												
needed												
Years Tested	1969-2011				NA ^c				1972-2011			
in KP Tests												



Table 1 (continued)

Unit root tests ^a		USA (1960-2020)			Massachusetts NHEA Data (1980-2014)			Massachusetts BEA Data (1963-2020)																
(1)	h_t	(2)	y_t	(3)	Δh_t	(4)	Δy_t	(5)	h_t	(6)	y_t	(7)	Δh_t	(8)	Δy_t	(9)	h_t	(10)	y_t	(11)	Δh_t	(12)	Δy_t	
Years Tested in EOS	Beginning 1963-1972	End 2009-2018	Beginning 1983-1992	End 2005-2014	Beginning 1968-1978	End 2006-2016	Beginning 1968-1978	End 2006-2016	Beginning 1968-1978	End 2006-2016	Beginning 1968-1978	End 2006-2016	Beginning 1968-1978	End 2006-2016	Beginning 1968-1978	End 2006-2016	Beginning 1968-1978	End 2006-2016	Beginning 1968-1978	End 2006-2016	Beginning 1968-1978	End 2006-2016	Beginning 1968-1978	End 2006-2016

^a5% critical values are in parentheses. All unit root tests include both a time trend and a constant, and all cointegration tests include a constant. h_t refers to the natural log of per capita real health care expenditures, while y_t refers to the natural log of per capita real GDP

^bNull hypothesis is that there is a unit root. Optimal lag structure is selected using the AIC or through the method proposed by Ng and Perron (2001)

^cNull hypothesis is that there is no cointegration. Optimal lag structure is selected using the AIC

^dNA = not available; with trimming and the required lags, the data series is too short to perform the KP tests. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$



The top two panels of Table 1 report the results of the unit root and cointegration tests. The battery of tests reported in the top panel overall suggests we cannot reject the null hypothesis of a unit root in the levels of either y_t or h_t for all three series. In unreported results, we find the same result for the other states and DC. This suggests that each series exhibits sustained persistence over time and that this serial correlation may bias our results if not properly treated.¹³ The second panel presents the results for both the preferred EG test and the Johansen test as confirmation. Both tests strongly confirm that the variables in all three series reported in Table 1 are cointegrated, as are the other states and DC. These time-series property conclusions are consistent with the literature (Blomqvist and Carter 1997; Gerdtham and Löthgren 2000; Freeman 2003; Carrion-i-Silvestre 2005; Wang and Rettenmaier 2007) and suggest a modification of our primary model is necessary before testing for stability.

Due to these time-series properties, standard ordinary least squares (OLS) regression estimation between h_t and y_t may be biased due to the correlation between the right-hand side variable and the cointegration error. To overcome this, Stock and Watson (1993) propose the dynamic OLS (DOLS) estimator, which they show to be asymptotically efficient, by including leads and lags of the first difference of the integrated right-hand side variable in the regression. Adding these variables not only helps mitigate issues with autocorrelation, but also accounts for some of the simultaneity bias, which may occur in a regression of cointegrated variables.¹⁴ However, using a Monte Carlo study, Hayakawa and Kurozumi (2008) investigate the finite sample properties of DOLS regressions and find that models without the first differenced leads are optimal if the cointegration errors do not Granger-cause the first difference of the integrated right-hand side variable. In other words, if past values of the cointegrating errors do not contain any information that could help predict the first differenced integrated regressor (y_t), then the simultaneity bias does not need to be addressed and the exclusion of the leads increases the degrees of freedom and the efficiency of the estimation, particularly in small samples.

We use the data driven procedures outlined by Hayakawa and Kurozumi (2008) to determine the optimal number of leads and lags to augment Equation 1 before proceeding with the structural break tests.¹⁵ We find differences across the three samples. While the national-level data suggests two leads and two lags, the MA NHEA sample implies no leads and the MA BEA sample requires four leads and

¹³ In the Massachusetts NHEA data, we are able to reject the null hypothesis of a unit root for the first difference of the y_t , but not initially for the first difference of h_t . However, when using the preferred Clemente et al. (1998) Additive Outlier test, we can reject the null hypothesis for h_t , suggesting the series is integrated of order one as well, but with a structural break in its level form. This finding is in line with Carrion-i-Silvestre (2005).

¹⁴ See Stock and Watson (1993) for a more detailed explanation of the DOLS estimation framework.

¹⁵ Following Hayakawa and Kurozumi (2008), we test for Granger causality between the residuals from a simple regression of y_t on h_t and the first difference of y_t . As is standard, we first use the Akaike information criteria (AIC) and Schwartz–Bayesian Information Criteria (SBIC) to determine the optimal number of leads and lags, and then we test whether or not the leads are necessary using the Granger causality procedure suggested by Hayakawa and Kurozumi (2008).



lags. The other states and DC mostly follow the national data with two leads and two lags (28 states) or one lead and lag (16 states). The remaining six states require at least three or four leads and lags and their higher number of lags preclude them from testing 1966. We explain this selection process in more detail further in the Appendix and discuss how the results differ across our three main time series in Appendix Tables 5, 6, 7.

Moving forward, we augment Equation 1 with the number of leads and lags as defined above for each of our series to reach our baseline DOLS models, and test for structural breaks using these augmented equations. The actual number of potential breakpoints, and where they occur, are not specified *a priori*, and both the intercept and the slope are allowed to vary in our tests.

Testing for Structural Change

To test the stability of the aggregate log–log HCE–GDP relationship over time, we first utilize the sup-Wald, the UD_{max} and the sequential multiple break testing procedures proposed by Kejriwal and Perron (2010).¹⁶ These procedures fit the scope of our question because they (1) allow for breaks in both the intercept and slope of the DOLS equation, (2) allow for multiple breaks over the sample, (3) do not require the break dates to be specified *a priori* and (4) are consistent even under the presence of non-stationary and cointegrated variables such as both we and the literature more generally has found for h_t and y_t (Bai and Perron 1998; Bai and Perron 2003; Perron 2006; Kejriwal and Perron 2010). It is also known that sequential procedures have a tendency to stop too early, but we alleviate the issue since we combine the UD_{max} test and the sequential procedure following Bai and Perron (1998), Perron (2006) and Kejriwal (2008).

While these testing procedures fit our question well, two issues may bias these tests toward a failure to find structural breaks. The first is that given our relatively small sample (1960–2020 or $T = 61$ for the USA, 1980–2014 or $T = 35$ for the NHEA Massachusetts sample and 1963–2020 or $T = 58$ for the BEA Massachusetts and the rest of the state samples), the size and power of these procedures may be limited. Second, in each test, the related DOLS equation is subject to “trimming” i.e., the removal of a certain portion of the beginning and the end of the sample, to determine the range over which the breaks will be searched. The trimming ensures that each testing segment does not get too small, which is necessary to ensure adequate power, especially when there is serial correlation in the data (Andrews 1993, 2003; Bai and Perron 2003; Perron 2006). While trimming helps increase power, it excludes years at both ends of the sample and, as such, eliminates two periods when large scale health reforms took place in our sample. For example (and as reported at the bottom of Table 1), the standard 15% trimming that we adopt means the years actually being tested for structural breaks in the US national data are 1969–2011,

¹⁶ The formulas for each of these three test statistics are located in the Appendix. For more discussion on any of these test statistics, or the dynamic algorithm used to estimate them, see Kejriwal and Perron (2010) or Casini and Perron (2019).



meaning the advent of Medicaid/Medicare in 1966 and the majority of the ACA rollout, which occurred well after its enactment in 2010, are not actually being investigated as potential points of structural change. The bottom of Table 1 reports the years of each sample that can be subjected to the KP structural break tests for each sample, followed by the years investigated with the EOS tests described next.

To address both of these concerns, we adopt the EOS structural change tests proposed by Andrews (2003) and extended by Andrews and Kim (2006). The P test developed in Andrews and Kim (2006) fits our research question because it addresses changes in short time periods, such as in the beginning or the end of the sample, unlike most structural change tests which are designed to identify breaks over a long span of data. Following Andrews and Kim (2006), we also know that the P test is appropriate for models with cointegrated variables, as h_t and y_t are here. These tests therefore allow us to check whether breaks occurred prior to 1969 (Medicare/Medicaid) or post 2011 (ACA). One criticism of this procedure is that it typically requires pre-specifying the break points. To mitigate this issue, we alternatively pre-specify the full range of possible break points, allowing us to test each year in both tails of the sample.

