The impact of biomedical innovation on U.S. mortality, 1999-2019: evidence partly based on 286 million descriptors of 27 million PubMed articles

Frank R. Lichtenberg
Columbia University.
National Bureau of Economic Research,
and CESifo
frl1@columbia.edu

Kriste Krstovski
Columbia University
kriste.krstovski@columbia.edu

9 December 2024

The impact of biomedical innovation on U.S. mortality, 1999-2019: evidence partly based on 286 million descriptors of 27 million PubMed articles

Abstract

We investigate econometrically the overall impact that biomedical innovation had on premature mortality in the U.S. during the period 1999-2019. We use a difference-in-differences research design: we investigate whether the diseases for which there was more biomedical innovation had larger reductions in premature mortality. Biomedical innovation related to a disease is measured by the change in the mean vintage of descriptors of PubMed articles about the disease. We analyze data on 286 million descriptors of 27 million articles about over 800 diseases.

Our estimates indicate that premature mortality from a disease is significantly inversely related to the lagged vintage of descriptors of articles about the disease. For example, the number of years of potential life lost before age 75 due to a disease is significantly inversely related to the mean vintage of descriptors 2-16 years earlier; it is most strongly inversely related to the vintage of descriptors 12 years earlier. This lag is not surprising: we show that the number of prescriptions for a drug is significantly positively related to the number of PubMed descriptors of the drug 6-14 years earlier. It is most strongly related to the number of PubMed descriptors of the drug 10 years earlier.

To explore whether biomedical innovation had different effects on the mortality of whites and blacks, we also estimate models using race-specific mortality data. Our estimates indicate that biomedical innovation reduced the mortality of white people sooner than it reduced the mortality of black people, and that the mortality of black people was not reduced by the most recent innovations.

From 1999 to 2019, the age-adjusted mortality rate (excluding deaths from unintentional injury, suicide, and homicide) from all diseases declined by approximately 20%--about 1% per year. Our estimates imply that, in the absence of biomedical innovation, age-adjusted mortality rates would not have declined. Also, from 1975 to 2019, the age-adjusted cancer mortality rate declined by approximately 27%. Our estimates also imply that, in the absence of biomedical innovation, the age-adjusted cancer mortality rate would not have declined. Some factors other than biomedical innovation (e.g., a decline in the smoking rate and an increase in educational attainment) probably contributed to the decline in mortality. But other factors (e.g., a rise in obesity and the prevalence of chronic conditions) undoubtedly contributed to an increase in mortality.

I. Introduction

Several previous studies have examined whether longevity has a positive effect on per capita income. Acemoglu and Johnson (2007) concluded that "there is no evidence that the large increase in life expectancy raised income per capita" (p. 925). However, Bloom et al (2014) found that "in a more general empirical framework in which we include initial life expectancy, the Acemoglu and Johnson results are reversed, with both levels and changes in health displaying positive associations with" per capita income growth. Kuhn et al (2024) also found there to be a highly significant positive effect of longevity on per capita income.

Even if longevity growth has no effect on per capita income growth, it has a positive effect on economic growth (or "human development"), broadly defined. Health and longevity are valued for their own sake, not merely because they may increase per capita income. Nordhaus (2003) argued that "to a first approximation, the economic value of increases in longevity in the last hundred years is about as large as the value of measured growth in non-health goods and services." Life expectancy at birth is one of the three components of the United Nations' Human Development Index (United Nations Development Programme (2024)), a summary measure of average achievement in key dimensions of human development.[1]

Many leading economists believe that economic growth is primarily driven by technological progress, which is generated by R&D investment. Romer (1990, p. S72) argued that "technological change…lies at the heart of economic growth." Jones (1998, pp. 89-90) argued that "technological progress is driven by research and development (R&D) in the advanced world." Jones (2002) presented a model in which long-run growth is driven by the discovery (via research effort) of new ideas throughout the world. His model built upon a large collection of previous research, including Grossman and Helpman (1991) and Aghion and Howitt (1992), as well as earlier contributions by Phelps (1966), Shell (1966), Nordhaus (1969), and Simon (1986). Cutler, Deaton and Lleras-Muney (2006) concluded that "knowledge, science, and technology are the keys to any coherent explanation" of mortality.

In this study, we propose to investigate econometrically the overall impact that biomedical innovation had on premature mortality in the U.S. during the period 1999-2019. We

[1] The other two components are per capita income and educational attainment.

will use a difference-in-differences research design: we will investigate whether the diseases for which there was more biomedical innovation had larger reductions in premature mortality (e.g., the number of years of potential life lost before age 75). Biomedical innovation related to a disease will be measured by the change in the mean vintage[2] of Medical Subject Headings (MeSH) descriptors in PubMed articles about the disease.[3]

Our methodology is similar to the one used in a previous study (Lichtenberg (2018)). The present study will build on the previous study in several important ways. First, the previous study only covered cancer. While cancer is certainly an important disease, it accounted for only 19% of years of potential life lost before age 75 in the U.S. in 2019. The present study will analyze data on all diseases. Second, the previous study did not examine whether biomedical innovation had different effects on the mortality of different racial groups. The present study will examine that issue. Third, the previous study analyzed cancer mortality during the period 1999-2013; the present study will analyze mortality from all diseases during the period 1999-2019. We will also reexamine cancer mortality during a longer period (1975-2019) and using an alternative mortality measure (the age-adjusted mortality rate).

In the next section, we explain how we constructed measures of biomedical innovation, by disease. In section III, we present our econometric model of mortality. Data sources and descriptive statistics are presented in section IV. Empirical results are presented in section V. Implications of the estimates are discussed in section VI. The final section provides a summary.

II. Measurement of biomedical innovation, by disease

Our measures of biomedical innovation, by disease, were constructed from data contained in PubMed, a literature database available to the public online since 1996, developed and maintained by the National Center for Biotechnology Information, part of the U.S. National Library of Medicine (NLM). The PubMed database contains more than 36 million citations and abstracts of biomedical literature. PubMed data can be downloaded as described on

[2] According to the Merriam-Webster dictionary, one meaning of the term vintage, and the one intended by us, is "the period of origin or manufacture."

[3] Kuhn et al (2024) used patent counts to measure biomedical innovation. They used data on total biomedical patents, not patents that pertain to specific diseases. Determining which disease(s) each biomedical patent pertains to would be extremely challenging. Also, the purpose of a patent is to *restrict access* to innovations, so an increase in patents will not necessarily reduce mortality.

the Download PubMed Data page. The PubMed database has three components: MEDLINE, PubMed Central (PMC), and Bookshelf.[4]

MEDLINE is the largest component of PubMed and consists primarily of citations from journals selected for MEDLINE. MEDLINE is NLM's premier bibliographic database that contains more than 31 million references to journal articles in life sciences with a concentration on biomedicine. Currently, there are citations from more than 5,200 worldwide journals in about 40 languages.

The subject scope of MEDLINE is biomedicine and health, broadly defined to encompass those areas of the life sciences, behavioral sciences, chemical sciences, and bioengineering needed by health professionals and others engaged in basic research and clinical care, public health, health policy development, or related educational activities. MEDLINE also covers life sciences vital to biomedical practitioners, researchers, and educators, including aspects of biology, environmental science, marine biology, plant and animal science as well as biophysics and chemistry. The majority of the publications in MEDLINE are scholarly journals; however, a small number of newspapers, magazines, and newsletters considered useful to particular segments of the NLM broad user community have historically been included.

