Social Science & Medicine 332 (2023) 116094

Contents lists available at ScienceDirect

Social Science & Medicine

journal homepage: www.elsevier.com/locate/socscimed

Reconsidering the relationship between health and income in the UK

Rosen Chowdhury [a], Steve Cook [a,*], Duncan Watson [b]

[a] School of Social Sciences, Swansea University, United Kingdom
[b] School of Economics, University of East Anglia, United Kingdom

ARTICLE INFO

Keywords:
Life expectancy
Income
Cointegration
Structural change

ABSTRACT

The present paper revisits and extends the examination of the long-run relationship between UK life expectancy and income provided by Tapia Granados (2012). Adopting a more detailed form of analysis, a clear break corresponding to the 1918–1919 Influenza Pandemic is identified in the long span of data examined. This finding of structural change, along with detected uncertainty regarding the orders of integration of the series examined, results in the application of split-sample analysis employing autoregressive distributed lag (ARDL) modelling. The results obtained reverse the 'no long-run relationship' conclusion of Tapia Granados (2012) with overwhelming evidence presented in support of a negative relationship between life expectancy and income. Our findings add to both health-income research and a burgeoning literature on the reproduction and replication of previously published empirical research.

1. Introduction

Examination of the potential relationship between health and income has generated an extensive empirical literature which is notable for its continued debate and dissent. This can be illustrated by consideration of the Preston Curve (Preston, 1975). Depicting a supposed relationship between higher levels of income and higher levels of health across a range of economies, its relevance and exact nature continue to attract attention (e.g. Bloom and Canning, 2007; Prados de la Escosura, 2023). Further examples of scholarly disagreement regarding the health-income relationship, including questioning of its very existence, are revealed in a number of exchanges including: Acemoglu and Johnson (2007, 2014) and Bloom et al. (2014); and Tapia Granados (2015), Tapia Granados and Ionides (2015), Catalano et al. (2011) and Catalano and Bruckner (2016). The present paper adds to these ongoing debates by focussing upon a 'time series' sub-literature examining the health-income relationship where uncertainty and a lack of consensus are apparent. The particular issue we consider is whether a long-run relationship exists between life expectancy and income. The conflict present in previous research exploring this issue is apparent in studies such as Arora (2001), Bishai (1995), Brenner (2005) and Tapia Granados (2005, 2012). While Brenner (2005) discusses cointegration in the analysis of infant mortality and macroeconomic indicators for the US, where cointegration refers to a long-run relationship between time series processes (see Hendry and Juselius, 2000; 2001), this work has been subsequently criticised by Tapia Granados (2005). Similarly, although Tapia Granados (2012) notes that Arora (2001) '*has asserted that income and health are cointegrated in Britain*' (Tapia Granados, 2012, p.690) this followed by the comment that '*I could not reproduce Arora's results*' (Tapia Granados, 2012, p.690). This uncertainty is compounded by studies which argue against the presence of a long-run relationship such as Bishai (1995) and Tapia Granados (2012). While the former study reports upon the absence of cointegration between measures of infant survival and income, the latter paper discounts the presence of cointegration between and life expectancy and income. In summary, the literature exploring the presence of long-run relationships contains conflicting results and criticism.

The present study is motivated by the above study of Tapia Granados (2015), hereafter referred to as TG. The specific and important issue revisited is TG's conclusion that the analysis of 160 years of data on life expectancy and gross domestic product (GDP) does '*not provide any evidence of cointegration*' (p.690). This paper questions the robustness or reliability of this 'no long-run relationship' conclusion. A particular feature of our analysis is the consideration of potential structural change. When analysing a long span (i.e. calendar period) of data, an empirical complication arises as a result of the increased possibility of capturing incidents of structural change, or 'breaks', within the sample considered. Like the research of Arora (2001), Bishai (1995) and Brenner (2005), TG employs a long span of data. This is problematic as such breaks can impact upon the properties of the time series econometrics

* Corresponding author.
E-mail addresses: r.a.chowdhury@swan.ac.uk (R. Chowdhury), s.cook@swan.ac.uk (S. Cook), duncan.watson@uea.ac.uk (D. Watson).

https://doi.org/10.1016/j.socscimed.2023.116094
Received 16 March 2023; Received in revised form 12 July 2023; Accepted 13 July 2023
Available online 19 July 2023

methods employed and result in incorrect inferences being drawn. Two issues make this possibility particularly pertinent for the analysis undertaken by TG. First, the cointegration methods employed by TG are based upon prior unit root testing to establish the orders of integration of the individual series considered. Unfortunately, it has long been recognised that unit root testing in the presence of breaks can result in the generation of spurious inferences. In particular, when considering whether the unit root null should be rejected against an alternative hypothesis of stationarity, the Dickey-Fuller test (Dickey and Fuller, 1979) employed by TG can generate misleading results in 'both possible directions': a stationary series can be deemed to be a unit root process (see Perron, 1989); and a unit root process can appear stationary (see Leybourne et al., 1998). In addition, Campos et al. (1996) provides a cautionary note on the implications of breaks for cointegration analysis itself, promoting the use of an approach based upon the application of error correction models. Second, the sample period examined by TG includes numerous key historical events. It is obviously possible that conflicts such as World Wars or the 1918–1919 Influenza Pandemic may impact significantly on data characteristics and cause a break or breaks in the series examined, particularly the life expectancy series. As will be shown later, the existence of a break in the life expectancy data corresponding to the Influenza Pandemic is indeed apparent from simple visual inspection of this series. Given the impact of breaks upon the reliability of inferences and the presence of an apparent break, the long-run relationship considered by TG must therefore be re-examined. It is the consideration of structural change and its analysis via a method allowing potential breakpoints to be determined endogenously by the data, rather than imposed exogenously, that constitutes the first major contribution of our research to the health-income literature. However, this leads to further extension of previous research via the use of split-sample analysis employing autoregressive distributed lag (ARDL) modelling to offer robustness to uncertainty regarding the integrated nature of the data examined. It is argued that the explicit recognition of structural change, along with the utilisation of split-sample analysis and ARDL modelling, extends an existing literature containing mixed results. Considering the specific work of TG, the more sophisticated analysis adopted in the current paper reverses the 'no long-run relationship' conclusion presented in this earlier research.

In addition to offering developments in relation to the specific work of TG and the more general health-income literature, the present paper also adds to the burgeoning literature on the reproduction and replication of previous empirical research. While perhaps initially associated with a replication crisis in psychology and medicine (see, inter alia, Wiggins and Christopherson, 2019; Rodgers and Collings, 2021), this issue has now gained prominence more generally (see, inter alia, Open Science Collaboration, 2015; NASEM, 2019; Tol, 2019; Chin et al., 2023). Our findings reinforce the importance of undertaking replication exercises, highlighting how original conclusions can be critically reversed and policy implications drastically changed.

To achieve its objectives, this paper proceeds as follows. In the following section, to explain our approach to replication, we provide a brief overview of this area of research. Section 3 then presents and discusses the initial data to be examined in our study, providing results from univariate analysis. The use of 'initial' here is important as we later examine the most recent available vintage of the GDP series to consider further the robustness of our inferences. Importantly, Section 3 confirms the detection of a statistically significant break in 1918 for both the life expectancy and GDP series, prompting the introduction of split-sample analysis about this date. Following split-sample examination of the univariate properties of the life expectancy and GDP series, Section 4 then considers their potential long-run relationship. In Section 5 the robustness of the identified long-run relationship between life expectancy and income is investigated further by extending the analysis beyond the 'initial' data considered by TG. Section 6 provides some concluding remarks concerning the findings of our analysis, their policy implications and potential lines of future research.