To carry out the P test, one first pre-selects a hypothesized break date at time $t = t_i$. The DOLS model is estimated over the entire sample, and the sum of squared residuals is calculated only for the post-change period ($t_i + 1, \dots, T$). This sum represents the P statistic associated with time t_i . Because the number of observations at the beginning or end of the sample is small, one cannot rely on the standard asymptotic critical distribution to generate critical values. The P test instead relies on an iterative, sub-sampling procedure to derive the empirical distribution of test statistics. Specifically, we follow Andrews and Kim (2006) by estimating $T - 2(T - t_i + 1) + 1$ hypothetical test statistics using a moving window of the pre-change period and reach their empirical distribution. If the P statistic is greater than the value at the “1-significance level” percentage of this empirical distribution, the null hypothesis of no structural change can be rejected in favor of a structural change at the year corresponding to time t_i at that significance.¹⁷

Results

We start by presenting formal structural break results for the national-level HCE–GDP relationship and income elasticity using the standard KP tests and a trimmed sample that spans 1969–2011. To explore whether the relationship is stable at the ends of the sample, when Medicare/Medicaid and the ACA are enacted, we use the EOS P test. We then subject these findings to three different robustness checks. The first estimates several variations of the log–log national-level HCE–GDP relationship and repeats the P tests. The second repeats this estimation and two types of tests for each of the other states and DC individually. Estimating each state independently helps avoid possible aggregation biases present in national

¹⁷ Stata code for the end-of-sample tests are available upon request.



data while the national scope of these reforms suggests that state-level relationships could also be impacted—with the possible exceptions of Massachusetts and Hawaii which enacted their own reforms. The final exercise focuses on Massachusetts and its 2006 reform, using two different sources of data and both types of tests. These complementary analyses permit testing for structural breaks during the reform period using both EOS P and KP tests, as well as abstracting from the effects of the Great Recession. Finding similar results for the national-, state-level and two focused Massachusetts analyses helps corroborate our key findings.

Structural Breaks in the US National log–log HCE–GDP Relationship

Table 2 presents the results for both sets of structural break tests applied to our main US HCE–GDP specification written in equation 1, estimated in line with the time-series properties described in Table 1. The KP tests with appropriately trimmed data tests the 1969–2011 period, while the EOS P-test looks for breaks between the years of 1963–1972 (Medicaid/Medicare era) and 2009–2018 (ACA era). As the constant slope in Figure 1b suggests, the different types of tests are unanimous in finding no evidence of structural breaks over the sample, including at the ends. For the inner 70% of the sample (1969–2011) examined with the KP tests, we fail to reject the null hypothesis of stability for all *SupF** tests, as well as the *UDmax* test; all tests fall far short of the 10% significance level. As a final level of confirmation, the sequential test finds no evidence of breaks either.

The last four columns of Table 2 report the results for the Andrews and Kim (2006) P test, which searches for breaks at the two critical ends of our sample. To conduct this test, we first must pre-specify a break year and then test the null hypothesis of no change at that year against the alternative hypothesis of a change at that year. The fact that we have to pre-specify allows us to conduct a falsification test in that we manually test all of the years in the beginning and the end of the sample, regardless of whether we expect them to be break years or not. If the EOS test finds many or all years to be significant, this casts doubt on the ability of the EOS test to reliably identify breaks. On the other hand, if only years reasonably expected to be breakpoints are selected, then we can be more confident that these are points of change in the curve. Similarly, having no years selected as breakpoints suggests that the relationship is in fact stable, even at the ends of the sample. The staggered roll-out of the different provisions of the ACA suggests a need to test over many years as well.

Starting with the beginning of our sample 1963–1972, which spans the enactment of Medicare/Medicaid, we find no evidence of a structural break in any year. For example, in 1966, the *p*-value is 0.50, meaning that this year was found less likely to be a point of change than 50% of the hypothetical break years constructed of different periods in the post-1966 sample. As such, we find no evidence that the US HCE–GDP relationship changed in the aftermath of Medicare/Medicaid’s enactment. Performing a similar set of tests for the end of the sample 2009–2018, which spans the Great Recession and the ACA, likewise yields no evidence of a structural break. In fact, each year tested is associated with high *p*-values (ranging from 0.34



Table 2 Kejriwal and Perron (2010) and end-of-sample structural break tests for the US log–log HCE–GDP relationship

Kejriwal and Perron (2010) Test ^a		End-of-sample tests ^b			
(Tested over 1969–2011)		Beginning of sample		End of Sample	
KP Tests	Test statistics	(Tested over 1963–1972)		(Tested over 2009–2018)	
	(10% Critical Value)	Year	<i>p</i> -value	Year	<i>p</i> -value
<i>SupF</i> *(1)	7.383 (12.11)	1963	0.50	2009	0.81
<i>SupF</i> *(2)	7.178 (9.96)	1964	0.34	2010	0.76
<i>SupF</i> *(3)	5.333 (8.60)	1965	0.46	2011	0.80
<i>SupF</i> *(4)	3.154 (7.36)	1966	0.50	2012	0.64
<i>SupF</i> *(5)	3.365 (5.90)	1967	0.55	2013	0.63
UDMax	7.383 (12.25)	1968	0.55	2014	0.52
		1969	0.53	2015	0.49
		1970	0.40	2016	0.34
		1971	0.33	2017	0.37
Sequential Test (# of breaks)	0	1972	0.26	2018	0.43

^aFollowing Kejriwal (2008) and Woodward and Wang (2012), a dynamic OLS model of the form $h_t = \alpha_{0j} + \beta_{1j}y_t + \sum_{i=1}^2 \beta_{i+1} \Delta y_{t+i} + \sum_{i=0}^2 \beta_{i+3} \Delta y_{t-i} + \epsilon_t$ (for $j = 1, \dots, M + 1$) is estimated in each case. Sequential tests using 2 leads and lags find no evidence of structural breaks. The null hypothesis is that there are no structural breaks. 10% critical values are in parentheses and come from Kejriwal and Perron (2010), Table 1A, nontrending case where $q_b=1$

^b*p*-values, calculated through the sub-sampling routine proposed by Andrews and Kim (2006), are reported. In each case, a dynamic OLS model of the form $h_t = \alpha_{0j} + \beta_{1j}y_t + \sum_{i=1}^2 \beta_{i+1} \Delta y_{t+i} + \sum_{i=0}^2 \beta_{i+3} \Delta y_{t-i} + \epsilon_t$ (for $j = 1, \dots, (M + 1)$) is tested. Each year is tested separately, and the null hypothesis is that there is no structural break. We test as far as we can to the beginning and end of the sample given the dynamic OLS lag/lead structure

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

in 2016 to 0.81 in 2009), suggesting none of these years is close to being statistically significantly related to a change in the health expenditure income elasticity or the log–log HCE–GDP relationship overall.

Additional Exercises with National Data

We first subject these findings to alternative specifications of equation 1. The first uses nominal instead of real values of HCE and GDP, as in Woodward and Wang (2012). Consistent with Table 2 and Woodward and Wang (2012), we find no



evidence of a structural break within the internal 70% of the samples.¹⁸ As shown in Table 3, the P-tests once again provide no evidence of a structural break during either the beginning or end of the sample and the p -values are similar to Table 2.

While past research varies widely in the factors included (e.g., see Hartwig and Sturm (2014)), including some measure of the age distribution or dependency ratio is common [e.g., Wang, 2009; Barati and Fariditavana, 2020]. We therefore redo the analyses including the percentage of the population aged between 18 and 64 and the percentage aged 65 and older as additional controls. Table 3 reports this exercise in the third column of each panel. While the p -values are lower, the results continue to show no evidence of a structural break. Next, we follow Blazquez-Fernandez et al. (2014), which explicitly investigates whether the elasticity has changed over time and finds their results sensitive to including a proxy for technological change (a temporal time trend). This addition, reported in the fourth columns, likewise has no qualitative impact.

To deal with both of those factors together along with the potential influence of other unobserved variables as well, we consider whether there are breaks in the log–log HCE–GDP relationship relative to the rest of the OECD countries in the next column. Data on HCE and GDP for the OECD countries is only available from 1970 to 2020, meaning the impact of Medicare/Medicaid at the beginning of our sample cannot be tested in this exercise.¹⁹ We calculate the median health expenditures and GDP by year across the OECD countries with data available over the full 1970–2020 time period and subtract that from the US values, giving US health expenditures and GDP net of the “global normal.”²⁰ Though not as stark as the USA, all other developed countries have also seen a rise in their health expenditures over time, suggesting the possibility that shared factors could at least partly be responsible for this growth across all developed nations. These common factors could therefore also be influencing the stability of the HCE–GDP relationship that has been documented here so far and subtracting out the yearly OECD median should help control for the potential effect those influences have. As before, the HCE–GDP relationship remains stable throughout this exercise with p -values ranging from 0.41 to 0.83 in the end of the sample P test. In unreported analyses, we also subject this net of “global normal” measure to the KP test, finding no evidence of a break in the interior of the sample either.²¹

¹⁸ In Appendix Table 8, we confirm Woodward and Wang (2012)’s key result by performing KP tests on both their trimmed sample period (1968–2001) and our more extended one (1969–2011).

¹⁹ Data for the other OECD countries on health expenditures, GDP and population comes from the OECD’s Databank (<https://data.oecd.org/>). Health expenditures and GDP are converted to real US dollar terms, matching the data from our other two sources.

²⁰ Only data for Austria, Australia, Belgium, Canada, Denmark, Finland, Germany, Iceland, Ireland, Japan, Netherlands, New Zealand, Norway, Portugal, Spain, Sweden, Switzerland and the UK is available over the full 1970–2020 period, and so, the yearly medians are based only off of those countries.

²¹ The results are robust to using the yearly OECD mean as opposed to the median, and the conclusions are similarly unchanged when using a less restrictive version of the exercise where the yearly OECD median (or mean) HCE/GDP ratio is included as a control variable instead of differencing the yearly median/mean out of the US measures. All results available upon request.