MEDLINE is the online counterpart to the MEDical Literature Analysis and Retrieval System (MEDLARS) that originated in 1964. MEDLINE includes literature published from 1966 to present, and an OLDMEDLINE subset that has selected coverage of literature prior to that period. The OLDMEDLINE subset represents journal article citations from two print indexes: the 1946-1959 *Current List of Medical Literature (CLML)*, and the 1960-1965 *Cumulated Index Medicus (CIM)*. There are approximately 2,011,000 article citations from international biomedical journals that cover the fields of medicine, preclinical sciences and allied health sciences from 1946 through 1965.[5]

A distinctive feature of MEDLINE/PubMED is that the records are indexed with Medical Subject Headings (MeSH). MeSH is the NLM's controlled vocabulary thesaurus; the NLM says

---

[4] Citations for PubMed Central (PMC) articles make up the second largest component of PubMed. PMC is a full text archive that includes articles from journals reviewed and selected by NLM for archiving (current and historical), as well as individual articles collected for archiving in compliance with funder policies. The third and last component of PubMed is citations for books and some individual chapters available on Bookshelf. Bookshelf is a full text archive of books, reports, databases, and other documents related to biomedical, health, and life sciences.

[5] The OLDMEDLINE subset does not include citations from the *Quarterly Cumulative Index Medicus (QCIM)* print indexes. A hand search of the *QCIM* print indexes is necessary to ensure comprehensive review of medical periodical literature of the world from 1946 through 1956.

that MeSH "is one of the most highly sophisticated thesauri in existence today." It consists of sets of terms ("descriptors") in a hierarchical structure that permits searching at various levels of specificity.[6] There were 30,454 descriptors in 2023 MeSH. The MeSH Section staff continually revise and update the MeSH vocabulary.[7] Staff subject specialists are responsible for areas of the health sciences in which they have knowledge and expertise. In addition to receiving suggestions from indexers and others, the staff collect new terms as they appear in the scientific literature or in emerging areas of research; define these terms within the context of existing vocabulary; and recommend their addition to MeSH. Professionals in various disciplines are also consulted regarding broad organizational changes and close coordination is maintained with various specialized vocabularies. Between 2003 and 2023, the number of MeSH descriptors increased from 21.4 thousand to 30.5 thousand. On average, about 453 descriptors were added per year. The mean number of descriptors per article is 10.6; the 27 million PubMed articles published by 2019 have 286 million descriptors.

The NLM's MeSH Descriptor file and website indicates the year in which each MeSH descriptor was established.[8] However, a descriptor can appear in PubMed articles that were published many years before the descriptor was established. For example, as shown in Appendix Figure 1, the descriptor "female" (unique ID D005260) was established in 1966 but appears in PubMed articles as early as 1942; it appeared in 12,370 articles in 1964. In this study, we will define the vintage of a descriptor as the earliest publication year of PubMed articles in which the descriptor occurs.

Although the number of MeSH descriptors has increased substantially, the average frequency at which recent descriptors occur in PubMed is much lower than the average frequency at which earlier-vintage descriptors occur. Consequently, the mean vintage of descriptors is quite old, and has not increased much. As shown in Figure 1, half of the descriptors that ever appeared in PubMed first appeared after 1960, but only 17% of the (frequency-weighted) descriptor citations occurring in post-2015 publications were for descriptors that first appeared after 1960.

---

[6] Most Descriptors indicate the subject of an indexed item, such as a journal article, that is, what the article is about. https://www.nlm.nih.gov/mesh/intro_record_types.html. The MeSH "tree" can be explored here: https://meshb.nlm.nih.gov/treeView.

[7] https://www.nlm.nih.gov/pubs/factsheets/mesh.html

[8] The MEDLINE co-occurrences file summarizes the MeSH Descriptors that occur together in MEDLINE citations from the MEDLINE/PubMed Baseline (National Library of Medicine (2024)).

A substantial number of MeSH descriptors (in section C of the MeSH tree) are about diseases. Hence, we can identify articles that are about different diseases. By using the NLM's Unified Medical Language System, we can determine the ICD10 codes corresponding to MeSH disease descriptors. Table 1 shows the top 25 (out of 1028) 3-digit ICD10 diseases, ranked by number of PubMed articles. (Some articles may discuss several diseases.) Not surprisingly, highly prevalent diseases including breast neoplasms, HIV, hypertension, lung neoplasms, acute myocardial infarction, and diabetes are included in this list.

The measure of disease-specific biomedical innovation we will employ is the long-run change in the mean vintage of descriptors of articles about each disease. We hypothesize that, in general, diseases for which there were larger increases in vintage had smaller increases, or larger declines, in mortality.

The hypothesis that later vintage goods and services tend to be of higher quality than earlier vintage goods has been advocated by many economists since it was first formulated in the 1950s.[9] Johansen (1959) developed a theoretical model of vintage capital in which there are technological improvements in capital in later vintages. Intriligator (1992, p. S77) said that "the newer capital is more productive than the older capital as a result of technological improvements in the later vintages." Bresnahan and Gordon (1996) said that "new goods are at the heart of economic progress." As noted by Jovanovic and Yatsenko (2012), in "the Spence–Dixit–Stiglitz tradition…new goods [are] of higher quality than old goods." Bohlman, Golder, and Debanjan (2002, p. 1177) said that "technology improves over time…As technology advances, later entrants can utilize a more recent and efficient vintage of technology than an earlier entrant who has committed to older technology…vintage effects will benefit later entrants…We refer to 'vintage effects' as any technology shift that results in lower costs for a later entrant, enabling it to achieve higher product quality." In 1987, the Royal Swedish Academy of Sciences awarded the Alfred Nobel Memorial Prize in Economic Sciences to Robert Solow for his contributions to the theory of economic growth. The Academy cited Solow's 1960 article, *Investment and Technical Progress,* in which he presented a

> new method of studying the role played by capital formation in economic growth. His basic assumption was that technical progress is "built into" machines and other capital

[9] The earliest paper attempting to quantitatively assess the vintage effect in capital model was by Benhabib and Rustichini (1993). Boucekkine, de la Croix, and Licandro (2018) highlight the salient characteristics and implications of the seminal contributions in the field of vintage capital growth theory.

> goods and that this must be taken into account when making empirical measurements of the role played by capital. This idea then gave birth to the "vintage approach" (a similar idea was discussed by Leif Johansen in Norway at about the same time)…The most important aspect of Solow's article was not so much the empirical outcome, but the method of analyzing "vintage capital." Nowadays, the vintage capital concept has many other applications and is no longer solely employed in analyses of the factors underlying economic growth…The vintage approach has proved invaluable, both from the theoretical point of view and in applications such as the analysis of the development of industrial structures.

We measure the vintage of "ideas" (MeSH descriptors), not the vintage of physical capital. Ideas may or may not be embodied in physical capital. If technological change is "disembodied", it affects output growth independently of capital accumulation. In contrast, "embodied" technological change requires investment to affect output. The "embodiment" controversy between Jorgenson and Solow in the 1960s centered on the importance of capital-embodied technological change (Hercowitz (1998)).

III. Econometric model of mortality

To investigate econometrically the impact that biomedical innovation had on premature mortality in the U.S. during the period 1999-2019, we will estimate many versions of the following model:

$$\ln(\text{mortality}_{dt}) = \beta_k \, \text{vintage_measure}_{d,t-k} + \alpha_d + \delta_t + \varepsilon_{dt} \qquad (1)$$

where mortality_{dt} is one of the following variables:

ypll85_{dt} = the number of years of potential life lost before age 85 due to disease d in year t (t = 1999, 2000, …, 2019)

ypll75_{dt} = the number of years of potential life lost before age 75 due to disease d in year t

ypll65_{dt} = the number of years of potential life lost before age 65 due to disease d in year t[10]

and $\text{vintage_measure}_{d,t-k}$ is one of the following variables:

[10] The U.S. CDC's WISQARS Years of Potential Life Lost (YPLL) Report website permits one to specify age thresholds of 65, 70, 75, 80, and 85. The World Health Organization has used YPLL to measure disease burden in its Global Burden of Disease (GBD) and Global Health Estimates (GHE) reports for many years. In the 2010 GBD, the WHO used an age threshold of 86.01 years for all persons. In the current GHE, the WHO uses an age threshold of 91.93 years for all persons.