2. Replication or reproduction?

The reproducibility of research findings has gained traction in both academic spheres and the popular media. While a 'replication crisis' is perhaps most closely associated with psychology and medicine (see, inter alia, Wiggins and Christopherson, 2019; Rodgers and Collings, 2021), discussion of these issues spans various disciplines. As an example, see the longstanding interest within criminology with studies ranging from Cook and Zarkin (1986) to Chin et al. (2023). Similarly, for economics, the special edition of de Marchi and Gilbert (1989), where numerous earlier empirical analyses are reconsidered to explore their robustness in light of econometric developments, provides an example of an early interest in reproducibility and replicability. This work pre-dates the more explicitly reproduction and replication focussed collections of studies in the *American Economic Review* in 2017 (see, inter alia, Berry et al., 2017; Duvendack et al., 2017) and *Energy Economics* in 2019 (see Tol, 2019). Further evidence of the increased importance of the reproducibility of research is illustrated by the emergence of data archives and dedicated sections within journals on replication (e.g. the Replication Network and the *Journal of Applied Econometrics*).

This emergent literature has also prompted research into the definition of 'replication' and the alternative terminology that is utilised (see Clemens, 2017; Machery, 2021; Nosek and Errington, 2020). This focus on terminology is apparent from the above discussion in this paper where the terms 'reproduction' and 'replication' both appear. To ensure transparency, we draw upon the stance of NASEM (2019). Therefore, 'reproduction' is taken to refer to the exact reproduction of results presented in empirical research using the same data and same method (s), while we take replication to involve the re-examination of a conclusion using alternative methods and/or data.

An important consideration is therefore whether, when revisiting TG's finding of no long-run relationship between life expectancy and income, we adopt an approach based upon replication or reproduction. The former approach is followed here for three reasons. First, despite obtaining data from the same sources, it is not possible to guarantee that exactly the same data as that employed by TG are examined. As will be discussed in the following section, this is due to uncertainty regarding the exact sample span of the data employed by TG. Second, the means of analysing these data differs from that of TG. As initial examination of the data confirms the existence of a break in both series considered, this prompts the analysis of two subsamples rather than just the full sample available. Third, given the findings of the univariate analysis undertaken, an alternative empirical method is employed which has the required property of being robust in the presence of uncertainty regarding the integrated nature of the series examined. In short, the analysis takes the form of a replication rather than reproduction for unavoidable and deliberate reasons as a result of data and robustness issues respectively.

3. Data and univariate properties

We utilise the following data series: annual measures of life expectancy at birth (LEB) for England and Wales; and real per capita gross domestic product for the United Kingdom (GDP). The LEB series was obtained from the Human Mortality Database (https://www.mortality.org/), while the GDP series was obtained from Maddison (2003). Although we employ the same sources as TG, we need to address some important matching issues that will inform our approach to the replication process. First, we should confirm start and end dates for the sample. TG states multiple dates here, referring both to 1840 and 1841 as start dates and 1999 and 2000 as end dates. As the earliest observation for LEB available to us is 1841, this is taken as the start date for the present analysis. Given key tables in TG (Tables 2 and 3) refer to 1999 as an end date, this is taken to be the final period for the 'whole sample'. Second, we should note issues concerning the 'geographical coverage' of the series. While TG refers to the GDP series being for Britain when introduced (p.689) and later for the UK (p.691), the latter is correct. However, in contrast to this, the LEB series relates to England and Wales. The differing levels of aggregation for the GDP and LEB series should be noted and might be a source of empirical bias. Third, the definition of the GDP series needs to be recognised. In this paper, real per capita gross domestic product is employed as the GDP variable as a 'per capita' series is arguably a superior measure not only in terms of standard economic applications but specifically when examining the health-income relationship. To support this latter argument reference can be made to, inter alia, the work of Acemoglu and Johnson (2007) where a detected negative relationship between health and economic growth is supported by reference to a reduction in steady state income due to an increased population under neoclassical growth theory. This importance of 'population' further supports our use of per capita GDP.

Figure One provides detail of the GDP and LEB series considered in this paper. From inspection of this figure, the upward trending nature of the two series is immediately apparent. When considering data on life expectancy and GDP, TG suggest that the possibility of a long-run relationship between life expectancy and GDP is unlikely on the basis of visual inspection of the series. In particular, TG refers to cointegration not being expected due to the 'shapes' (p.690) of plots of GDP and LEB, with the convexity of GDP noted. It is important to reflect upon this issue. It can be argued that discounting potential cointegration purely on the basis of visual inspection is problematic. Fig. 1 could be interpreted as providing evidence of the series moving apart slightly in the first part of the sample, before closing to intersect ahead of then pulling apart again, and finally drawing together in the final part of the sample. This could then be argued to reflect movements about an underlying attractor that indicate the *presence*, rather than *absence*, of cointegration. Difficulty in assessing the presence of cointegration visually is further complicated by the alternative forms it might take such as, inter alia, stochastic and deterministic cointegration (see Ogaki and Park, 1997) and asymmetric cointegration (Enders and Siklos, 2001). However, there is a more important issue to consider when reflecting upon potential cointegration for the present replication. When undertaking empirical analysis of the life expectancy and GDP data, TG considered these series in their natural logarithmic, rather than raw, forms. Such an approach follows conventional practice and is adopted in the present analysis. However, converting the series to their natural logarithmic forms has a linearising effect. The natural logarithmic forms of LEB and GDP are denoted here as *leb* and *gdp* respectively and presented in Fig. 2. From inspection of this graph the previous arguments regarding different 'shapes' and convexity can be discounted, with similarity between *leb* and *gdp* very apparent. In summary, the visual inspection which led TG to discount the possibility of cointegration was based upon the consideration of raw data rather than the logged data to be employed in the empirical analysis.

To examine the univariate properties of *leb* and *gdp*, we begin our analysis with application of the augmented Dickey-Fuller (ADF) unit root test (Dickey and Fuller, 1979, 1981) and GLS-based ADF (ADF-GLS) unit root test (Elliott et al., 1996) to the series over the full sample. Given the trending nature of the series, our unit root testing employs intercept and linear trend terms as deterministics to mitigate potential spurious inferences. Denoting the series under examination as y_t, the testing equation employed for the ADF test is specified as given in (1) below:

$$\Delta y_t = \alpha + \beta t + \varphi\, y_{t-1} + \sum_{j=1}^{p} \lambda_j \Delta y_{t-j} + v_t \tag{1}$$

where the unit root null hypothesis $(H_0 : \varphi = 0)$ is tested against an alternative hypothesis of asymptotic stationarity $(H_1 : \varphi < 0)$ via the test statistic $\tau_\varphi = \frac{\widehat{\varphi}}{s.e.(\widehat{\varphi})}$. The analogous testing equation for the ADF-GLS test is specified as given in (2):

$$\Delta y_t^d = \gamma\, y_{t-1}^d + \sum_{j=1}^{p} \delta_j \Delta y_{t-j}^d + e_t \tag{2}$$

where the null and alternative hypotheses are given as $H_0 : \gamma = 0$ and $H_1 : \gamma < 0$, the test statistic employed is given as $\tau_\gamma = \frac{\widehat{\gamma}}{s.e.(\widehat{\gamma})}$ and the series y_t^d is the 'GLS detrended' version of y_t obtained from application of the quasi-differencing and regression-based detrending approach presented in Elliott et al. (1996).