Table 3 Estimated p-values from end-of-sample structural break tests for the US log-log HCE-GDP relationship: alternative specifications

Year	Beginning of Sample					End of Sample					
	Nominal variables	+ Age Controls	+ Techno-logical change control	Net of OECD Median ^a	Income Asymmetry	Year	Nominal variables	+ Age controls	+ Techno-logical change control	Net of OECD Median ^a	Income asymmetry
1963	0.54	0.31	0.87		0.53	2009	0.86	0.53	0.36	0.85	0.81
1964	0.46	0.22	0.66		0.40	2010	0.82	0.50	0.32	0.86	0.79
1965	0.48	0.21	0.81		0.52	2011	0.88	0.58	0.28	0.87	0.83
1966	0.22	0.20	0.72		0.52	2012	0.86	0.52	0.19	0.78	0.69
1967	0.21	0.27	0.68	NA	0.59	2013	0.93	0.52	0.18	0.76	0.68
1968	0.21	0.31	0.67		0.60	2014	0.89	0.44	0.17	0.68	0.52
1969	0.28	0.40	0.63		0.60	2015	0.92	0.48	0.17	0.66	0.52
1970	0.16	0.45	0.55		0.53	2016	0.84	0.34	0.14	0.48	0.40
1971	0.19	0.44	0.42		0.36	2017	0.87	0.39	0.15	0.41	0.44
1972	0.18	0.41	0.41		0.27	2018	0.94	0.56	0.22	0.43	0.45

p-values, calculated through the sub-sampling routine proposed by Andrews and Kim (2006), are reported. In each case, a dynamic OLS model of the form $h_t = \alpha_{0j} + \beta_{1j}y_t + \sum_{i=1}^2 \beta_{i+1} \Delta y_{t+i} + \sum_{i=0}^2 \beta_{i+3} \Delta y_{t-i} + \epsilon_t$ (for $j = 1, \dots, (M + 1)$) is estimated. Each year is tested separately and the null hypothesis is that there is no structural break. We test as far as we can to the beginning and end of the sample given the dynamic OLS lag/lead structure. Age controls include the share of the population between 18 and 64 and the share of the population over age 65. Similar to Blazquez-Fernandez et al. (2014), technological change is controlled for using a temporal time trend

^aData comes from the OECD's Databank and spans 1970–2020, preventing the beginning of sample tests from capturing the start of Medicare/Medicaid. Both HCE and GDP capture the US experience net of the yearly median across OECD countries where data was available

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$



Finally, we allow the effect of GDP on HCE to be asymmetric in the last column in each panel as in Barati and Fariditavana (2020). Across all of these different model specifications, we find no evidence that the log–log HCE–GDP relationship changed during the rollout of Medicare/Medicaid in the early part of our sample or during the Great Recession and the ACA in the latter part. The stability over the 1960–2020 period remains unchanged. As a final check, we also re-estimate these national-level models excluding 2020, as the COVID-19 pandemic may have had an undue effect on both HCE and GDP. The results are nearly identical.

The last set of exercises compare the estimated income elasticities obtained from different time periods and in past research. Whether nominal or real values are used, our estimated income elasticities barely budge when the sample is extended and are similar to past work that uses shorter series. For the main specification that uses real HCE–GDP values, we find a large income elasticity of 2.59 for the full 1960–2020 sample, similar to what Lee, Oh and Meng (2019, Table 5) find for 1960–1997 (2.34) and in line with the predictions made in Hall and Jones (2007). When we eliminate the post-ACA period (drop 2011 and years after), the estimated income elasticity changes by only 0.03 to 2.62. Woodward and Wang (2012), using nominal data instead, report an estimated income elasticity of 1.388 for their 1960–2008 sample, which excludes the entire post-ACA period. Our sample using updated data adds 12 years and increases the sample by more than 20%, yet the corresponding estimate is strikingly similar at 1.40.²² In another exercise similar to the recursive approach of Blazquez-Fernandez et al. (2014), we systematically add and subtract the years included in the sample and re-estimate the model.²³ Across both nominal and real national data, this exercise yields elasticity estimates that barely budge, differing by at most 0.042. The bigger difference in elasticities comes with whether one estimates the relationship in real or nominal terms. In sum, neither the structural break tests nor the estimated elasticities themselves provide evidence that the national HCE–GDP relationship has changed during 1960–2020.

Estimates and Structural Break Tests Using State-Level Data

Our next set of robustness checks take advantage of the BEA state-level data spanning nearly the same time period (1963–2020) to repeat this estimation (including the diagnostic tests outlined in Table 1) and the two types of structural break tests for each individual state. This exercise may help alleviate possible aggregation bias; given the national scope of these reforms, it also provides additional, credible evidence as to whether the HCE–GDP relationship was affected at a disaggregated

²² Using our (updated) data and limiting the sample to their years (1960–2008) yields an even closer elasticity estimate of 1.399, which suggests that most of the small difference in estimates is due to updated data, not adding another 12 years.

²³ Specifically, we re-estimate the models systematically delaying the beginning of the sample by one year (e.g., 1961–2020, 1962–2020, etc. through 1973–2020) and curtail the end of the sample by one year (e.g., 1960–2019, 1960–2018, etc. through 1960–2007). We then calculate the difference between the maximum and minimum estimates. We take a similar approach with the Massachusetts samples.



level. However, it also yields an enormous number of tests, summarized here and available in greater detail upon request, and its results are complicated by the fact that the prescribed leads and lags—and thus the possible years subjected to EOS tests—differ across states. Because the data does not begin until 1963 and most states require two or more lags, testing the period just before and during the Medicaid/Medicare enactment is not feasible for most states. We therefore limit our structural break analyses to the 70% interior of the sample (1972–2011) that can use traditional KP tests and apply the EOS tests to the period just before, during and after the ACA rollout. As the highest number of leads found is four, all states can test at least through 2016 and the overwhelming majority (44) can test through 2018.

Performing the KP tests for the 1972–2011 period for each of the other states and DC yields only seven states with possible breaks with no real pattern as the breaks occur at different times and none are even close to 2011.²⁴ Turning to the EOS tests and the end of the sample, we focus on the four years prior to the ACA (2007–2010), the four years during the rollout (2011–2014) and the four years (when possible) afterward (2015–2018).²⁵ Figure 2 summarizes the results of the EOS structural test results for 2007–2018 in all 50 states and DC. We once again find seven states with possible breaks and the timing and states appear fairly random. Two of the breaks occur in the pre-period, three in the rollout period and two in the post-period. Only one break occurs in the critical years of 2011 and 2014. Two of the states (AZ and DE) had suggested breaks in the internal period too, while the remaining five display no obvious pattern (AK, MD, MI, NH and SC). Both types of structural break tests therefore suggest a small number of breaks but with no systematic pattern linking them to policy reforms. Finding more structural breaks is perhaps to be expected in what seems to be less stable series in general; individual states are buffeted with forces that may have little effect in the aggregate. As an additional robustness check, we also use the structural break test proposed by Ditzen et al. (2021) that extend the Bai and Perron (1998) and Bai and Perron (2003) method to the panel level on our full set of states together, finding no evidence of a structural break in the interior 70% of the sample.²⁶

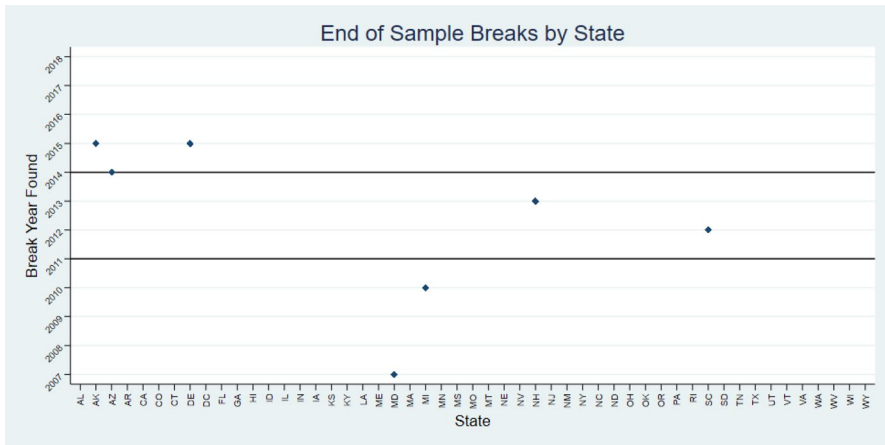
The estimated income elasticities are also reasonably stable across the states and, as in the national data, over time. The top panel of Figure 3 reports the estimated income elasticities and their 95% confidence intervals for each state; the two horizontal lines show the 95% confidence interval income elasticities estimates with the national BEA data. With the exception of Alaska, the elasticities fall in a reasonably close range of 1.51 and 3.42 and most of the confidence intervals overlap. Even Alaska's outlier estimate of 0.92 is imprecise such that its confidence interval overlaps with most others. These estimates are also in the same general range as we find

²⁴ Instead, most of the breaks occur in the 1976–1977 period, with the others being 1972, 1986 and 2003. The states with breaks also seem random and are AZ, DE, GA, IL, NV, PA and WI. All unreported results from our state-level analyses are available upon request.

²⁵ Given the ACA was not enacted until August 2010 and many of the policies did not come into play until 2011 at the earliest, we consider 2010 part of the before period.

²⁶ To our knowledge, this test statistic is not shown to be consistent under the presence of non-stationary and cointegrated variables, unlike the KP test.