$\text{vint_mean}_{d,t-k}$ = the mean vintage (year of first appearance in PubMed) of descriptors of articles about disease d published in year t-k (k = 0, 2, …, 20)

$\text{post1990\%}_{d,t-k}$ = the fraction of descriptors of articles about disease d published in year t-k that first appeared in PubMed after 1990

$\text{vint_mean}_{d,t-k}$ was computed as follows:

$$\text{vint_mean}_{d,t-k} = \frac{\sum_a \sum_j \text{article_descriptor}_{aj} * \text{article_disease}_{ad} * \text{article_year}_{a,t-k} * \text{vintage}_j}{\sum_a \sum_j \text{article_descriptor}_{aj} * \text{article_disease}_{ad} * \text{article_year}_{a,t-k}}$$

where

$\text{article_descriptor}_{aj}$ = 1 if article a has descriptor j
= 0 if article a does not have descriptor j

$\text{article_disease}_{ad}$ = 1 if article a is about disease d
= 0 if article a is not about disease d

$\text{article_year}_{a,t-k}$ = 1 if article a was published in year t-k
= 0 if article a was not published in year t-k

vintage_j = vintage (first year of PubMed appearance) of descriptor j

$\text{post1990\%}_{d,t-k}$ was computed as follows:

$$\text{post1990\%}_{d,t-k} = \frac{\sum_a \sum_j \text{article_descriptor}_{aj} * \text{article_disease}_{ad} * \text{article_year}_{a,t-k} * \text{post1990}_j}{\sum_a \sum_j \text{article_descriptor}_{aj} * \text{article_disease}_{ad} * \text{article_year}_{a,t-k}}$$

where

post1990_j = 1 if $\text{vintage}_j > 1990$
= 0 if $\text{vintage}_j \leq 1990$

In eq. (1), α_d is a fixed effect for disease d, and δ_t is a fixed effect for year t. The year fixed effects (δ_t's) in eq. (1) control for the effects of changes in aggregate demographic and macroeconomic variables (e.g. population size and age structure, GDP, educational attainment), to the extent that those variables have similar effects on mortality from different diseases. Eq. (1) will be estimated by weighted least squares, weighting by ($\sum_t \text{mortality}_{dt}$). Disturbances will be clustered within diseases.

$\text{vint_mean}_{d,t-k}$ is our principal measure of vintage. But the data on the vintages of MeSH descriptors are clearly left-censored: only 2% of MeSH descriptors first appeared in PubMed

before 1945, and 40% first appeared during 1945-1947. Therefore, we will also estimate some versions of eq. (1) in which vintage_measure$_{d,t-k}$ = post1990%$_{d,t-k}$.

Estimates of eq. (1) will not capture cross-disease spillover effects: the potential effects of biomedical innovation for one disease (e.g., hypertension) on mortality from other diseases (e.g., acute cerebrovascular disease (stroke)). Although some spillover effects are adverse, others—perhaps most—are positive. For example, Prince et al (2007) argued that "mental disorders increase risk for communicable and non-communicable diseases, and contribute to unintentional and intentional injury. Conversely, many health conditions increase the risk for mental disorder, and comorbidity complicates help-seeking, diagnosis, and treatment, and influences prognosis." Also, the NIH National Institute on Aging (2024) says that "conditions such as diabetes, depression, and stroke may increase a person's risk for Mild Cognitive Impairment."

The relationship between biomedical innovation, as measured by the change in PubMed descriptor vintage, and the change in mortality is unlikely to be contemporaneous and may be subject to a substantial lag. Innovations may be discussed in biomedical literature several years before they are most frequently used, and improvement in outcomes may occur several years after changes in treatment (van de Glind et al (2016), Barter and Waters (2018)). For one important type of descriptors—drug descriptors—we can obtain insight into the timing of the relationship between the frequency of PubMed descriptors and the frequency of use of the drug. From the Medical Expenditure Panel Survey Prescribed Medicines files, we can obtain estimates of the number of U.S. outpatient prescriptions, by (Multum MediSource Lexicon) generic drug name and year (1996-2021). We can also compute the annual number of PubMed descriptors of each of these drugs.[11] Using these data, we estimated the following model, using data on about 600 drugs:

$$n_rx_{st} = \rho_k \, n_descriptors_{s,t-k} + \alpha_s + \delta_t + \varepsilon_{st} \qquad (2)$$

where

n_rx_{st} = the estimated number (in millions) of U.S. outpatient prescriptions for drug (chemical substance) s in year t (t = 1996, 1997, …, 2021)

$n_descriptors_{s,t-k}$ = the number of times the descriptor of drug s occurred in PubMed in year t-k (k = 0, 1, 2, …, 20)

[11] The NLM's Unified Medical Language System was used to link Multum MediSource Lexicon generic drug names to MeSH drug descriptors.

Estimates of ρ_k from eq. (2) are shown in Table 2 and plotted in Figure 2. Each estimate is from a separate regression. For $k \leq 5$, the estimates of ρ_k are not statistically significant: the number of prescriptions for a drug is not significantly related to the number of PubMed descriptors of the drug 0-5 years earlier. However, for $6 \leq k \leq 14$, the estimates of ρ_k are positive and statistically significant: the number of prescriptions for a drug is significantly positively related to the number of PubMed descriptors of the drug 6-14 years earlier. It is most strongly related to the number of PubMed descriptors of the drug 10 years earlier. The point estimate of ρ_{10} (.0036) indicates that one additional descriptor of a drug is associated with 3600 additional outpatient prescriptions for the drug 10 years later.

Considering this evidence, it would not be surprising if mortality from a disease is most strongly inversely related to the mean vintage of descriptors of articles about the disease years (e.g., 6-14 years) earlier.

As noted above, in addition to estimating eq. (1) using data on mortality of the entire population, we will estimate eq. (1) using data on mortality by race for two groups—whites and blacks—to explore whether biomedical innovation had different effects on the mortality of the two groups. Such disparities could occur because some groups may have greater access to innovations than other groups. Wang et al (2006) found significant racial disparities in use of drugs approved within the previous 5 years. Data from the 2016–2021 Medical Expenditure Panel Survey Prescribed Medicines Files indicate that the mean FDA approval year of outpatient antineoplastic drugs taken by black Americans was over 4 years earlier than the mean FDA approval year of outpatient antineoplastic drugs taken by white Americans (1970.2 vs. 1974.4). Lichtenberg (2024) found that the approval of one additional drug for a cancer site 6–10 years earlier reduced the black female mortality rate 58% as much as it reduced the white female mortality rate.

The change in mortality from a disease is likely to depend on the change in incidence of the disease as well as on biomedical innovation for the disease. The change in incidence might be correlated across diseases with biomedical innovation (the change in descriptor vintage).[12] Unfortunately, data on incidence, by disease and year, are not available for all diseases.

[12] If changes in incidence are positively correlated across diseases with changes in descriptor vintage (e.g., because the sudden outbreak of a disease induces more innovation related to the disease), failure to control for incidence would likely bias estimates of β_k towards zero.

However, annual 1975-2019 data on age-adjusted incidence rates[13] (and on age-adjusted mortality rates), by disease and year, are available for an important subset of diseases: 43 types of cancer (e.g., breast, colon, lung). These data enable us to determine if and how controlling for incidence affects estimates of the effect of biomedical innovation on mortality.