With regard to the degree of our augmentation of the testing equations as given by the value of p, this is determined via use of the modified

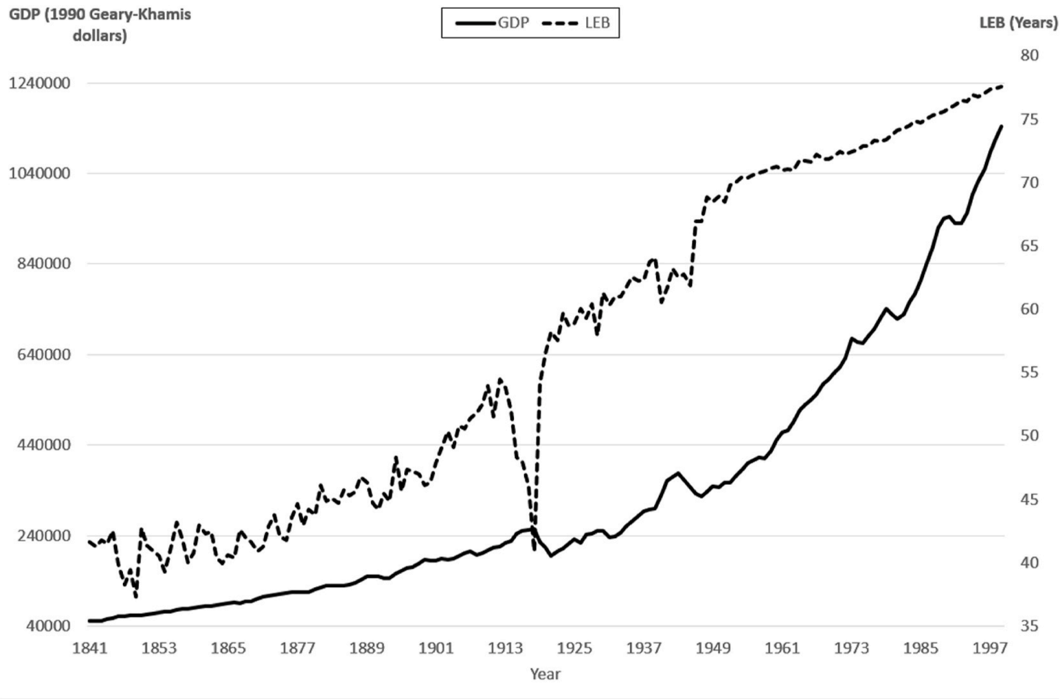

Fig. 1. Income and life expectancy at birth.

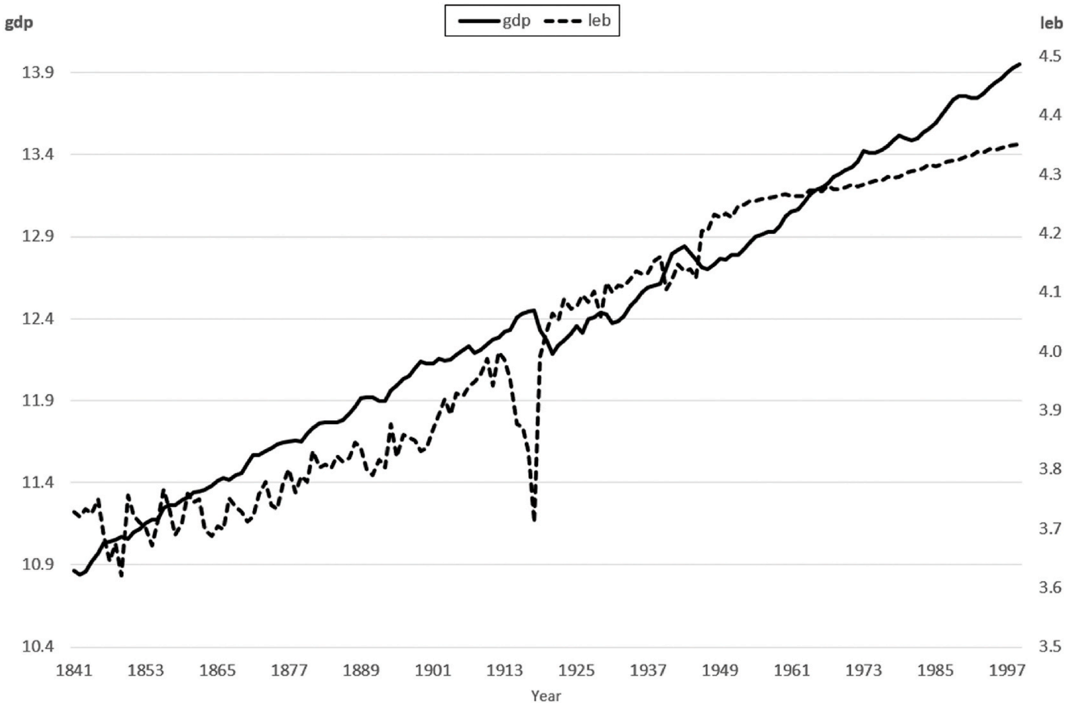

Fig. 2. Income and Life Expectancy at Birth (in logarithmic form).

Akaike Information Criterion (MAIC) with the maximum lag length considered given by the sample-based Schwert (1989) rule. For the 1841–1999 sample considered, this results in consideration of lag lengths from 0 to 13, with the optimised value *p* being that which returns the lowest value of the MAIC for the testing equations estimated. To increase the robustness of our analysis, the ADF-GLS test is employed in the 'OLS-GLS' form proposed by Perron and Qu (2007). Under this approach, OLS detrended data are employed to determine the degree of augmentation of the ADF-GLS test, while GLS detrended data are employed for inferences (see Perron and Qu, 2007; Sephton, 2022). The results obtained from application of the ADF and ADF-GLS tests are reported in Table 1. Inspection of the p-values obtained shows that the ADF and ADF-GLS tests do not reject the unit root null hypothesis for either series at conventionally considered levels of significance.

Before concluding that *leb* and *gdp* are unit root processes, the impact of structural change upon unit root tests should be recognised. Following Perron (1989), it has long been recognised that otherwise stationary series can be mis-classified as unit root processes on the basis of the application of an ADF test if they are subject to structural change or exhibit a 'break'. Alternatively expressed, the findings of Perron (1989) demonstrate that the presence a neglected break in a series increases the probability of the ADF test failing to correctly reject the unit root hypothesis. The subsequent research of Leybourne et al. (1998) has further emphasised the impact of breaks by demonstrating that a neglected break in a unit root process can cause spurious rejection of the unit root hypothesis by the ADF test. While the findings of Leybourne et al. (1998) show the impact of a break to depend not only upon its size but also its nature (whether it is a break in level or drift) and the point at which it occurs in the sample considered, they demonstrate that a neglected break can cause the ADF test to mistakenly classify a unit root process as stationary. The results of these seminal studies forcefully demonstrate the importance of considering potential structural change when examining the unit root hypothesis due its impact upon both the power (Perron, 1989) and size (Leybourne et al., 1998) of the ADF test. Further to this, it has been shown that misleading inferences can be drawn using other unit root tests in the presence of structural change (see, inter alia, Cook and Manning, 2004, 2005). These findings are particularly relevant for the present analysis as perhaps the most striking feature of Fig. 2 is the apparent break in *leb* coinciding with the 1918–1919 Influenza Pandemic. In an analysis of this pandemic, Pearce et al. (2011) identify three 'waves' relating to the periods June 1918–September 1918, September 1918–January 1919 and February 1919–May 1919 respectively. It is noted that the second of these was the most severe in terms of mortality, dwarfing the first and third waves in terms of weekly death rates (see Pearce et al., 2011, Fig. 1 p.90). It appears that this second wave is reflected in the *leb* series. To re-examine the initial 'unit root inference' in the presence of potential structural change, the unit root test of Perron (1997) is adopted using the innovational outlier model. The test is employed with endogenous determination of intercept and trend break dates via use of the minimum *t*-test criterion, with the MAIC again employed to select the degree of augmentation of the testing equations. The relevant testing equation employed is therefore given by (3) below:

$$y_t = \mu + \beta t + \theta\, DU_t(T_B) + \pi\, DT_t(T_B) + \psi\, D_t(T_B) + \rho\, y_{t-1} + \sum_{j=1}^{p} \vartheta_j \Delta y_{t-j} + u_t \tag{3}$$

where the unit root null hypothesis $(H_0 : \rho = 1)$ is tested against an alternative hypothesis of asymptotic stationarity $(H_1 : |\rho| < 1)$ via the test statistic $\tau_\rho = \frac{\widehat{\rho}}{s.e.(\widehat{\rho})}$. The breakpoint (T_B) is employed to create the

Table 1
Unit root test results (full sample analysis).