Notes: Data for each state comes from the BEA. Break years are calculated through the sub-sampling routine proposed by Andrews and Kim (2006). In each case, a dynamic OLS model of the form $h_t = \alpha_{0,j} + \beta_{1,j}y_t + \sum_{i=1}^n \beta_{i+1}\Delta y_{t+i} + \sum_{i=0}^n \beta_{i+3}\Delta y_{t-i} + \epsilon_t$ (for $j = 1, \dots, (M + 1)$) is estimated. We test over the years 2007 to 2018 and each state-year is tested separately. The null hypothesis is that there is no structural break.

Fig. 2 End-of-sample breaks across the States

in the national NHEA data (2.62), despite using an entirely different data source. Also similar to the national analyses, re-estimating these models using nominal data yields substantially lower estimated elasticities, ranging from 1.17 to 1.52, which are again similar to those produced in the national analysis (1.40). These analyses therefore highlight that the little-utilized BEA state-level data yields estimates similar to the NHEA data and that the elasticities do appear sensitive to whether real or nominal values are used.

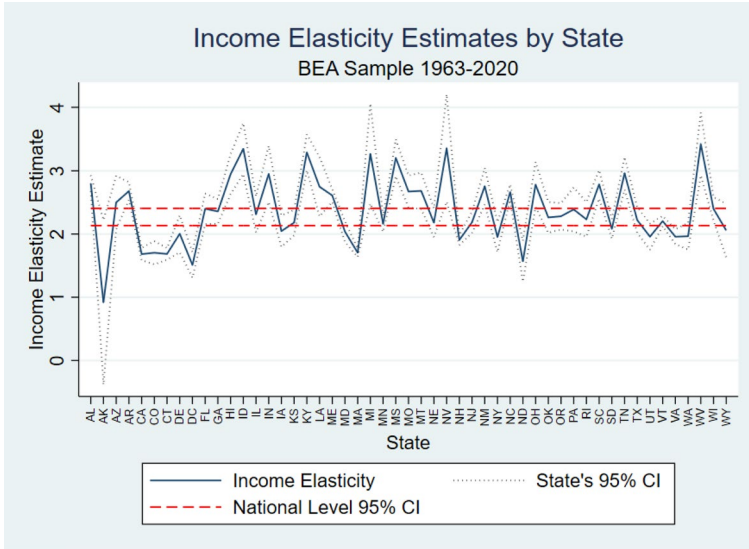
A final question is whether these estimates are stable over time, especially with respect to the period during and after the ACA. We therefore re-estimate the income elasticities for every state ending the sample first in 2010 before the ACA, and also in 2014 (thus omitting the years after most ACA policies were implemented). The bottom panel reports these three sets of estimates (using 1963–2010, 1963–2014 and 1963–2020) and clearly shows—once again—how little the estimated income elasticities are affected by extending the sample into the ACA era.

Evidence of Structural Breaks in Massachusetts’ log–log HCE–GDP Relationship

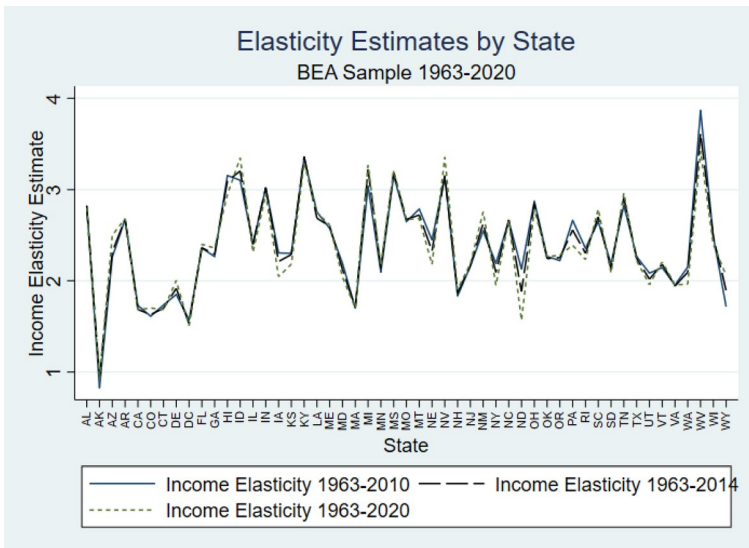
We now turn to testing for structural breaks in the state of Massachusetts, especially in the period immediately following its 2006 health care reform law. We analyze MA in more detail, as we view this exercise as a complement to the US analyses because of the similarity of the reform to the ACA; that the law occurred before the Great Recession and more time has elapsed makes it more straightforward to test than the ACA. Table 4 reports the structural break test results for both of the Massachusetts samples. The more commonly used NHEA sample is too short to permit testing for structural breaks around the 2006 reform with the traditional KP tests, and so, we



(a) State Income Elasticities & 95% Confidence Intervals



(b) State Income Elasticities Across Different Time Periods



Notes: Data for each state comes from the BEA. In panel (a), the full 1963-2020 time period is utilized. In panel (b), the straight line represents the elasticity estimated over the pre-ACA period or 1963-2010, while the long dashed line represents the elasticity for the whole time period up until the rollout of the ACA was mostly complete (1963-2014). The short dashed line represents the elasticity over the full time period (1963-2020).

Fig. 3 Health expenditure income elasticity across states and time



Table 4 Kejriwal and Perron (2010) & End-of-Sample Structural Break Tests for the Massachusetts log–log HCE–GDP Relationship

NHEA sample				BEA sample					
End-of-sample tests ^a				Kejriwal & Perron (2010) Test ^b		End-of-sample tests ^a			
Beginning of sample (Tested over 1983–1992)	End of sample (Tested over 2005–2014)			(Tested over 1972–2011)		Beginning of sample (Tested over 1968–1978)	End of sample (Tested over 2006–2016)		
Year	<i>p</i> -value	Year	<i>p</i> -value	KP tests	Test statistics (10% critical value)	Year	<i>p</i> -value	Year	<i>p</i> -value
1983	0.96	2005	0.86	<i>SupF</i> *(1)	7.344 (12.11)	1968	0.40	2006	0.74
1984	0.92	2006	0.69	<i>SupF</i> *(2)	4.403 (9.96)	1969	0.12	2007	0.86
1985	0.90	2007	0.78	<i>SupF</i> *(3)	6.060 (8.60)	1970	0.12	2008	0.84
1986	0.90	2008	0.60	<i>SupF</i> *(4)	4.544 (7.36)	1971	0.13	2009	0.82
1987	0.78	2009	0.73	<i>SupF</i> *(5)	10.907 (5.90)	1972	0.16	2010	0.80
1988	0.69	2010	0.83			1973	0.14	2011	0.89
1989	0.64	2011	0.81	UDMax	10.907 (12.25)	1974	0.18	2012	0.90
1990	0.58	2012	0.82			1975	0.16	2013	0.93
1991	0.70	2013	0.93	Sequential Test	0	1976	0.28	2014	0.93
1992	0.38	2014	0.81	(# of breaks)		1977	0.28	2015	0.96
						1978	0.20	2016	0.98

^a*p*-values, calculated through the sub-sampling routine proposed by Andrews and Kim (2006), are reported. In each case, a dynamic OLS model of the form $h_t = \alpha_{0j} + \beta_{1j}y_t + \sum_{i=1}^k \beta_{i+k}\Delta y_{t+i} + \sum_{i=0}^p \beta_{i+p}\Delta y_{t-i} + \epsilon_t$ (for $j = 1, \dots, (M + 1)$) is estimated. For the NHEA sample, the dynamic OLS model uses 0 leads and 2 lags, while the BEA sample uses 4 leads and 4 lags. Each year is tested separately and the null hypothesis is that there is no structural break. We test as far as we can to the beginning and end of the sample given the dynamic OLS lag/lead structure

^bFollowing Kejriwal (2008) and Woodward and Wang (2012), a dynamic OLS model of the form $h_t = \alpha_{0j} + \beta_{1j}y_t + \sum_{i=1}^2 \beta_{i+1}\Delta y_{t+i} + \sum_{i=0}^2 \beta_{i+3}\Delta y_{t-i} + \epsilon_t$ (for $j = 1, \dots, (M + 1)$) is tested in each case. Sequential tests using 2 leads and lags find no evidence of structural breaks. The null hypothesis is that there are no structural breaks. 10% critical values are in parentheses and come from Kejriwal and Perron (2010), Table 1A, nontrending case where $q_b = 1$

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

perform only the EOS P-test. While we are primarily interested in the end of the sample—the period after the reform—for completeness, we test at the beginning as well. Neither end of the sample provides evidence of a structural break.²⁷ The *p*-values during the reform period, in particular, are all quite high (0.60 or greater).

²⁷ We cannot use the KP tests for the Massachusetts NHEA sample because the sample size is too small to estimate given the amount of right-hand side variables needed for the DOLS estimation.



The BEA series, with its greater length, permits the full set of KP tests, and its interior 70% includes several years from the post-reform period (it includes 1972–2011). We can therefore perform two different types of tests to search for a structural break after the 2006 policy. Moreover, comparing the results of these more traditional structural break tests to those from EOS P-tests provides evidence on the validity of the P-tests used to test the years immediately following Medicare/Medicaid and the Great Recession/ACA in our US analyses. The full set of KP tests are reported in the first two columns of the BEA panel in Table 4, and the remaining columns report the EOS P-tests. For completeness, we again test both ends of the sample. None of these tests yield evidence of a structural break during either the ends of the sample or in the interior. In particular, there is no evidence of a structural break in 2006 or any subsequent year, corroborating the findings from the Great Recession/ACA era in the US analyses. Not only do the traditional KP tests find no breaks but the p -values from the EOS P-tests also strongly support this finding.