First, we examined the "simple" relationship between incidence and mortality by estimating the following equation:

$$\ln(\text{aa_mort_rate}_{dt}) = \gamma_j \ln(\text{aa_inc_rate}_{d,t-j}) + \alpha_d + \delta_t + \varepsilon_{dt} \quad (3)$$

where

aa_mort_rate_{dt} = the age-adjusted mortality rate from disease (cancer site) d in year t (t = 1975, 1976, …, 2019)

$\text{aa_inc_rate}_{d,t-j}$ = the age-adjusted incidence rate of disease (cancer site) d in year t-j

Eq. (3) was estimated by weighted least squares, weighting by ($\sum_t \text{aa_mort_rate}_{dt}$). Disturbances were clustered within cancer sites. The estimates are shown in Table 3. The growth in mortality is significantly positively correlated with the growth in incidence 0-8 years earlier; it is most strongly correlated with the contemporaneous growth in incidence. A 10% increase in the incidence rate in year t is associated with a 5.8% increase in the mortality rate in year t.

Next, we examined the "simple" relationship between incidence and descriptor vintage by estimating the following equation:

$$\text{vint_mean}_{d,t} = \gamma_j \ln(\text{aa_inc_rate}_{d,t-j}) + \alpha_d + \delta_t + \varepsilon_{dt} \quad (4)$$

Eq. (4) was estimated by weighted least squares, weighting by $\text{n_descriptors}_{d,t}$ = the total number of descriptors of articles about disease (cancer site) d in year t. Disturbances were clustered within cancer sites. The estimates are shown in Appendix Table 1. The change in mean descriptor vintage is not significantly related to either contemporaneous or lagged changes in incidence.

This suggests that controlling for incidence will not have a significant effect on estimates of the effect of biomedical innovation on mortality. This can be verified by estimating two

[13] A cancer incidence rate is the number of new cancers of a specific site/type occurring in a specified population during a year, usually expressed as the number of cancers per 100,000 population at risk. A cancer mortality rate is the number of deaths, with cancer as the underlying cause of death, occurring in a specified population during a year. Cancer mortality is usually expressed as the number of deaths due to cancer per 100,000 population.

versions of the following equation, one excluding and the other including $\ln(\text{aa_inc_rate}_{dt})$ as a regressor:[14]

$$\ln(\text{aa_mort_rate}_{dt}) = \beta_k \text{ vintage_measure}_{d,t-k} + \gamma_j \ln(\text{aa_inc_rate}_{dt}) + \alpha_d + \delta_t + \varepsilon_{dt} \quad (5)$$

Eq. (5) was estimated by weighted least squares, weighting by $(\sum_t \text{aa_mort_rate}_{ct})$. Disturbances were clustered within cancer sites. Estimates of both versions of eq. (5) are shown in Table 4. When we don't control for incidence, the estimates of β_k are negative and significant (p-value < .04) when $k \leq 2$, and negative and at least marginally significant (p-value < .08) when $k \leq 6$. When we control for incidence, the estimates of β_k are negative and significant (p-value < .05) for all values of k but one. This suggests that failure to control for incidence will not result in overestimation of the effect of biomedical innovation on mortality.

IV. Data sources and descriptive statistics

Mortality data (yll85_{dt}, yll75_{dt}, yll65_{dt}) for all diseases were computed from 1999-2019 Multiple Cause-of-Death Mortality Data files posted on the NBER website (National Bureau of Economic Research (2024)).

Vintage data ($\text{vintage_measure}_{d,t-k}$, $\text{post1990\%}_{d,t-k}$) and other attributes of PubMed articles ($\text{n_descriptors}_{s,t}$) were computed from the PubMed annual baseline files.

Prescription drug data. Estimates of the number of U.S. outpatient prescriptions, by (Multum MediSource Lexicon) generic drug name and year (1996-2021) (n_rx_{st}) were computed from Medical Expenditure Panel Survey Prescribed Medicines files (Agency for Healthcare Research and Quality (2024)).

Cancer incidence and mortality data (aa_mort_rate_{ct}, $\text{aa_inc_rate}_{c,t-j}$) for 1975-2019 were obtained from Cancer Query Systems (National Cancer Institute (2024)).

Mappings from MeSH descriptors to ICD10 codes and to Multum MediSource Lexicon generic drug names were computed from the Unified Medical Language System Concept Names and Sources (MRCONSO) File (National Library of Medicine (2024)).

Mappings from ICD10 codes to SEER cancer site codes were obtained from the SEER Cause of Death Recode 1969+ (National Cancer Institute (2024)).

[14] The equation includes $\ln(\text{aa_inc_rate}_{dt})$ because, as shown in Table 3, mortality growth is most strongly correlated with the contemporaneous growth in incidence.

Descriptive statistics. Aggregate data on years of potential life lost before ages 85, 75, and 65 in 1999 and 2019 are shown in Table 5. Between 1999 and 2019, YPLL before age 85 (YPLL85) increased by 15.0%, from 33.5 million to 38.6 million. When deaths from three external causes (unintentional injury, suicide, and homicide) are excluded, YPLL85 increased by 8.7%, from 28.0 million to 30.4 million. During that period, the population below age 85 increased, so the age-adjusted YPLL85 rate declined by 13.5%; excluding deaths from three external causes, the YPLL85 rate declined by 21.7%--about 1% per year. Excluding deaths from the three external causes, the age-adjusted rates of years of potential life lost before ages 75 and 65 (YPLL75 and YPLL65, respectively) declined by similar amounts: 20.1% and 19.8%, respectively.

Table 5 also shows the "age-adjusted population below age 85" which we define as the ratio of YPLL85 to the age-adjusted YPLL85 rate. The age-adjusted population below age 85 increased by 33.0% (38.9% when deaths from three external causes are excluded).

Aggregate data on PubMed descriptor vintage during 1979-2019 are shown in Appendix Table 2. The data shown are weighted means across diseases; diseases are weighted by total YPLL75 during 1999-2019. From 1979 to 1999, vint_mean increased by 3.0 years, from 1943.1 to 1946.1. From 1999 to 2019, vint_mean increased by about half that much, by 1.4 years. From 1979 to 1999, post1990% increased from 0.0% to 0.9%. From 1999 to 2019, post1990% increased from 0.9% to 3.5%.

Appendix Table 3 shows the 20 most frequently occurring MeSH descriptors that first appeared in PubMed after 1990.

Appendix Table 4 shows data on mortality and descriptor attributes for the top 25 diseases, ranked by total ypll75 during 1999-2019.

V. Empirical results

Estimates of many versions of eq. (1) are reported in Table 6 and plotted in Figure 3. Each estimate is from a separate equation. On the left side of the table and figure are estimates when the vintage measure is vint_mean; on the right side are estimates when the vintage measure is post1990%. In the table, estimates in bold are statistically significant (p-value < .05). In the figure, solid squares denote statistically significant estimates, the large solid squares denote the most significant estimates, and hollow squares denote statistically insignificant estimates.

Rows 1-11 of the table and panels A and B of the figure show estimates when the dependent variable is $\ln(YPLL85_{dt})$. When the vintage measure is vint_mean, the estimates are negative and statistically significant when $2 \leq k \leq 16$: YPLL85 is significantly inversely related to the mean vintage of descriptors 2-16 years earlier. It is most strongly inversely related to the vintage of descriptors 6 years earlier. The point estimate of β_6 when the vintage measure is vint_mean (-0.086) implies that a one-year increase in vintage is associated with an 8.3% reduction in YPLL85 6 years later. Further implications of the magnitudes of the estimates will be discussed in the next section. When the vintage measure is post1990%, the estimates are negative and statistically significant when $0 \leq k \leq 8$: YPLL85 is significantly inversely related to the fraction of post-1990 descriptors 0-8 years earlier. It is most strongly inversely related to the vintage of descriptors 2-4 years earlier.