Series	ADF	ADF-GLS	Perron	
			Test stat	TB
gdp	– 1.654 [0.767]	– 1.657 [0.516]	– 6.310 [<0.01]	1918
leb	– 2.097 [0.543]	– 1.212 [0.789]	– 6.173 [<0.01]	1918

Notes: The above tabulated negative figures are calculated test statistics for the ADF, ADF-GLS and Perron tests. Figures in square brackets are p-values for the relevant test statistics, with the remaining figures under 'TB' denoting calculated break dates.

three breaks in (3) according to:

$$DU_t(T_B) = 1 \text{ if } t \geq T_B, \text{zero otherwise} \tag{4}$$

$$DT_t(T_B) = t - T_B + 1 \text{ if } t \geq T_B, \text{zero otherwise} \tag{5}$$

$$D_t(T_B) = 1 \text{ if } t = T_B, \text{zero otherwise} \tag{6}$$

Therefore, expressions (4) to (6) denote an intercept break, trend break and one-off dummy break respectively. The results obtained from application of the Perron test are presented in Table 1. In contrast to the findings obtained from application of 'no-break' unit root tests, the unit root null can now be seen to be overwhelming rejected for both the *leb* and *gdp* series. In both cases a highly significant break is detected for 1918 with the p-values associated with tests of significance of the intercept, trend and dummy break coefficients having p-values of {0.000, 0.084, 0.000} and {0.000, 0.001, 0.012} for *leb* and *gdp* respectively.

The results obtained identify a clear break in the series considered which matches the break that is apparent visually in Figs. 1 and 2, particularly for *leb*. However, to explore the possibility that the series might possess two breaks, the two-break Lee and Strazicich (2003) unit root test, hereafter referred to as the LS test, is considered in its 'Model C' form with two breaks in both intercept and trend. Following this seminal research, we employ the t-statistic rule to determine the degree of augmentation of the underlying testing equation, again using the Schwert rule to determine the maximum possible lag length. While full details of the mechanics of the LS test can be obtained from Lee and Strazicich (2003), a key issue here is whether this test identifies two significant breaks in *leb*. To examine this, the significance of the identified breaks resulting from application of the test to *leb* can be considered. Due to its design, the LS test will return two breaks and for this application they correspond to 1920 and 1945 respectively. However, what is important is whether these breaks are statistically significant. Consulting the relevant t-ratios of 4.908 and −0.351 for the coefficients on these breaks, it is clear that while the first break is significant, the second is highly insignificant. Our failure to detect two breaks when applying the LS test is obviously not unique in empirical research. For example, following its introduction in Lee and Strazicich (2003), a subsequent application of the LS test in Strazicich et al. (2004) found a subset of series examined possessed only a single significant break. In summary, the findings obtained from the LS test demonstrate that a single break is apparent in *leb* but that the date identified does not fit with the break apparent from visual inspection of the series. Obvious potential explanations for this apparent 'misdating' include the attempt to detect a second break when this is not present and the general issue that the date is determined via optimisation (minimisation) of the unit root test statistic rather than an explicit breakpoint dating procedure. However, the application of the LS test does nonetheless provide important information that supports the earlier findings obtained using the Perron (1997) test. With a calculated test statistic of −7.305, the LS test clearly rejects unit root hypothesis beyond the 1% level of significance. In addition to agreeing with the 'no unit root' inference drawn from application of the Perron (1997) test, this rejection provides evidence on the robustness of the earlier findings. More precisely, the research of Kim et al. (2000) has shown that spurious rejection of the unit root hypothesis can occur under application of a single-break test if a second break is present but ignored. The results of the LS test counter this potential accusation for the present analysis as it does not identify the presence of a significant second break. Application of the LS test to *gdp* resulted in similar findings in the sense that while the unit root hypothesis was rejected at the 10% level of significance and almost at the 5% level of significance (the calculated test statistic and 5% critical value being −5.671 and −5.772 respectively), the coefficients attached to the intercept and break dummies were not all significant at even the 10% level. As a consequence of the Perron test providing overwhelming evidence of a single break that exactly matches the break overwhelmingly apparent from visual inspection of the series and the LS test not detecting a significant second break, we proceed on the basis of the presence of a single significance break in both series occurring in 1918.

Given the detection of a significant break in both series in 1918, a split-sample analysis is undertaken with unit root tests applied to two subsamples for 1841–1917 and 1918–1999. Following the previously adopted approach, an intercept and linear trend are employed as the deterministic terms in the underlying testing equation and the degree of augmentation is determined using the Schwert rule and minimisation of the MAIC. The results obtained from this analysis are reported in Table 2. The ADF test results for 1841–1917 provide some evidence of *leb* and *gdp* both being stationary, with rejection of the unit root null hypothesis occurring beyond the 5% level for *leb* and almost the 5% level for *gdp*. Although the ADF-GLS test does not reject the null at standard levels of significance for *leb* (p-value = 25.9%), the null is rejected at the 5% level for *gdp* under this test. Although the ADF-GLS test has been shown to be more powerful that the ADF test in a number of circumstances and hence might be considered to generate increased rejection in empirical application, it has been recognised that there is no uniformly most powerful unit root test (see, inter alia, Müller and Elliott, 2003). Similarly, it has been shown that the finite-sample power of the ADF-GLS test can be influenced by a variety of issues, including initial conditions as considered by Müller and Elliott (2003). As a result, the findings from 'standard' (no break) unit root tests will be considered on balance. Adopting such an approach results in some uncertainty regarding the orders of integration of both series.

Considering the results for the second subsample at conventionally employed levels of significance, there is a single rejection of the unit root null resulting from application of the ADF test to *gdp*. As the results from the application of the ADF-GLS test statistic to *gdp* do not result in rejection, it could be that the ADF test result has been influenced by the 'initial condition' issue examined by Müller and Elliott (2003). Here, if the first observation of a series deviates from its underlying deterministic components, this can increase rejection of the null by the ADF test.

Given the overall findings from the application of the ADF and ADF-GLS unit root tests for the two subsamples, there is some evidence that a change in order of integration occurs for the series, or at least the classification of the series as either (asymptotically) stationary or unit root processes is difficult to confirm. These findings are important it is well recognised that care has to be exercised when considering unit root processes as a spurious relationship can be detected when none exists (see, Granger and Newbold, 1974; Phillips, 1986). This prompts our subsequent application of autoregressive distributed lag (ARDL) modelling to consider the potential long-run relationship between these series due to the robustness of this approach in the presence of uncertainty regarding the integrated nature of the series considered.

4. The long-run relationship between life expectancy and income

While not providing full details on the cointegration analysis undertaken, TG does report on the application of the cointegrating

Table 2
Unit root test results (split-sample analysis).