As in the state-level analyses, we also re-estimate the income elasticity both with and without the post-reform years included (i.e., ending in 2006 versus ending in 2014 for the NHEA or 2020 for the BEA). The income elasticity estimates are similar across time periods and data sets. Using the NHEA data, the estimated income elasticity changes from 1.70 to 1.61 when 2007–2014 is dropped. In the BEA data, the estimate changes from 1.71 to 1.70 when 2007–2020 is dropped. Both changes are strikingly similar and fall well within the estimated confidence intervals.

Concluding Remarks

The US has enacted many health care reforms over the last 60 years, most with the goal of expanding access to care and/or controlling cost, which in turn could presumably alter the relationship between GDP and health care expenditures. We explore the possible impact of these reforms by testing for structural breaks in the US health care expenditure–GDP relationship using the longest possible time series of annual data (1960–2020). However, because the most substantial reforms—Medicaid/Medicare and the ACA—occurred near the beginning and the end of this data, respectively, we also use tests that perform well over short time periods, such as the end of the sample. Neither the traditional tests on the internal 70% of the sample, nor the EOS tests yield evidence of a structural break. These results are consistent with the hypothesis that the relationship in general and income elasticity in particular is unchanged over the 1960–2020 period. This conclusion is robust to using both real and nominal measures and controlling for additional factors.

This stability and our consistent finding that the US income elasticity of health expenditures is greater than one (and thus a luxury good) also provides empirical support for the conclusions from Hall and Jones (2007). That research presents a model based on standard assumptions and shows that a rising health expenditure share is the rational response to increased income because while the marginal utility of consumption falls with income, the marginal utility of extending life (obtained through more health spending) does not decrease. As income rises, it is therefore



utility maximizing to shift spending from general consumption toward the relatively more valuable health expenditures, a response that would explain the stable and high-income elasticity we document here. While some aspect of rising health expenditures is likely the result of the various inefficiencies detailed in the “1 Percent Steps for health care Reform Project,” the fact that the income elasticity remains stable and high even when controlling for the impact of technological change, the rising age share and the influence of other unobserved factors suggests that there is also another, more fundamentally entrenched side to the driving forces behind the rise in health expenditures.

We caution, however, against interpreting our findings about past reforms as necessarily being predictive of any impacts of future policies. Put simply, our analyses suggest that the major reforms put in place so far do not appear to have fundamentally changed the US HCE–GDP relationship. Future reforms such as those offered on the 1% steps website could very well have an effect, as the studies listed there suggest. Indeed, the empirical methodology presented here offers one approach to testing their success, if enacted.

A challenge to investigating the ACA era is its staggered rollout and coincidental timing with the Great Recession and subsequent recovery. Our parallel analyses using Massachusetts data address these challenges by testing for a structural break in the years following its 2006 health care reform, a reform that is widely viewed as the blueprint for the ACA (Gruber 2013). Its earlier adoption means that more years of post-reform data are available, which permits both traditional KP and EOS structural break tests. The Great Recession is also less of a confounding factor given the reform’s earlier timing and the lesser impact of the Great Recession on Massachusetts. Similarly, the tests provide no evidence of a break in Massachusetts’ HCE–GDP relationship.

All of our analyses therefore lend support to the hypothesis that neither the initiation of Medicare/Medicaid nor the ACA altered the health expenditure income elasticity or overall log–log HCE–GDP relationship. A final concern is that these tests may simply suffer from a lack of statistical power; while we use the best available data, our data series are still fairly short. In addition, the structural break tests consider the entire relationship, including the intercept, rather than only the income elasticity (the slope). Reassurance regarding these concerns comes from the remarkable precision, consistency and stability of the estimated income elasticities from these samples. Taken together, our results suggest that US health care is a luxury good and that the HCE–GDP relationship in general and health expenditure income elasticity in particular have remained quite stable since 1960 despite numerous reforms to health care policy.

Appendix

Specifying the DOLS Estimating Equation

As illustrated in Table 1 and described in the Time Series Properties section, h_t and y_t are non-stationary and cointegrated in all three data series used in this paper. As



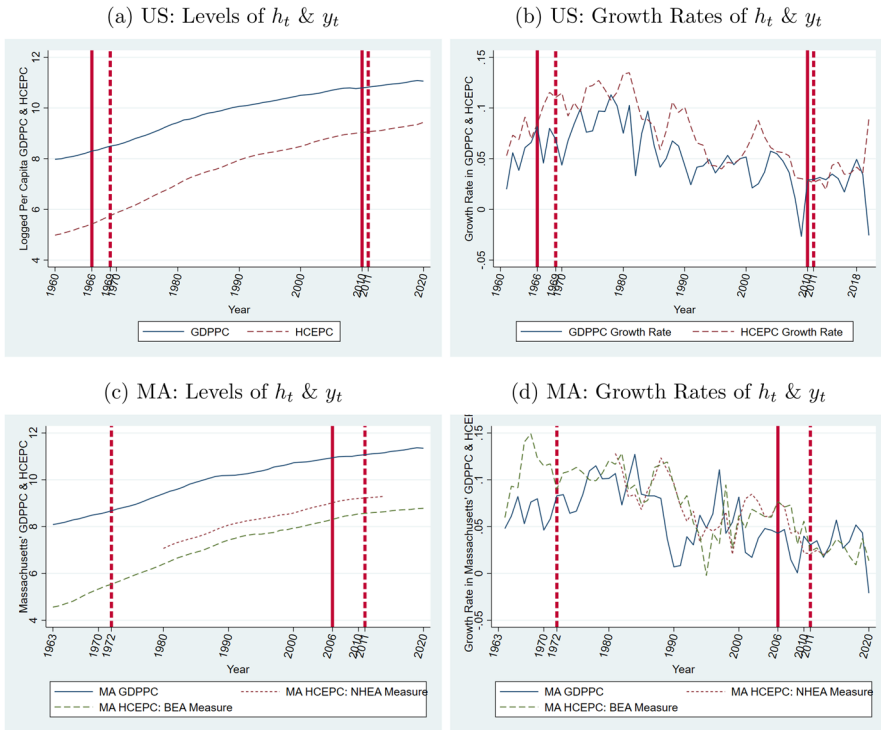


Fig. A1 Illustrating the Time-Series Properties of h_t & y_t . Notes: US estimation sample is annual from 1960 to 2020. Massachusetts estimation sample is from 1980 to 2014 when using the National Health Expenditure Account’s measure and 1963–2020 when using the Bureau of Economic Analysis’s measure. Health expenditures (HCE) and GDP are both in per capita terms in all figures. Solid vertical lines represent major health care reforms at the national level (Medicare/Medicaid in 1966 and the Affordable Care Act in 2010) and for Massachusetts (Massachusetts health reform in 2006). Years in between the dashed lines represent testable years once the standard 15% trimming is applied for the Kejriwal & Perron (2010) tests, while years outside the dashed lines are untestable by that procedure (Table 8)

also noted in the Time Series Properties section, to better account for the effect of these time series properties, we appeal to the DOLS estimation procedure developed by Stock and Watson (1993) by augmenting our basic estimating equation 1 with leads and lags of first differenced y_t . As is standard, we use the AIC and SBIC to determine the optimal amount of leads and lags needed for each of the three data series independently. We then consider whether we can eliminate the leads to gain efficiency as proposed by Hayakawa and Kurozumi (2008). Here, we first lay out the steps needed to make this decision in general. We then discuss the findings on the need for leads in the DOLS specification for each data series separately and present these results in the corresponding Appendix Tables 5, 6, 7.

To test whether or not leads are required in the DOLS estimating equation, we start by estimating the simple regression presented in the Data and Empirical Strategy section as equation 1.



Table 5 DOLS Model Specification: US (1960–2020)

	(1)	(2)	(3)	(4)
Step 1: Lag Order of Δy_t^a				
Testing Equation: $\Delta y_t = \rho_0 + \rho_1 \Delta y_{t-1} + \rho_2 \Delta y_{t-2} + \rho_3 \Delta y_{t-3} + \rho_4 \Delta y_{t-4} + u_t$				
	Coefficient Estimates			
Δy_{t-1}	0.432*** (2.99)	-4.766	-4.738	-4.694†
Δy_{t-2}	-0.228 (-1.46)	-4.792†	-4.750†	-4.683
Δy_{t-3}	-0.050 (-0.32)	-4.760	-4.704	-4.616
Δy_{t-4}	-0.042 (-0.29)	-4.726	-4.656	-4.545
Step 2: Does $\hat{\varepsilon}_t$ Granger-cause Δy_t ? ^b				
Testing Equation: $\Delta y_t = \rho_0 + \rho_1 \Delta y_{t-1} + \eta_1 \hat{\varepsilon}_{t-1} + \eta_2 \hat{\varepsilon}_{t-2} + e_t$				
	Coefficient Estimates			
Δy_{t-1}	1.365*** (4.63)			
$\hat{\varepsilon}_{t-1}$	0.535*** (4.04)			
$\hat{\varepsilon}_{t-2}$	-0.401*** (-3.10)			
Step 3: Joint Test that Coefficients on Lags of $\hat{\varepsilon}_t=0$ ^b				
Joint Test: $\eta_1 = \eta_2 = 0$				
	F-Statistic			
	13.92***			
Step 4: Are Lead-less DOLS Residuals Serially Uncorrelated? ^c				
Testing Equation: $h_t = a_{0,j} + \beta_{1,j} y_t + \sum_{i=0}^2 \beta_{i+1,j} \Delta y_{t-i} + v_t$				
		(6)		
		Cumby-Huizinga		
		χ^2 Statistic		



Table 5 (continued)

Step 1: Lag Order of Δy_t^a	(1)	(2)	(3)	(4)
Testing Equation: $\Delta y_t = \rho_0 + \rho_1 \Delta y_{t-1} + \rho_2 \Delta y_{t-2} + \rho_3 \Delta y_{t-3} + \rho_4 \Delta y_{t-4} + u_t$				
	Coefficient Estimates			
		AIC	HQIC	SBIC
		38.142***		

All estimation samples are at the national level and annual from 1960 to 2020. t-statistics are in parentheses

^aDependent variable is the first difference of per capita real GDP (Δy_t). In column (1), the first four lags of Δy_t are included simultaneously. In columns (2), (3) and (4), the information criteria come from testing models lag orders of $k = 1, 2, 3, 4$ lags against a model with none. The optimal model is designated by the lowest information criteria and is denoted by †.