Rows 12-22 of the table and panels C and D of the figure show estimates when the dependent variable is $\ln(YPLL75_{dt})$. When the vintage measure is vint_mean, the estimates are negative and statistically significant when $2 \leq k \leq 16$: YPLL75 is significantly inversely related to the mean vintage of descriptors 2-16 years earlier. It is most strongly inversely related to the vintage of descriptors 12 years earlier. When the vintage measure is post1990%, the estimates are negative and statistically significant when $0 \leq k \leq 20$: YPLL75 is significantly inversely related to the fraction of post-1990 descriptors 0-20 years earlier. It is most strongly inversely related to the vintage of descriptors 4 years earlier.

Rows 23-33 of the table and panels E and F of the figure show estimates when the dependent variable is $\ln(YPLL65_{dt})$. When the vintage measure is vint_mean, the estimates are negative and statistically significant when $6 \leq k \leq 16$: YPLL65 is significantly inversely related to the mean vintage of descriptors 6-16 years earlier. It is most strongly inversely related to the vintage of descriptors 12 years earlier. When the vintage measure is post1990%, the estimates are negative and statistically significant when $0 \leq k \leq 20$: YPLL65 is significantly inversely related to the fraction of post-1990 descriptors 0-20 years earlier. It is most strongly inversely related to the vintage of descriptors 14 years earlier.

As discussed above, to explore whether biomedical innovation had different effects on the mortality of whites and blacks, eq. (1) can be estimated using race-specific mortality data. (The same descriptor vintage data are used to estimate the white and black mortality equations.) Estimates of eq. (1) when $mortality_{dt} = YPLL75_{dt}$, by race, are shown in Table 7 and plotted in

Figure 4. In rows 1-11 of the table, the vintage measure is vint_mean. The estimates of β_k from the white mortality equation are negative and statistically significant when $2 \leq k \leq 16$: YPLL75 of whites is significantly inversely related to the mean vintage of descriptors 2-16 years earlier. It is most strongly inversely related to the vintage of descriptors 12 years earlier. The estimates of β_k from the black mortality equation are negative and statistically significant when $6 \leq k \leq 14$: YPLL75 of blacks is significantly inversely related to the mean vintage of descriptors 6-14 years earlier. It is most strongly inversely related to the vintage of descriptors 12 years earlier.

In rows 12-22 of the table, the vintage measure is post1990%. All 11 estimates of β_k from the white mortality equation are negative and statistically significant. None of the 11 estimates of β_k from the black mortality equation are statistically significant. These estimates indicate that biomedical innovation reduced the mortality of white people sooner than it reduced the mortality of black people, and that the mortality of black people was not reduced by the most recent innovations.

VI. Discussion

Estimates of two versions of eq. (1)—including and excluding $\text{vintage_measure}_{d,t-k}$ as a regressor—enable us to estimate how much biomedical innovation reduced age-adjusted mortality rates during the period 1999-2019. When $\text{vintage_measure}_{d,t-k}$ is included in eq. (1), δ_{2019} is an estimate of the 1999-2019 change in mortality, holding constant $\text{vintage_measure}_{d,t-k}$, i.e., in the *absence* of biomedical innovation. (δ_{1999} is normalized to zero.) When $\text{vintage_measure}_{d,t-k}$ is excluded from eq. (1), δ_{2019} is an estimate of the 1999-2019 change in mortality in the *presence* of biomedical innovation. Changes in age-adjusted mortality *rates* can be computed by subtracting the corresponding log-change in age-adjusted population shown in Table 5 (e.g., 0.329 for the population below age 85) from the δ_{2019} estimates.

These calculations are shown for each of the three mortality measures and both of the vintage measures in Figure 5. In each case, we use the lag for which the relationship is most significant. Panel A shows estimates of the growth in the age-adjusted YPLL85 rate in the presence and absence of the change in vint_mean 6 years earlier. The growth (log-change) in the presence of biomedical innovation is -0.20 and is significantly different from zero. The growth in the absence of biomedical innovation is 0.02 and is not significantly different from zero. We

cannot reject the null hypothesis that in the absence of biomedical innovation, as measured by the lagged change in vint_mean, the age-adjusted YPLL85 rate would not have declined during the period 1999-2019. The estimates for the other mortality and vintage measures are quite similar. The growth (ln change) in the presence of biomedical innovation is -0.19 to -0.16 and is significantly different from zero; the growth in the absence of biomedical innovation is not significantly different from zero.

Similarly, by estimating 3 versions of eq. (5), we can estimate how much biomedical innovation and changes in cancer incidence reduced the age-adjusted *cancer* mortality rate during the period 1975-2019. Those calculations are shown in Figure 6. In the presence of changes in cancer incidence and biomedical innovation, the log-change in the age-adjusted cancer mortality rate was -0.37. Only a small part of that decline was due to the decline in cancer incidence: controlling only for cancer incidence, the estimated log-change in the age-adjusted cancer mortality rate was -0.33, and is significantly different from zero. But when we control for $vint_mean_{d,t-6}$ as well as for $\ln(aa_inc_rate_{dt})$, the estimated log-change in the age-adjusted cancer mortality rate was 0.04, and is not significantly different from zero.

From 1999 to 2019, age-adjusted mortality rates (excluding deaths from unintentional injury, suicide, and homicide) declined by approximately 20%--about 1% per year. Our estimates imply that, in the absence of biomedical innovation, age-adjusted mortality rates would not have declined. Some factors other than biomedical innovation probably contributed to the decline in mortality. The adult cigarette smoking rate declined from 23.5% in 1999 to 13.7% in 2018 (American Lung Association (2024)). The fraction of adults who had at least a bachelor's degree increased from 23.9% in 1997 to 34.2% in 2017 (American Council on Eduction (2024)). However, other factors probably contributed to an increase in mortality. From 1999-2000 through 2017-March 2020, US obesity prevalence increased from 30.5% to 41.9%, and the prevalence of severe obesity increased from 4.7% to 9.2%. Obesity-related conditions include heart disease, stroke, type 2 diabetes and certain types of cancer. These are among the leading causes of preventable, premature death (Centers for Disease Control (2024)). And between 2007 and 2018, the prevalence of 14 out of 21 chronic conditions among male Medicare beneficiaries increased, and the prevalence of 16 out of 21 chronic conditions among female beneficiaries increased (Center for Medicare and Medicaid Services (2024)).

VII. Summary

We investigated econometrically the overall impact that biomedical innovation had on premature mortality in the U.S. during the period 1999-2019. We used a difference-in-differences research design: we investigated whether the diseases for which there was more biomedical innovation had larger reductions in premature mortality. Biomedical innovation related to a disease was measured by the change in the mean vintage of descriptors of PubMed articles about the disease. We analyzed data on 286 million descriptors of 27 million articles about over 800 diseases.

Our estimates indicated that premature mortality from a disease is significantly inversely related to the lagged vintage of descriptors of articles about the disease. For example, the number of years of potential life lost before age 75 due to a disease is significantly inversely related to the mean vintage of descriptors 2-16 years earlier; it is most strongly inversely related to the vintage of descriptors 12 years earlier. This lag is not surprising: we showed that the number of prescriptions for a drug is significantly positively related to the number of PubMed descriptors of the drug 6-14 years earlier. It is most strongly related to the number of PubMed descriptors of the drug 10 years earlier.