Series	1841 – 1917		1918 – 1999	
	ADF	ADF-GLS	ADF	ADF-GLS
gdp	− 3.370 [0.063]	− 2.932 [0.044]	− 3.849 [0.019]	− 1.530 [0.643]
leb	− 4.094 [0.010]	− 2.162 [0.259]	− 2.333 [0.411]	− 1.920 [0.384]

Notes: The above tabulated negative figures are calculated test statistics for the ADF and ADF-GLS tests. Figures in square brackets are p-values for the relevant test statistics.

regression Durbin-Watson test and the Johansen method not leading to a rejection of the null of no cointegration. This conclusion of no cointegration is revisited here, with the decision on the approach adopted to perform this analysis guided by the insights gained from the univariate analysis of the previous section.

The results presented in the previous section present uncertainty regarding the order of integration of the series examined, with a potential change apparent when moving between subsamples and some conflicting results found within the subsamples. These findings prompt investigation of a potential long-run relationship between *leb* and *gdp* based upon split-sample analysis, together with the use of an empirical method applicable irrespective of whether the series under examination are unit root or stationary processes. The method employed is the ARDL-based bounds testing approach of Pesaran et al. (2001). Application of this method begins with the formulation of an appropriate ARDL model of the series which in this bivariate context simply involves regressing a variable upon its own lags, appropriate deterministic terms, and the current value and lags of another variable. This ARDL model is then respecified as a conditional error correction model (CECM). As with the unit root testing conducted in the previous section, decisions are therefore required concerning the deterministic terms to employ in this analysis and lag augmentation. With regard to deterministic terms, the trending nature of the series considered leads to inclusion of a linear trend in our analysis. Following the approach considered in seminal research of Pesaran et al. (2001), we consider the inclusion of a trend in both restricted and unrestricted forms. This results in consideration of CECMs referred to as Cases IV and V respectively by Pesaran et al. (2001). These specifications are presented in equations (7) and (8) below.

$$\Delta gdp_t = c_0 + \pi_g\left(gdp_{t-1} - \gamma_y t\right) + \pi_l(leb_{t-1} - \gamma_x t) + w\Delta leb_{t-j} + \sum_{i=1}^{p-1} \delta_i \Delta gdp_{t-i} + \sum_{j=1}^{q-1} \gamma_j \Delta leb_{t-j} + u_t \tag{7}$$

$$\Delta gdp_t = c_0 + c_1 t + \pi_g gdp_{t-1} + \pi_l leb_{t-1} + w\Delta leb_{t-j} + \sum_{i=1}^{p-1} \delta_i \Delta gdp_{t-i} + \sum_{j=1}^{q-1} \gamma_j \Delta leb_{t-j} + u_t \tag{8}$$

The next issue to consider is the degree of augmentation of the ARDL models underlying the CECMs. Our consideration of subsamples containing 77 and 82 observations results in a need to be mindful of the degrees of freedom implications of higher lag orders. We therefore consider ARDL (p, q) models where $1 \leq p, q \leq 4$. While this lower bound is chosen to ensure the presence of lagged level terms, the upper bound is sufficiently large to capture dynamics given the annual frequency of the data examined. The optimised values of p and q employed are determined via minimisation of the AIC.

The existence of a long-run relationship between *leb* and *gdp* can be tested in two ways. First, it can be examined via an F-like statistic examining the joint insignificance of the coefficients on the lagged levels terms in (7) and (8). Formally, this involves examination of the null hypothesis $H_0 : \pi_g = \pi_l = 0$. However, for model (8) a second option is available via a t-like statistic based upon the null $H_0 : \pi_g = 0$. Given the F- and t-like structure of the tests and their use of models referred to as Cases IV and V in Pesaran et al. (2001), the resulting test statistics are denoted here as F_{IV}, F_V and t_V respectively. Critical values for these tests follow non-standard distributions, with the use of 'bounds' arising due to the presence of two critical values based upon the series examined being either stationary or unit root processes. Only when the relevant upper bound is exceeded is the null of no long-run relationship rejected. Finally, drawing upon the above models, the long-run relationship between *leb* and *gdp* is given by $\beta = -\pi_l/\pi_g$.

The results obtained from application of the ARDL approach to the two subsamples are reported in Table 3. From inspection of these results it can be seen that relatively low order ARDL models are selected for both subsamples considered with the Breusch-Godfrey test statistics indicating that these models do not exhibit any issues regarding serial correlation.

The calculated F_{IV}, F_V and t_V statistics presented in Table 3 provide overwhelming evidence of a long-run relationship between *leb* and *gdp* in both subsamples, with the relevant nulls rejected beyond the 1% level of significance in all but one case. The single exception to rejection at the 1% level occurs under application of the F_{IV} test in the first subsample. However, the failure to reject at this level is marginal as the calculated test statistic of 8.884 is compared to an upper bound 1% critical value of 8.905. Two interesting features can be commented upon in relation to the long-run coefficients, as given by the calculated β. First, it can be seen that the long-run coefficients are negative for both subsamples. This detection of a negative relationship between life expectancy and GDP is supported by previous research. While studies such as Kunze (2014) refer to the complexity of the life expectancy-GDP relationship, Boucekkine et al. (2002) and Echevarría (2004) both note the potential of a rise in 'higher levels' of life expectancy to have a negative impact economic growth. This negative relationship is also present in the research of Tapia Granados and Ionides (2015) and Hansen and Lønstrup (2015), while Acemoglu and Johnson (2007) make reference to neo-classical growth theory to support an identified negative relationship and Acemoglu and Johnson (2014) report directly upon 'a negative impact of life expectancy on GDP per capita' (p.1370). In attempt to explain this negative relationship, Ruhm (2000) discusses the roles of obesity, smoking, physical activity and diets as factors underlying the procyclical nature of mortality. A second feature of the results is the increased sensitivity of income to life expectancy in the latter subsample as demonstrated by the near doubling (in absolute terms) of the estimated values of β (-0.4244 compared to -0.8079). Again this finding is consistent with previous research in which the negative relationship between the series is noted as more apparent for the higher levels of income. Clearly, higher levels of income corresponds to the second, rather than first, subsample considered here. This finding is also apparent in the results of the panel data analysis of Acemoglu and Johnson (2014) where life expectancy coefficients in a regression of log per capita GDP are more negative for sample covering 1940–2000 compared to results for a 1940–1980 sample. Considering the size of the reported long-run coefficients in Table 3, these lie in a range reported for a series of panel data models presented by Acemoglu and Johnson (2014).

In summary, adopting an approach based upon split-sample analysis about an identified breakpoint shows the presence of long-run relationships between *leb* and *gdp*, and these differ in nature between the two periods considered.

Table 3
Long-run relationships between ***gdp*** and ***leb***.

Sample	ARDL	β	F_{IV}	t_V	F_V	BG
1841–1917	(2,2)	$-0.4244^{\dagger}$	8.884*	$-4.815^{\dagger}$	$12.928^{\dagger}$	0.286
1918–1999	(2,1)	$-0.8079^{\dagger}$	$9.011^{\dagger}$	$-4.532^{\dagger}$	$13.212^{\dagger}$	0.296

Notes: Figures under the heading 'ARDL' denote the dimensions of the ARDL models underlying the estimated conditional ECMs. Figures under the headings F_{IV} and F_V are calculated F-like bounds test statistics for Cases IV and V of Pesaran et al. (2001). t_V denotes the calculated t-like bounds test for Case V of Pesaran et al. (2001). The results under the heading 'BG' are p-values for a second order Breusch-Godfrey test of serial correlation. * and † denote significance at the 5% and 1% levels of significance respectively.

5. Exploring robustness using an alternative data vintage

To further examine the relationship between life expectancy and income, the above analysis is repeated using the most recent vintage of per capita GDP available from the Maddison Project (https://www.rug.nl/ggdc/historicaldevelopment/maddison/?lang=en), namely the 2020 release. Clearly, this is a subsequent vintage of the data unavailable to TG.