^bDependent variable is the first difference of per capita real GDP (Δy_t). $\hat{\epsilon}_t$ are the residuals from Equation 1 (a regression of h_t on y_t). The null hypothesis is that $\hat{\epsilon}_t$ does not Granger-cause Δy_t or that past lags of $\hat{\epsilon}_t$ do not help predict Δy_t or that $\eta_1 = \eta_2 = 0$.

^cDependent variable is per capita real national health expenditures (h_t). The null hypothesis is that ϵ_t is not serially correlated

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 6 DOLS model specification: MA NHEA sample (1980-2014)

	(1)	(2)	(3)	(4)
Step 1: Lag Order of Δy_t ^a				
Testing Equation: $\Delta y_t = \rho_0 + \rho_1 \Delta y_{t-1} + \rho_2 \Delta y_{t-2} + \rho_3 \Delta y_{t-3} + \rho_4 \Delta y_{t-4} + u_t$				
	Coefficient Estimates			
Δy_{t-1}	0.428** (2.52)	-4.602 [†]	-4.572 [†]	-4.508 [†]
Δy_{t-2}	0.049 (0.26)	-4.540	-4.495	-4.400
Δy_{t-3}	0.236 (1.29)	-4.517	-4.457	-4.330
Δy_{t-4}	-0.405** (-2.52)	-4.512	-4.437	-4.278
Step 2: Does $\hat{\epsilon}_t$ Granger-cause Δy_t ? ^b				
Testing Equation: $\Delta y_t = \rho_0 + \rho_1 \Delta y_{t-1} + \eta_1 \hat{\epsilon}_{t-1} + \eta_2 \hat{\epsilon}_{t-2} + \epsilon_t$				
	Coefficient Estimates			
Δy_{t-1}	0.812* (1.88)			
$\hat{\epsilon}_{t-1}$	0.176 (0.69)			
$\hat{\epsilon}_{t-2}$	-0.066 (-0.30)			
	F-Statistic 1.95			
Step 3: Joint Test that Coefficients on Lags of $\hat{\epsilon}_t=0$ ^b				
	Joint Test: $\eta_1 = \eta_2 = 0$			
Step 4: Are Lead-less DOLS Residuals Serially Uncorrelated? ^c				
			(6)	
Testing Equation: $h_t = \alpha_{0,j} + \beta_{1,j} y_t + \sum_{i=0}^2 \beta_{i+1,j} \Delta y_{t-i} + v_t$				Cumby-Huizinga χ^2 Statistic



Table 6 (continued)

Step 1: Lag Order of Δy_t ^a	(1)	(2)	(3)	(4)
Testing Equation: $\Delta y_t = \rho_0 + \rho_1 \Delta y_{t-1} + \rho_2 \Delta y_{t-2} + \rho_3 \Delta y_{t-3} + \rho_4 \Delta y_{t-4} + u_t$				
	Coefficient Estimates			
		AIC	HQIC	SBIC
		3.402		

All estimation samples use data from the National Health Expenditure Accounts for Massachusetts and are annual from 1980 to 2014. t-statistics are in parentheses

^aDependent variable is the first difference of per capita real state GDP (Δy_t). In column (1), the first four lags of Δy_t are included simultaneously. In columns (2), (3) and (4), the information criteria come from testing models lag orders of $k = 1, 2, 3, 4$ lags against a model with none. The optimal model is designated by the lowest information criteria and is denoted by †

^bDependent variable is the first difference of per capita real state GDP (Δy_t). $\hat{\epsilon}_t$ are the residuals from Equation 1 (a regression of h_t on y_t). The null hypothesis is that $\hat{\epsilon}_t$ does not Granger-cause Δy_t or that past lags of $\hat{\epsilon}_t$ do not help predict Δy_t , or that $\eta_1 = \eta_2 = 0$

^cDependent variable is per capita Massachusetts real health expenditures (h_t). The null hypothesis is that ϵ_t is not serially correlated

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 7 DOLS Model Specification: MA BEA Sample (1963-2020)

Step 1: Lag Order of Δy_t ^a	(1)	(2)	(3)	(4)
Testing Equation: $\Delta y_t = \rho_0 + \rho_1 \Delta y_{t-1} + \rho_2 \Delta y_{t-2} + \rho_3 \Delta y_{t-3} + \rho_4 \Delta y_{t-4} + u_t$				
	Coefficient Estimates	AIC	HQIC	SBIC
	Δy_{t-1} 0.593*** (4.06)	-8.955	-8.870	-8.734
	Δy_{t-2} -0.340** (-2.03)	-9.335	-9.914 [†]	-8.967 [†]
	Δy_{t-3} 0.212 (1.26)	-9.375	-9.176	-8.859
	Δy_{t-4} -0.257* (-1.78)	-9.420 [†]	-9.165	-8.758
Step 2: Does $\hat{\varepsilon}_t$ Granger-cause Δy_t ? ^b	(5)			
Testing Equation: $\Delta y_t = \rho_0 + \rho_1 \Delta y_{t-1} + \rho_2 \Delta y_{t-2} + \eta_1 \hat{\varepsilon}_{t-1} + \eta_2 \hat{\varepsilon}_{t-2} + \eta_3 \hat{\varepsilon}_{t-3} + \eta_4 \hat{\varepsilon}_{t-4} + e_t$				
	Coefficient Estimates			
	Δy_{t-1} 1.059*** (3.76)			
	Δy_{t-2} -0.309 (-1.15)			
	$\hat{\varepsilon}_{t-1}$ 0.373** (2.26)			
	$\hat{\varepsilon}_{t-2}$ -0.318 (-1.26)			
	$\hat{\varepsilon}_{t-3}$ -0.137 (-0.78)			
	$\hat{\varepsilon}_{t-4}$ 0.223** (2.69)			
Step 3: Joint Test that Coefficients on Lags of $\hat{\varepsilon}_t=0$ ^b				
Joint Test: $\eta_1 = \eta_2 = \eta_3 = \eta_4 = 0$	F-Statistic			
	3.18**			
Step 4: Are Lead-less DOLS Residuals Serially Uncorrelated? ^c		(6)		
Testing Equation: $h_t = \alpha_{0,j} + \beta_{1,j} y_t + \sum_{i=0}^4 \beta_{i+1,j} \Delta y_{t-i} + v_t$				
		Cumby-Huizinga		
		χ^2 Statistic		
		38.43***		

All estimation samples use data from the Bureau of Economic Analysis for Massachusetts and are annual from 1963 to 2020. t-statistics are in parentheses

^aDependent variable is the first difference of per capita real state GDP (Δy_t). In column (1), the first four lags of Δy_t are included simultaneously. In columns (2), (3) and (4), the information criteria come from testing models lag orders of $k = 1, 2, 3, 4$ lags against a model with none. The optimal model is designated by the lowest information criteria and is denoted by [†]

^bDependent variable is the first difference of per capita real state GDP (Δy_t). $\hat{\varepsilon}_t$ are the residuals from Equation 1 (a regression of h_t on y_t). The null hypothesis is that $\hat{\varepsilon}_t$ does not Granger-cause Δy_t or that past lags of $\hat{\varepsilon}_t$ do not help predict Δy_t or that $\eta_1 = \eta_2 = \eta_3 = \eta_4 = 0$

^cDependent variable is per capita Massachusetts real health expenditures (h_t). The null hypothesis is that ε_t is not serially correlated

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$



Table 8 Replicating and Updating Woodward and Wang (2012)

KP Tests	(1) Woodward and Wang (2012) Sample 1960-2008	(2) Our Sample 1960-2008	(3) Our Sample 1960-2020
(10% Critical Values)	(Tested over 1968-2001)	(Tested over 1968-2001)	(Tested over 1969-2011)
<i>SupF</i> *(1)	11.367 (12.11)	10.760 (12.11)	11.653 (12.11)
<i>SupF</i> *(2)	5.061 (9.96)	5.439 (9.96)	3.567 (9.96)
<i>SupF</i> *(3)	4.663 (8.60)	4.942 (8.60)	4.772 (8.60)
<i>SupF</i> *(4)	7.123 (7.36)	8.839 (7.36)	5.318 (7.36)
<i>SupF</i> *(5)	4.569 (5.90)	6.467 (5.90)	4.533 (5.90)
UDMax	11.367 (12.25)	10.760 (12.25)	11.653 (12.25)
Sequential Test (# of breaks)	0	0	0