To explore whether biomedical innovation had different effects on the mortality of whites and blacks, we also estimated models using race-specific mortality data. Our estimates indicated that biomedical innovation reduced the mortality of white people sooner than it reduced the mortality of black people, and that the mortality of black people was not reduced by the most recent innovations.

From 1999 to 2019, the age-adjusted mortality rate (excluding deaths from unintentional injury, suicide, and homicide) from all diseases declined by approximately 20%--about 1% per year. Our estimates implied that, in the absence of biomedical innovation, age-adjusted mortality rates would not have declined. Also, from 1975 to 2019, the age-adjusted cancer mortality rate declined by approximately 27%. Our estimates also implied that, in the absence of biomedical innovation, the age-adjusted cancer mortality rate would not have declined. Some factors other than biomedical innovation (e.g., a decline in the smoking rate and an increase in educational attainment) probably contributed to the decline in mortality. But other factors (e.g., a rise in

obesity and the prevalence of chronic conditions) undoubtedly contributed to an increase in mortality.

Our estimates do not capture cross-disease spillover effects: the potential effects of biomedical innovation for one disease on mortality from other diseases. Although some spillover effects are adverse, others are positive.

Statements and declarations

- The authors did not receive any financial support for this research.
- The authors do not have any conflicts of interest.
- All data used in this research are publicly available.
- Ethical approval was not required.

References

Acemoglu D, Johnson S (2007). Disease and Development: The Effect of Life Expectancy on Economic Growth. *Journal of Political Economy* 115 (6): 925-985, December.

Agency for Healthcare Research and Quality (2024). Medical Expenditure Panel Survey Prescribed Medicines files.

Aghion P, Howitt P (1992). A Model of Growth through Creative Destruction. *Econometrica 60*(2):323–51, March.

American Council on Education (2024)). United States Population Trends and Educational Attainment.

American Lung Association (2024). Overall Tobacco Trends.

Barter PJ, Waters DD (2018). Variations in time to benefit among clinical trials of cholesterol-lowering drugs. *J Clin Lipidol*. 12(4):857-862.

Benhabib J, Rustichini A (1993). A vintage capital model of growth and investment: Theory and evidence. In Becker R, Boldrin M, Jones R, Thomson W (Eds.), *General Equilibrium, Growth and Trade*, pp. 248-300. Academic Press.

Bloom DE, Canning D, Fink G (2014). Disease and Development Revisited *Journal of Political Economy* 122(6): 1355-1366.

Bohlmann JD, Golder PN, Debanjan M (2002). Deconstructing the Pioneer's Advantage: Examining Vintage Effects and Consumer Valuations of Quality and Variety. *Management Science* 48(9): 1175-1195, September.

Boucekkine R, de la Croix D, Licandro O (2018). Vintage Capital, in *The New Palgrave Dictionary of Economics*, 2018 Edition (Macmillan Publishers Ltd.).

Bresnahan TF, Gordon RJ, eds. (1997). *The Economics of New Goods*, v. 58 Studies in income and wealth, Chicago: University of Chicago Press.

Center for Medicare and Medicaid Services (2024). Medicare Chronic Conditions.

Centers for Disease Control (2024). WISQARS Years of Potential Life Lost (YPLL) Report, 1981 – 2020.

Centers for Disease Control (2024). Adult Obesity Facts.

Cutler D, Deaton A, Lleras-Muney A (2006). The Determinants of Mortality. *Journal of Economic Perspectives* 20(3): 97-120, Summer.

Grossman GM, Helpman E (1991). *Innovation and growth in the global economy*. Cambridge, MA: MIT Press.

Hercowitz Z (1998). The embodiment controversy: a review essay. *Journal of Monetary Economics* 41: 217-224

Intriligator MD (1992). Productivity and the Embodiment of Technical Progress. *The Scandinavian Journal of Economics* 94 (Supplement, Proceedings of a Symposium on Productivity Concepts and Measurement Problems: Welfare, Quality and Productivity in the Service Industries): S75-S87.

Johansen L (1959). Substitution versus Fixed Production Coefficients in the Theory of Economic Growth: A Synthesis. *Econometrica* 27(2): 157-176, April.

Jones, CI (1998). *Introduction to Economic Growth*. New York: W.W. Norton.

Jones CI (2002). Sources of U.S. Economic Growth in a World of Ideas. *American Economic Review* 92 (1): 220-239, March.

Jovanovic B, Yatsenko Y (2012). Investment in vintage capital. *Journal of Economic Theory* 147(2): 551-569.

Kuhn M, Minniti A, Prettner K, Venturini F (2024). Medical Innovation, Life Expectancy, and Economic Growth, SSRN working paper, May 26.

Lichtenberg (2018). The long-run impact of new medical ideas on cancer survival and mortality, *Economics of Innovation and New Technology* 28(7): 722-740 (2019).

Lichtenberg F (2024). Racial and gender disparities in the effect of new drug approvals on U.S. cancer mortality. *Academia Medicine* 2024;1.

National Bureau of Economic Research (2024). Mortality Data - Vital Statistics NCHS Multiple Cause of Death Data.

National Cancer Institute (2024). SEER Cause of Death Recode 1969+.

National Cancer Institute (2024). Cancer Query Systems.

National Library of Medicine (2024). Download PubMed Data.

National Library of Medicine (2024). MEDLINE® database.

National Library of Medicine (2024). OLDMEDLINE Data.

National Library of Medicine (2024). MEDLINE Co-Occurrences (MRCOC) Files.

National Library of Medicine (2024). Unified Medical Language System Concept Names and Sources (MRCONSO) File.

NIH National Institute on Aging (2024). What Is Mild Cognitive Impairment?

Nordhaus WD (1969). An Economic Theory of Technological Change. *American Economic Review* (*Papers and Proceedings*) 59(2):18–28, May.

Nordhaus WD (2003). The Health of Nations: The Contribution of Improved Health to Living Standards. In *Measuring the Gains from Medical Research: An Economic Approach*, Edited by KM Murphy and RH Topel (Chicago: University of Chicago Press).

Phelps ES (1966). Models of Technical Progress and the Golden Rule of Research. *Review of Economic Studies* 33(2):133–45, April.

Prince M, Patel V, Saxena S, et al (2007). No health without mental health. *Lancet* 370(9590): 859–77.

Romer P (1990). Endogenous technological change. *Journal of Political Economy* 98(5).

Royal Swedish Academy of Sciences (1987). The Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel 1987.

Shell K (1966). Toward a Theory of Inventive Activity and Capital Accumulation. *American Economic Review* (*Papers and Proceedings*) 56(2):62–68, May.

Simon JL (1986). *Theory of population and economic growth*. New York: Blackwell.

Solow R (1960). Investment and Technological Progress. In: Arrow K, Karlin S, Suppes P (eds). *Mathematical methods in the social sciences, 1959*. Stanford, CA: Stanford University Press.

United Nations Development Programme (2024). Human Development Index.

van de Glind EM, Willems HC, Eslami S, Abu-Hanna A, Lems WF, Hooft L, de Rooij SE, Black DM, van Munster BC (2016). Estimating the Time to Benefit for Preventive Drugs with the Statistical Process Control Method: An Example with Alendronate. *Drugs Aging* 33(5):347-53.

Wang J, et al (2006). Disparities in access to essential new prescription drugs between non-Hispanic whites, non-Hispanic blacks, and Hispanic whites. *Med Care Res Rev.* 63(6):742–63.