The issue of revisions to data, and the implications this has for econometric analysis, are prominent within a particular approach to modelling that has become known as the LSE or Hendry methodology (see, inter alia, Cook, 1999; Gilbert, 1986; Hendry et al., 1990; Mizon, 1995). Under this approach reference is made to the notion of the Data Generation Process (DGP) which gives rise to the data employed in empirical research. This is deemed to arise as a result of a measurement system being imposed upon an actual underlying economic mechanism. In simple terms, this can be interpreted as an uncontroversial statement proposing that: actual activity and actions take place, an attempt to record this requires the use of a particular measurement system, and this use of this measurement system results in data being generated. Therefore as the measurement system changes, the data obtained change or are revised. This has clear implications for the 'truth' of models and the ability of tests to evaluate specific issues or events as actual phenomena are not examined directly but rather via measures of them obtained using specific measurement systems that are subject to potential revision or change. By extending our analysis to consider a later vintage of data, we explore the impact of data revision upon our initial inferences.

Fig. 3 presents the raw forms of the 2020 release of per capita GDP and life expectancy. The similarity between this graph and Figure One demonstrates the similarity of the two vintages of per capita GDP, albeit that the series are measured in different units (the latter version being measured in 2011 US dollars rather than 1990 Geary-Khamis dollars). A further comparison of the two per capita GDP series in provided by Fig. 4 in which both series appear in their natural logarithmic forms. Before considering the relationship between the life expectancy and the 2020 vintage of GDP in their logged forms, these series are presented in Fig. 5. This final figure offers a more direct comparison of the two series and makes explicit their similarity.

The similarity between the two vintages of GDP discussed above is reflected in the results of unit root testing provided in Table 4. Here, the results are very similar to those for the earlier vintage of GDP reported in Tables One and Two. In summary, considering the 'no break' unit root tests, the unit root null hypothesis is not rejected over the full sample but there is some evidence against it in the second subsample and more still in the first. As with the earlier version of the GDP series, application of the Perron test results in the detection of a significant break in 1918 and rejection of the unit root hypothesis.

Table 5 presents the results obtained from examination of a potential long-run relationship between life expectancy and the new vintage of GDP. Again, the results provided are very similar to those obtained when considering the earlier vintage of reported in Table 3. In summary, highly significant evidence in support of a long-run relationship is apparent for both subsamples, with the long-run coefficients again showing life expectancy to have a greater negative impact upon income in the second subsample.

6. Conclusion

The existence of a large literature exploring the relationship between life expectancy and income is unsurprising given its potentially important policy implications. Our paper has revisited a very definite conclusion presented in relevant research which states UK life expectancy and income do not share a long-run relationship. At the heart of our analysis is the detection of clear structural breaks in both the life expectancy and income data examined. These breaks correspond to the peak of the 1918–1919 Influenza Pandemic in the UK. This discovery shapes the empirical approach that we subsequently adopt. Prompting the utilisation of a split-sample analysis, it discloses uncertainty regarding the orders of integration of the series examined thus encouraging the adoption of ARDL modelling. The central outcome of this analysis is the reversal of TG's conclusion of no long-run relationship between life expectancy and income. However, our empirical analysis provides further important findings. First, the significant long-run relationship detected is found to be negative. Second, it is shown that the relationship was more negative in the second of the two samples

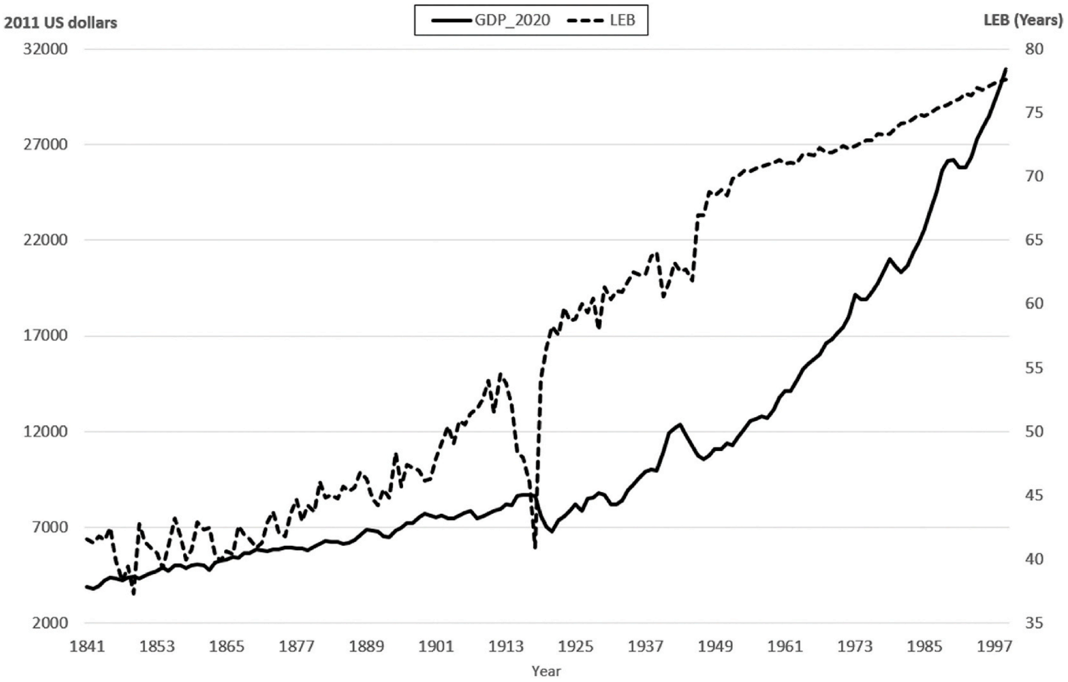

Fig. 3. 2020 income and life expectancy.

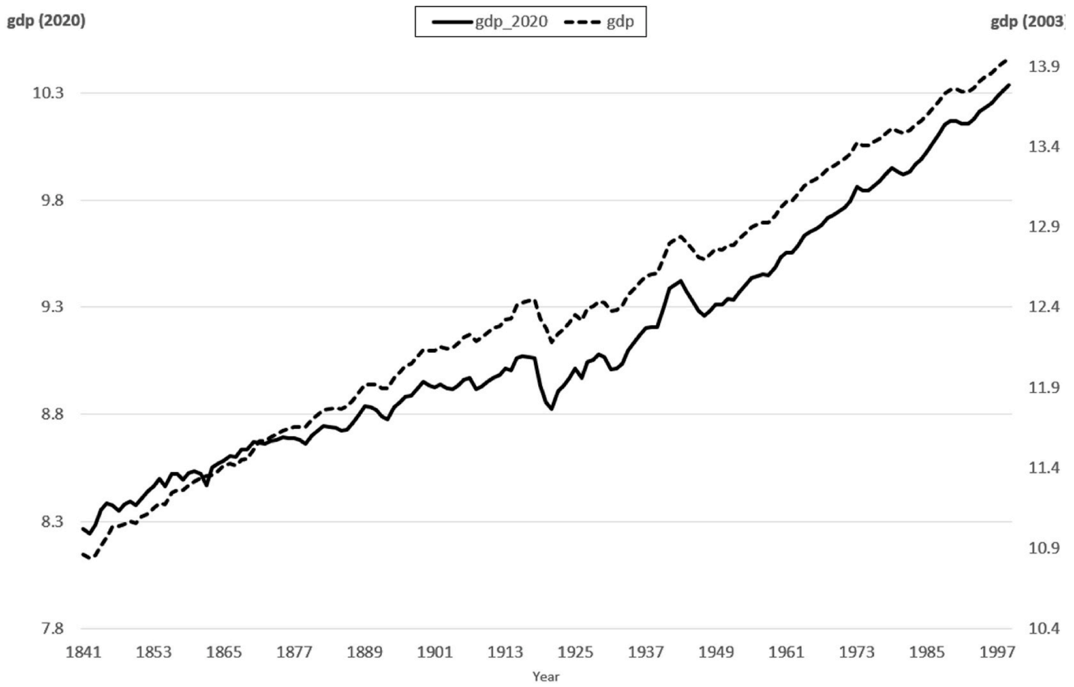

Fig. 4. Two vintages of GDP.