Following Woodward and Wang (2012), both h_t and y_t are specified in nominal terms in each column. Following Kejriwal (2008) and Woodward and Wang (2012), a dynamic OLS model of the form $h_t = \alpha_{0,j} + \beta_{1,j}y_t + \sum_{i=1}^5 \beta_{i+1} \Delta y_{t+i} + \sum_{i=0}^5 \beta_{i+6} \Delta y_{t-i} + \epsilon_t$ (for $j = 1, \dots, (M + 1)$) is estimated in each case. Sequential tests using five leads and lags find no evidence of structural breaks. The null hypothesis is that there are no structural breaks. 10% critical values are in parentheses and come from Kejriwal and Perron (2010), Table 5, nontrending case where $q_b=1$. Results in column (1) come from Appendix Table 5 of Woodward and Wang (2012)

$$h_t = \alpha + \beta y_t + \epsilon_t \text{ for } t = 1, \dots, T$$

To determine whether or not we can improve the efficiency of the DOLS estimation through eliminating the leads of Δy_t , we consider a Granger causality test between ϵ_t (proxied for using the residuals $\hat{\epsilon}_t$ from the simple regression above), and the first differenced right-hand side variable, Δy_t . To test the null hypothesis that ϵ_t does not Granger-cause Δy_t , we first find the appropriate autoregressive specification for Δy_t by determining when past values of Δy_t no longer help predict Δy_t via an AR(k) model. Before being estimated, this Δy_t AR(k) function is augmented with lagged values of $\hat{\epsilon}_t$.

$$\Delta y_t = \rho_0 + \sum_{k=1}^K \rho_k \Delta y_{t-k} + \sum_{p=1}^P \eta_p \hat{\epsilon}_{t-p} + e_t$$

If the lagged values of $\hat{\epsilon}_t$ collectively add no explanatory power to the Δy_t AR(k) function, then the past values of $\hat{\epsilon}_t$ do not help predict Δy_t , and therefore, the cointegration errors do not Granger-cause the first differenced right-hand side variable.



To actually test the explanatory power of the cointegrating errors, we consider a joint test on whether all the lags of $\hat{\varepsilon}_t$ are equal to 0. If we reject the null hypothesis here, then it suggests we do need to include the leads to mitigate the cointegration bias.

As an additional confirmation, we follow the suggestion by Hayakawa and Kurozumi (2008) to also test the DOLS specification's errors for serial correlation when leads are not included. If adding the lags is enough to eliminate the persistence, then it is another sign that the power gains from eliminating the leads does not introduce much bias. To do so, we estimate a DOLS equation without leads and then test the errors from this estimation for serial correlation as described above.

DOLS Specification Results for the National Sample

The results for whether leads need to be included when estimating structural change for the national sample are presented in Appendix Table 5. As shown in columns (1) to (4), a single lag is optimal when modeling the autocorrelation structure of Δy_t . Based on the AIC, HQIC and SBIC criterion (not reported in Appendix Table 5 but available upon request), we add two lags of ε_t to that Δy_t AR(1) specification. We then test whether those two lags of ε_t are jointly equal to 0, and as the results in column (5) show, we are able to reject this claim at the 1% significance level. As such, we conclude that the cointegration errors Granger-cause Δy_t for the national sample, meaning we must include both leads and lags of Δy_t in that DOLS specification. As column (6) illustrates, this conclusion is further confirmed when testing for serial correlation in the lead-less DOLS specification. Combining these findings with the AIC and SBIC criterion on the lag structure, we estimate a DOLS model with 2 leads and 2 lags when using data from the national sample.

DOLS Specification Results for the Massachusetts NHEA Sample

Appendix Table 6 presents the results for the Massachusetts NHEA sample. Similar to the national level sample, the evidence suggests that a single lag of Δy_t is optimal (columns (1) to (4)). The information criteria also suggest adding two lags of ε_t to the AR(1) specification. However, the Massachusetts NHEA sample differs from its national counterpart in that we fail to reject the null hypothesis that the cointegration errors are jointly 0, suggesting that they do not Granger-cause Δy_t . The serial correlation results presented in column (6) suggest the same conclusion. It is therefore optimal to eliminate the leads when estimating the DOLS model on the Massachusetts NHEA sample. Combining these findings with the AIC and SBIC criterion on the lag structure, we estimate a DOLS model with 0 leads and 2 lags when using data from the Massachusetts NHEA sample.

DOLS Specification Results for the Massachusetts BEA Sample

Finally, Appendix Table 7 presents the results for the Massachusetts BEA sample. Here, the autocorrelation structure of Δy_t is slightly different, as the information



criterion in columns (1) to (4) suggest either two lags or four lags are optimal. The error structure tested is also longer, as the different information criteria suggest adding 4 lags of ε_t to the AR specification. As shown in column (5), we are able to reject the null hypothesis that the cointegration errors are jointly equal to 0 at the 5% level and therefore must add leads to the DOLS specification. The serial correlation result in column (6) is strongly significant, even at the 1% level, confirming the need for leads. Based on the AIC and SBIC criterion on the lag structure, we estimate a DOLS model with four leads and four lags when using data from the Massachusetts BEA sample.

Structural Break Test Statistics

Kejriwal and Perron (2010) Sup-Wald Test

The null hypothesis of no structural breaks is tested against the alternative hypothesis of a predetermined number of breaks (l breaks). The test statistic is given by:

$$\sup F_T^*(l) = \sup_{\lambda \in \Lambda} \frac{SSR_0 - SSR_l}{\hat{\sigma}^2}$$

where $\hat{\sigma}^2$ is an estimate of the long-run variance, SSR_0 is the sum of squared residuals with no structural breaks, SSR_l is the sum of squared residuals with l breaks, λ represents the break fraction or $\frac{T_i}{T}$ (where T_i is the fraction of the sample corresponding to the break and T is the total sample) and Λ represents the entire set of potential break fractions.

Kejriwal and Perron (2010) UD Max Test

The null hypothesis of no structural breaks is tested against the alternative hypothesis of an unknown number of breaks (as in Kejriwal (2008), the 15% trimming means that the maximum number of breaks (l) is bounded at 5: $1 \leq l \leq 5$). The test statistic is given by:

$$UD \max F_T^* = \max_{1 \leq j \leq 5} F_T^*(l)$$

where $F_T^*(l)$ are the sup-Wald statistics calculated in the initial sup-Wald tests.

Kejriwal and Perron (2010) Sequential Test

The null hypothesis of l structural breaks is tested against the alternative hypothesis of $l + 1$ breaks. The test statistic is given by:

$$SEQ_T(l + 1 | l) = \max_{1 \leq k \leq l+1} \sup_{\tau \in \Lambda} T\{SSR_T(\hat{T}_1, \dots, \hat{T}_l) - SSR_T(\hat{T}_1, \dots, \hat{T}_{k-1}, \tau, \hat{T}_k, \dots, \hat{T}_l)\} / SSR_{l+1}$$

where $\Lambda = \{\tau : \hat{T}_{k-1} + (\hat{T}_k - \hat{T}_{k-1})\epsilon \leq \tau \leq \hat{T}_k - (\hat{T}_k - \hat{T}_{k-1})\epsilon\}$.



Andrews and Kim (2006) End-of-Sample Structural Break P Test Statistic

The null hypothesis of no structural breaks is tested against 4 the alternative hypothesis that a structural break occurred at the pre-specified point of t_j . The test statistic is given by:

$$\hat{\epsilon}_t = h_t - \hat{\alpha}_0 - \hat{\beta}_1 y_t - \sum_{i=1}^k \hat{\beta}_{i+1} \Delta y_{t+i} - \sum_{i=0}^q \hat{\beta}_{i+k+2} \Delta y_{t-i} \quad \text{for } t = 1, \dots, T$$

$$P = \sum_{t=T-(T-t_j)+1}^T \hat{\epsilon}_t^2$$

where k and q denote the number of leads and lags, respectively.

Funding The authors have not received any funding or support for this research, nor are there any sponsors of this project.

Declarations

Conflict of interest As such, there are no potential conflicts of interest for any of the authors.