Figure 1
Fraction of 2023 MeSH descriptors first appearing in PubMed by year (1940, 1950, ..., 2020): unweighted, and weighted by frequency in post-2015 PubMed articles

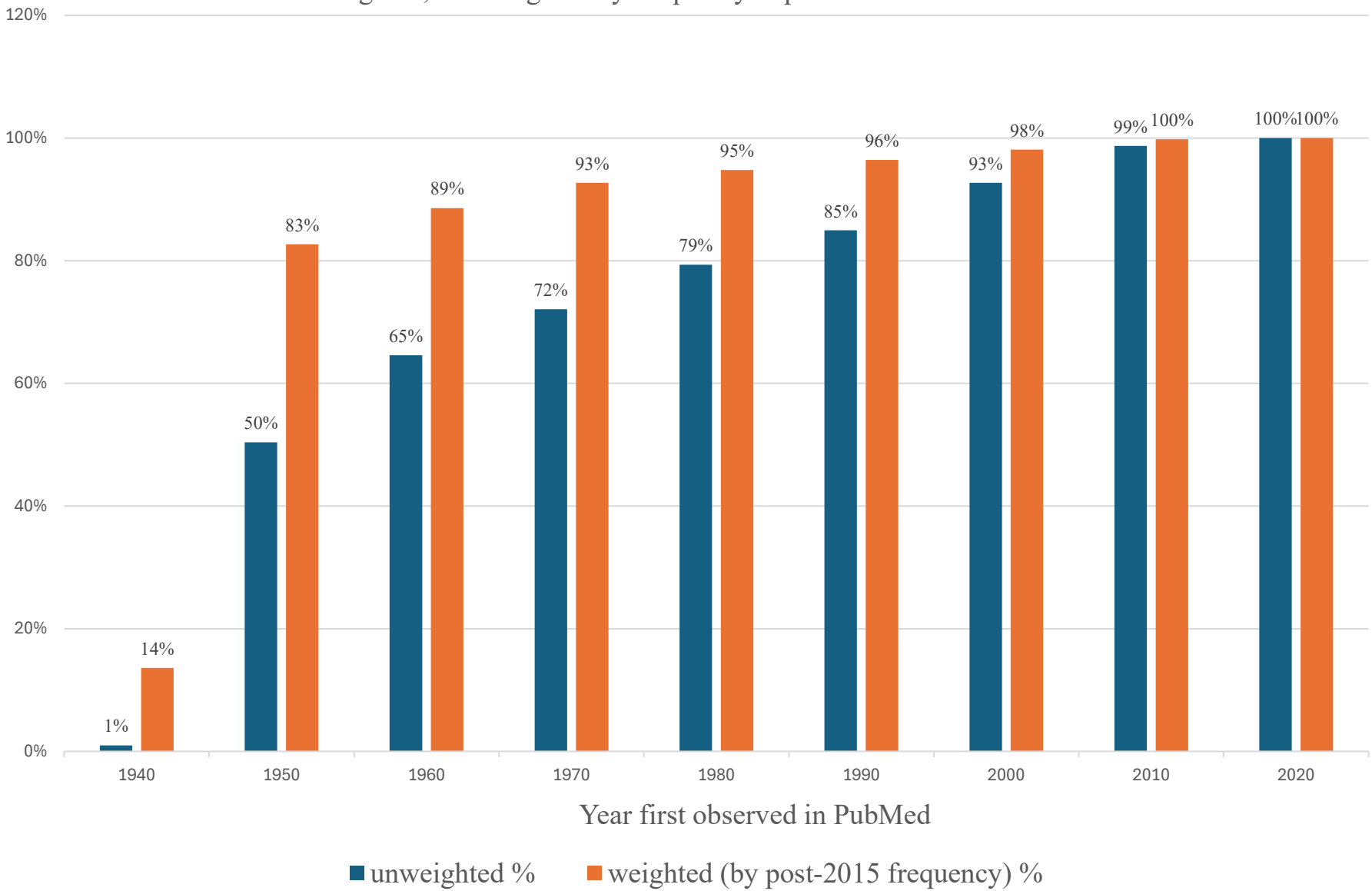
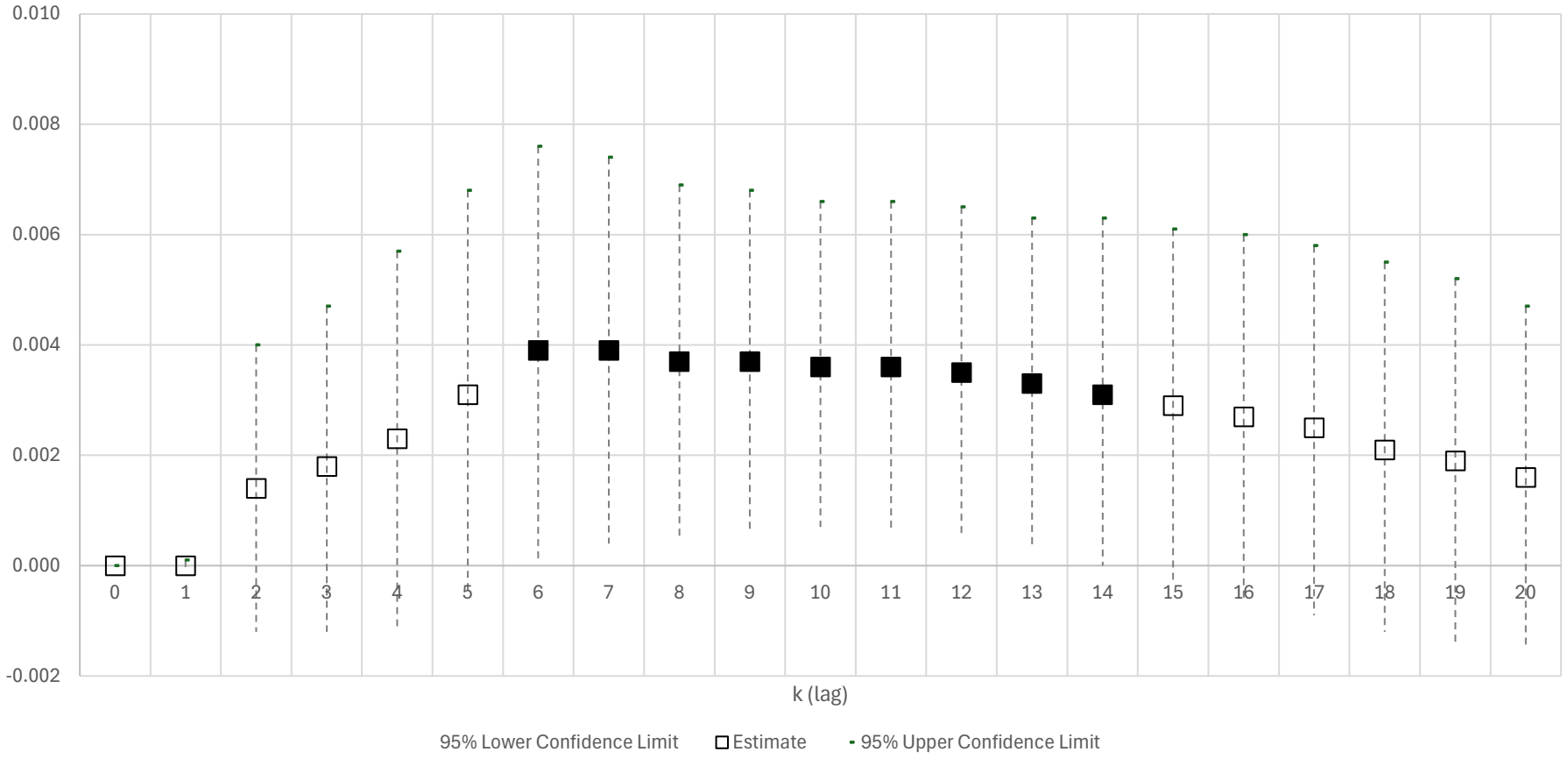


Figure 2

Estimates of ρ_k from eq. (2), based on data on about 600 drugs:

$$n_rx_{st} = \rho_k \, n_descriptors_{s,t-k} + \alpha_s + \delta_t + \varepsilon_{st}$$



Each estimate is from a separate regression.