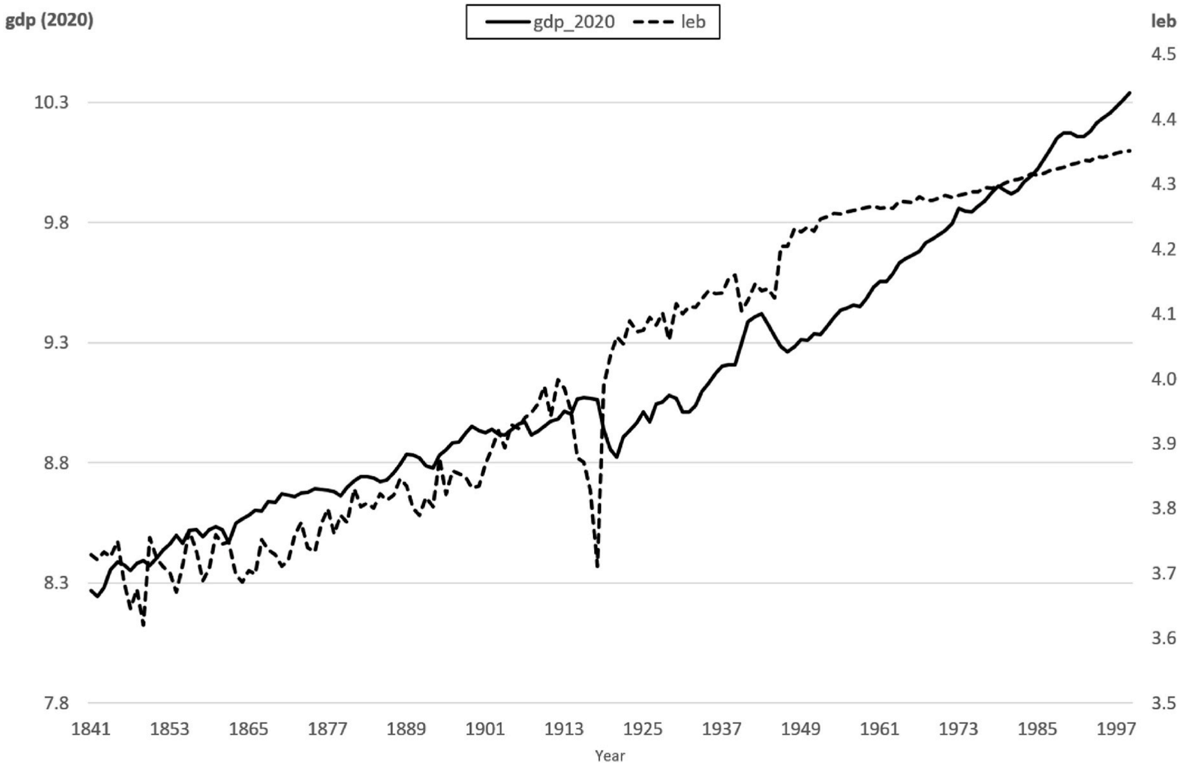

Fig. 5. 2020 Income and Life Expectancy at Birth (in logarithmic form).

considered. As noted, both of these findings have support in previous research. A further feature of our analysis is the demonstration of robustness when extended to consider a later vintage of income data unavailable to TG.

The empirical findings presented have clear policy implications. While consideration of the relationship between life expectancy and income might typically focus upon socio-economic factors such as income inequalities, education and housing (see, inter alia, Purnell et al., 2016), the negative relationship revealed in our analysis of a ‘higher income, higher life expectancy’ economy directs attention to additional issues. Prominent amongst these are the factors explored in studies such as Ruhm (2000) relating to obesity, diet, activity and smoking. While we have deliberately adopted analysis at a highly aggregated level to highlight the value of replication and reproduction, our findings

Table 4
Unit Root Test Results for 2020 vintage ***gdp***.

	ADF	ADF-GLS	Perron	
			Test stat	TB
1841–1999	– 0.067 [0.995]	– 0.771 [0.943]	– 5.772 [<0.01]	1918
1841–1917	– 4.028 [0.012]	– 3.226 [0.020]		
1918–1999	– 3.584 [0.038]	– 1.147 [0.856]		

Notes: The above tabulated negative figures are calculated test statistics for the ADF, ADF-GLS and Perron tests. Figures in square brackets are p-values for the relevant test statistics, with the remaining figures under 'TB' denoting calculated break dates.

Table 5
Long-run Relationships between the 2020 vintage of ***gdp*** and ***leb***.

Sample	ARDL	β	F_{IV}	t_V	F_V	BG
1841–1917	(2,2)	−0.3830*	9.377†	−5.059†	13.660†	0.930
1918–1999	(2,1)	−0.7498†	8.973†	−4.512†	13.239†	0.123

Notes: Figures under the heading 'ARDL' denote the dimensions of the ARDL models underlying the estimated conditional ECMs. Figures under the headings F_{IV} and F_V are calculated F-like bounds test statistics for Cases IV and V of Pesaran et al. (2001). t_V denotes the calculated t-like bounds test for Case V of Pesaran et al. (2001). The results under the heading 'BG' are p-values for a second order Breusch-Godfrey test of serial correlation. * and † denote significance at the 5% and 1% levels of significance respectively.

demonstrate that a shift in focus to a more disaggregated level might be fruitful for future research to explore the importance income inequality and the life style factors which are not captured by the consideration of aggregate GDP.

Data availability

The data are available (with ease) to interested readers via sources stated in the paper, but cannot be supplied by the authors.

Acknowledgements

We are very grateful to the editor and two anonymous referees for comments which have improved both the content and presentation of our work.

References

Acemoglu, D., Johnson, S., 2007. Disease and development: the effect of life expectancy on economic growth. J. Polit. Econ. 115, 925–985.

Acemoglu, D., Johnson, S., 2014. Disease and development: a reply to Bloom, canning, and fink. J. Polit. Econ. 122, 1367–1375.

Arora, S., 2001. Health, human productivity, and long-term economic growth. J. Econ. Hist. 61, 699–749.

Berry, J., Coffman, L., Hanley, D., Gihleb, R., Wilson, A., 2017. Assessing the rate of replication in Economics. Am. Econ. Rev. 107, 27–31.

Bishai, D., 1995. Infant mortality time series are random walks with drift: are they cointegrated with socioeconomic variables? Health Econ. 4, 157–167.

Bloom, D., Canning, D., 2007. Commentary: the Preston Curve 30 years on: still sparking fires. Int. J. Epidemiol. 36, 498–499.

Bloom, D., Canning, D., Fink, G., 2014. Disease and development revisited. J. Polit. Econ. 122, 1355–1366.

Boucekkine, R., de la Croix, D., Licandro, O., 2002. Vintage human capital, demographic trends and endogenous growth. J. Econ. Theor. 104, 340–375.

Brenner, M., 2005. Commentary: economic growth is the basis of mortality rate decline in the 20th century- Experience of the United States 1901-2000. Int. J. Epidemiol. 34, 1214–1221.

Campos, J., Ericsson, N., Hendry, D., 1996. Cointegration tests in the presence of structural breaks. J. Econom. 70, 187–220.

Catalano, R., Bruckner, T., 2016. Clear and simple: No association between the Great Recession and period life expectancy at birth in the USA. Int. J. Epidemiol. 45, 1681–1683.