References

- Andrews, Donald. 1993. "Tests for Parameter Instability and Structural Change with Unknown Change Point." *Econometrica*, 821–856.
- Andrews, Donald. 2003. End of sample instability tests. *Econometrica* 71 (6): 1661–1694.
- Andrews, Donald, and Jae-Young Kim. 2006. Tests for cointegration breakdown over a short time period. *Journal of Business & Economic Statistics* 24 (4): 379–394.
- Bai, Jushan, and Pierre Perron. 1998. "Estimating and Testing Linear Models with Multiple Structural Changes." *Econometrica*, 47–78.
- Bai, Jushan, and Pierre Perron. 2003. Computation and analysis of multiple structural change models. *Journal of Applied Econometrics* 18 (1): 1–22.
- Baltagi, Badi, and Francesco Moscone. 2010. Health care expenditure and income in the OECD reconsidered: Evidence from panel data. *Economic Modelling* 27 (4): 804–811.
- Baltagi, Badi, Raffaele Lagravinese, Francesco Moscone, and Elisa Tosetti. 2017. Health care expenditure and income: A global perspective. *Health Economics* 26 (7): 863–874.
- Barati, Mehdi, and Hadiseh Fariditavane. 2020. Asymmetric effect of income on the US healthcare expenditure: Evidence from the nonlinear autoregressive distributed lag (ARDL) approach. *Empirical Economics* 58 (4): 1979–2008.
- Blazquez-Fernandez, Carla, David Cantarero, and Patricio Perez. 2014. Disentangling the heterogeneous income elasticity and dynamics of health expenditure. *Applied Economics* 46 (16): 1839–1854.
- Blomqvist, Åke, and Richard Carter. 1997. Is health care really a luxury? *Journal of Health Economics* 16 (2): 207–229.
- Carrion-i-Silvestre, Josep Lluís. 2005. Health care expenditure and GDP: Are they broken stationary? *Journal of Health Economics* 24 (5): 839–854.
- Casini, Alessandro, and Pierre Perron. 2019. *Structural Breaks in Time Series*. Oxford University Press.



- Chandra, Amitabh, Jonathan, Holmes, and Jonathan, Skinner. 2013. "Is This Time Different? The Slowdown in Healthcare Spending." National Bureau of Economic Research.
- Cheng, Chak Hung Jack, and Nopphol, Witvorapong. 2019. "Health Care Policy Uncertainty, Real Health Expenditures and Health Care Inflation in the USA." *Empirical Economics*, 1–21.
- Clemente, Jesus, Antonio Montanes, and Marcelo Reyes. 1998. Testing for a unit root in variables with a double change in the mean. *Economics Letters* 59 (2): 175–182.
- CMS. 2019. "Health Expenditures by State of Provider, 1980–2014." <https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/NationalHealthExpendData/Downloads/prov-tables.zip>. Accessed June 2, 2019.
- CMS. 2021a. "NHE Summary Including Share of GDP, CY 1960–2020." NHE Summary including share of GDP, CY 1960–2020. <https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/NationalHealthExpendData/Downloads/NHEGDP20.zip>. Accessed December 23, 2021.
- CMS. 2021b. "National Health Expenditures by Type of Service and Source of Funds, CY 1960–2020." <https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/NationalHealthExpendData/Downloads/NHE2020.zip>. Accessed December 23, 2021.
- Cutler, David. 2010. How health care reform must bend the cost curve. *Health Affairs* 29 (6): 1131–1135.
- Ditzen, Jan, Yiannis Karavias, and Joakim Westerlund. 2021. "Testing and Estimating Structural Breaks in Time Series and Panel Data in Stata."
- Elliot, Graham, Thomas Rothenberg, and James Stock. 1996. Efficient tests for an autoregressive unit root. *Econometrica* 64 (8): 13–36.
- Farag, Marwa, Allyala NandaKumar, Stanley Wallack, Dominic Hodgkin, Gary Gaumer, and Can Erbil. 2012. The income elasticity of health care spending in developing and developed countries. *International Journal of Health Care Finance and Economics* 12 (2): 145–162.
- Freeman, Donald. 2003. Is health care a necessity or a luxury? Pooled estimates of income elasticity from US state-level data. *Applied Economics* 35 (5): 495–502.
- Gaffney, Adam, Danny McCormick, David Bor, Anna Goldman, Steffie Woolhandler, and David Himmelstein. 2019. The effects on hospital utilization of the 1966 and 2014 health insurance coverage expansions in the United States. *Annals of Internal Medicine* 171 (3): 172–180.
- Gerdtham, Ulf-G., and Mickael Löthgren. 2000. On stationarity and cointegration of international health expenditure and GDP. *Journal of Health Economics* 19 (4): 461–475.
- Gerdtham, Ulf-G., Jes Søgaard, Fredrik Andersson, and Bengt Jönsson. 1992. An econometric analysis of health care expenditure: A cross-section study of the OECD countries. *Journal of Health Economics* 11 (1): 63–84.
- Gruber, Jonathan. 2013. Evaluating the massachusetts health care reform. *Health Services Research* 48 (6 Pt 1): 1819.
- Hall, Robert, and Charles Jones. 2007. The value of life and the rise in health spending. *The Quarterly Journal of Economics* 122 (1): 39–72.
- Hartman, Michah, Robert Kornfield, and Aaron Catlin. 2010. "Health care expenditures in the national health expenditures accounts and in gross domestic product: A reconciliation." *Bureau of Economic Analysis Working Paper*.
- Hartwig, Jochen. 2011. Can Baumol's model of unbalanced growth contribute to explaining the secular rise in health care expenditure? An Alternative Test. *Applied Economics* 43 (2): 173–184.
- Hartwig, Jochen, and Jan-Egbert Sturm. 2014. Robust determinants of health care expenditure growth. *Applied Economics* 46 (36): 4455–4474.
- Hayakawa, Kazuhiko, and Eiji Kurozumi. 2008. The role of "Leads" in the dynamic OLS estimation of cointegrating regression models. *Mathematics and Computers in Simulation* 79 (3): 555–560.
- Hitiris, Theo. 1997. Health care expenditure and integration in the countries of the European Union. *Applied Economics* 29 (1): 1–6.
- Kaiser Family Foundation. 2011. Timeline: History of Health Reform in the US. <https://www.kff.org/wp-content/uploads/2011/03/5-02-13-history-of-health-reform.pdf>. Accessed June 2, 2019.
- Kaiser Family Foundation. 2013. Health Reform Implementation Timeline: <https://www.kff.org/interactive/implementation-timeline>. Accessed June 2, 2019.
- Kejriwal, Mohitosh. 2008. "Cointegration with structural breaks: An application to the Feldstein-Horioka Puzzle." *Studies in Nonlinear Dynamics & Econometrics*, 12(1).
- Kejriwal, Mohitosh, and Pierre Perron. 2010. Testing for multiple structural changes in cointegrated regression models. *Journal of Business & Economic Statistics* 28 (4): 503–522.



- Lee, Hyejin, Oh, Dong-Yop, and Ming Meng. 2019. Stationarity and cointegration of health care expenditure and GDP: Evidence from tests with smooth structural shifts. *Empirical Economics* 57 (2): 631–652.
- Mosccone, Francesco, and Elisa Tosetti. 2010. Health expenditure and income in the United States. *Health Economics* 19 (12): 1385–1403.
- Mosccone, Francesco, and Elisa Tosetti. 2010b. “Testing for error cross section independence with an application to US health expenditure.” *Regional Science and Urban Economics*, 40(5): 283–291.
- Newhouse, Joseph. 1977. Medical-care expenditure: A cross-national survey. *Journal of Human Resources* 12 (1): 115–125.
- Newhouse, Joseph. 1992. Medical care costs: How much welfare loss? *Journal of Economic Perspectives* 6 (3): 3–21.
- Ng, Serena, and Pierre Perron. 2001. Lag length selection and the construction of unit root tests with good size and power. *Econometrica* 69 (6): 1519–1554.
- Obama, Barack. 2010. “Remarks At the House Democratic Caucus.” <http://www.presidency.ucsb.edu/ws/index.php?pid=78955&st=curve&st1=#axzz1c15DPjMx>.
- OECD. 2020. Data includes “Expenditure on health, All financing schemes, Current expenditures, All providers, Share of gross domestic product, All countries, All years.”. <https://stats.oecd.org/#>.
- Panopoulou, Ekaterini, and Theologos Pantelidis. 2013. Cross-state disparities in US health care expenditures. *Health Economics* 22 (4): 451–465.
- Parkin, David, Alistair McGuire, and Brian Yule. 1987. Aggregate health care expenditures and national income: Is health care a luxury good? *Journal of Health Economics* 6 (2): 109–127.
- Perron, Pierre. 1990. Testing for a unit root in a time series with a changing mean. *Journal of Business & Economic Statistics* 8 (2): 153–162.
- Perron, Pierre. 2006. Dealing with structural breaks. *Palgrave Handbook of Econometrics* 1: 278–352.
- Perron, Pierre, and Serena Ng. 1996. Useful modifications to some unit root tests with dependent errors and their local asymptotic properties. *Review of Economic Studies* 63 (3): 435–463.
- Piehl, Anne Morrison, Suzanne Cooper, Anthony Braga, and David Kennedy. 2003. Testing for structural breaks in the evaluation of programs. *Review of Economics and Statistics* 85 (3): 550–558.
- Skinner, Jonathan, and Amitabh Chandra. 2016. The past and future of the affordable care act. *JAMA* 316 (5): 497–499.
- Stock, James, and Mark Watson. 1993. A simple estimator of cointegrating vectors in higher order integrated systems. *Econometrica* 61 (4): 783–820.
- Wang, Zijun. 2009. The determinants of health expenditures: Evidence from US state-level data. *Applied Economics* 41 (4): 429–435.
- Wang, Zijun, and Andrew Rettenmaier. 2007. A note on cointegration of health expenditures and income. *Health Economics* 16 (6): 559–578.
- Weiner, Janet, Clifford Marks, and Mark Pauly. 2017. Effects of the ACA on health care cost containment. *LDI issue brief* 24 (4): 1–7.
- Woodward, Robert, and Le. Wang. 2012. The Oh-So straight and narrow path: Can the health care expenditure curve be bent? *Health Economics* 21 (8): 1023–1029.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