Catalano, R., Goldman-Mellor, S., Saxton, K., Margerison-Zilko, C., Subbaraman, M., LeWinn, K., Anderson, E., 2011. The health effects of economic decline. Annu. Rev. Publ. Health 32, 431–450.

Chin, J.M., Pickett, J., Vazire, S., Holcombe, A., 2023. Questionable research practices and open science in quantitative criminology. J. Quant. Criminol. 39, 21–51.

Clemens, M., 2017. The meaning of failed replications: a review and proposal. J. Econ. Surv. 31, 326–342.

Cook, P., Zarkin, G., 1986. Homicide and economic conditions: a replication and critique of M. Harvey Brenner's new report to the U.S. Congress. J. Quant. Criminol. 2, 69–80.

Cook, S., 1999. Methodological aspects of the encompassing principle. J. Econ. Methodol. 6, 61–78.

Cook, S., Manning, N., 2004. The disappointing properties of GLS-Based unit root tests in the presence of structural breaks. Commun. Stat. Simulat. Comput. 33, 585–596.

Cook, S., Manning, N., 2005. Unobserved heterogeneity in Markovian analysis of the size distortion of unit root tests. J. Stat. Comput. Simulat. 75, 709–729.

de Marchi, N., Gilbert, C., 1989. The history and methodology of econometrics. Oxf. Econ. Pap. 41, 1–11.

Dickey, D., Fuller, W., 1979. Distribution of the estimators for autoregressive time series with a unit root. J. Am. Stat. Assoc. 74, 427–431.

Dickey, D., Fuller, W., 1981. Likelihood ratio statistics for autoregressive time series with a unit root. Econometrica 49, 1057–1072.

Duvendack, M., Palmer-Jones, R., Reed, W., 2017. What is meant by "replication" and why does it encounter resistance in economics? Am. Econ. Rev. 107, 46–51.

Elliott, G., Rothenberg, T., Stock, J., 1996. Efficient tests for an autoregressive unit root. Econometrica 64, 813–836.

Enders, W., Siklos, P., 2001. Cointegration and threshold adjustment. J. Bus. Econ. Stat. 19, 166–176.

Echevarría, C., 2004. Life expectancy, schooling time, retirement, and growth. Econ. Inq. 42, 602–617.

Gilbert, C., 1986. Professor Hendry's econometric methodology. Oxf. Bull. Econ. Stat. 48, 283–307.

Granger, C., Newbold, P., 1974. Spurious regressions in econometrics. J. Econom. 2, 111–120.

Hansen, C., Lønstrup, L., 2015. The rise in life expectancy and economic growth in the 20th century. Econ. J. 125, 838–852.

Hendry, D., Juselius, K., 2000. Explaining cointegration analysis: Part I. Energy J. 21, 1–42.

Hendry, D., Juselius, K., 2001. Explaining cointegration analysis: Part II. Energy J. 22, 75–120.

Hendry, D., Leamer, E., Poirer, D., 1990. A conversation on econometric methodology. Econom. Theor. 6, 171–261.

Kim, T., Leybourne, S., Newbold, P., 2000. Spurious rejections by Perron tests in the presence of a misplaced or second break. Oxf. Bull. Econ. Stat. 62, 433–444.

Kunze, L., 2014. Life expectancy and economic growth. J. Macroecon. 39, 54–65.

Lee, J., Strazicich, M., 2003. Minimum Lagrange Multiplier unit root test with two structural breaks. Rev. Econ. Stat. 85, 1082–1089.

Leybourne, S., Mills, T., Newbold, P., 1998. Spurious rejections by Dickey–Fuller tests in the presence of a break under the null. J. Econom. 87, 191–203.

Machery, E., 2021. What is a replication? Philos. Sci. 87, 545–567.

Mizon, G., 1995. Progressive modelling of macroeconomic time series: the LSE methodology. In: Hoover, K. (Ed.), Macroeconometrics: Developments, Tensions and Prospects. Kluwer Academic Press, Dordrecht.

Müller, U., Elliott, G., 2003. Tests for unit roots and the initial condition. Econometrica 71, 1269–1286.

National Academies of Sciences, Engineering and Medicine (NASEM), 2019. Reproducibility and Replicability in Science [online]. The National Academies Press, Washington, DC. https://doi.org/10.17226/25303.

Nosek, B., Errington, T., 2020. What is replication? PLoS Biol. 183, e3000691.

Open Science Collaboration, 2015. Estimating the reproducibility of psychological science. Science 349, aac4716. https://doi.org/10.1126/science.aac4716.

Ogaki, M., Park, J., 1997. A cointegration approach to estimating preference parameters. J. Econom. 82, 107–134.

Pearce, D., Pallaghy, P., McCaw, J., McVernon, J., Mathews, J., 2011. Understanding mortality in the 1918–1919 influenza pandemic in England and Wales. Influenza Other Respir. Virus. 5, 89–98.

Perron, P., 1989. The Great Crash, the oil price shock, and the unit root hypothesis. Econometrica 57, 1361–1401.

Perron, P., 1997. Further evidence on breaking trend functions in macroeconomic variables. J. Econom. 80, 355–385.

Perron, P., Qu, Z., 2007. A simple modification to improve the finite sample properties of Ng and Perron's unit root tests. Econ. Lett. 94, 12–19.

Pesaran, M., Shin, Y., Smith, R., 2001. Bounds testing approaches to the analysis of level relationships. J. Appl. Econom. 16, 289–326.

Prados de la Escosura, L., 2023. Health, income, and the Preston curve: a long view. Econ. Hum. Biol. 48, 101212.

Phillips, P., 1986. Understanding spurious regressions in econometrics. J. Econom. 33, 311–340.

Preston, S., 1975. The changing relation between mortality and level of economic development. Popul. Stud. 29, 231–248.

Purnell, J., Simon, S., Zimmerman, E., Camberos, G., Fields, R., 2016. Policy implications of social determinants of health. In: Eyler, A., Chriqui, J. (Eds.), Prevention, Policy, and Public Health. Oxford Academic, New York.

Rodgers, P., Collings, A., 2021. Reproducibility in cancer biology: what have we learned? Elife 10, e75830.

Ruhm, C., 2000. Are recessions good for your health? Q. J. Econ. 115, 617–650.

Schwert, G., 1989. Tests for unit roots: a Monte Carlo investigation. J. Bus. Econ. Stat. 7, 147–159.

Sephton, P., 2022. Finite sample lag adjusted critical values of the ADF-GLS test. Comput. Econ. 59, 177–183.

Strazicich, M., Lee, J., Day, E., 2004. Are incomes converging among OECD Countries? Time series evidence with two structural breaks. J. Macroecon. 26, 131–145.

Tapia Granados, J., 2005. Response: on economic growth, business fluctuations, and health progress. Int. J. Epidemiol. 34, 1226–1233.

Tapia Granados, J., 2012. Economic growth and health progress in England and Wales: 160 years of a changing relation. Soc. Sci. Med. 74, 688–695.

Tapia Granados, J., 2015. Commentary: william Ogburn, Dorothy Thomas and the influence of recessions and expansions on mortality. Int. J. Epidemiol. 44, 1484–1490.

Tapia Granados, J., Ionides, E., 2015. Statistical evidence shows that mortality tends to fall during recessions: a rebuttal to Catalano and Bruckner. Int. J. Epidemiol. 45, 1683–1686.

Tol, R., 2019. Special issue on replication. Energy Econ. 82, 1–3.

Wiggins, B., Christopherson, C., 2019. The replication crisis in psychology: an overview for theoretical and philosophical psychology. J. Theor. Phil. Psychol. 39, 202–217.