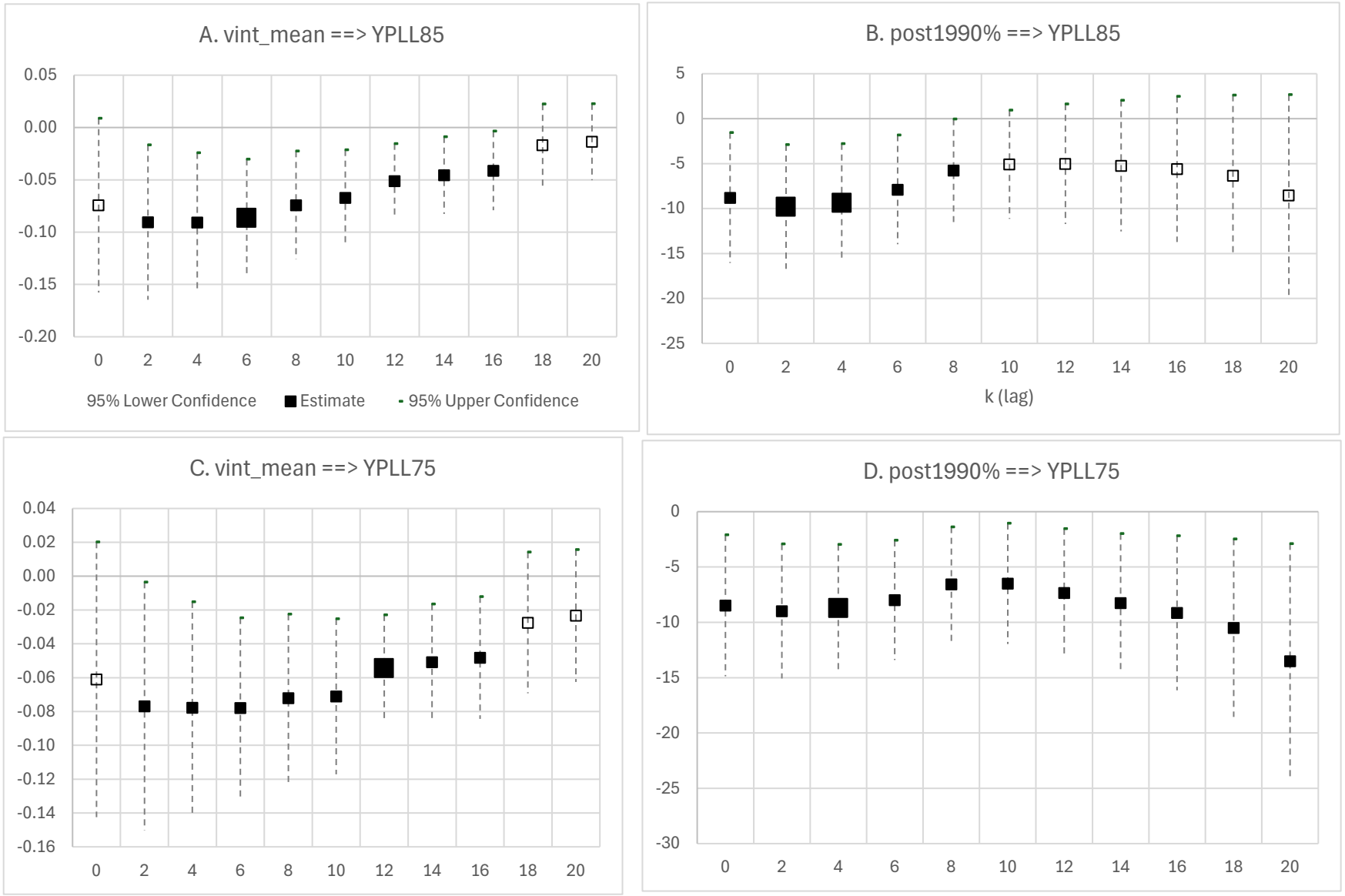
Disturbances were clustered within drugs.

n_rx_{st} = the estimated number (in millions) of U.S. outpatient prescriptions for drug (chemical substance) s in year t (t = 1996, 1997, ..., 2021)

$n_descriptors_{s,t-k}$ = the number of times the descriptor of drug s occurred in PubMed in year t-k (k = 0, 1, 2, ..., 20)

Figure 3

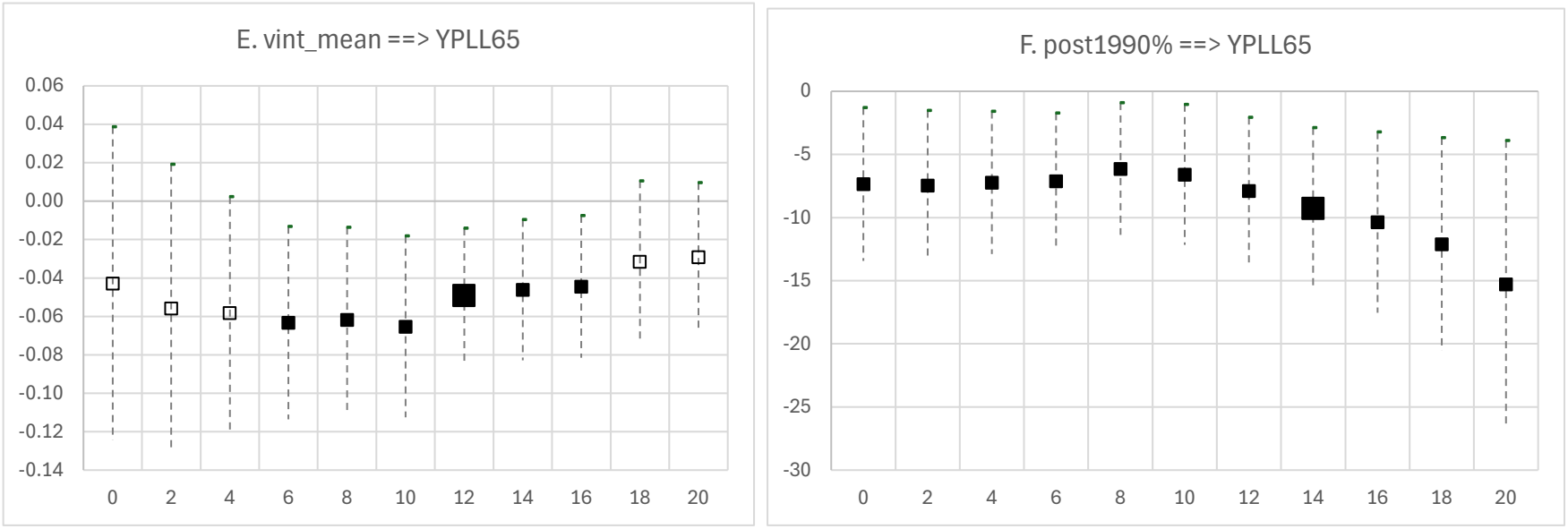
Estimates of β_k from eq. (1): $\ln(\text{mortality}_{dt}) = \beta_k \text{ vintage_measure}_{d,t-k} + \alpha_d + \delta_t + \varepsilon_{dt}$



Continued on next page.

Figure 3

Estimates of β_k from eq. (1): $\ln(\text{mortality}_{dt}) = \beta_k \text{ vintage_measure}_{d,t-k} + \alpha_d + \delta_t + \varepsilon_{dt}$

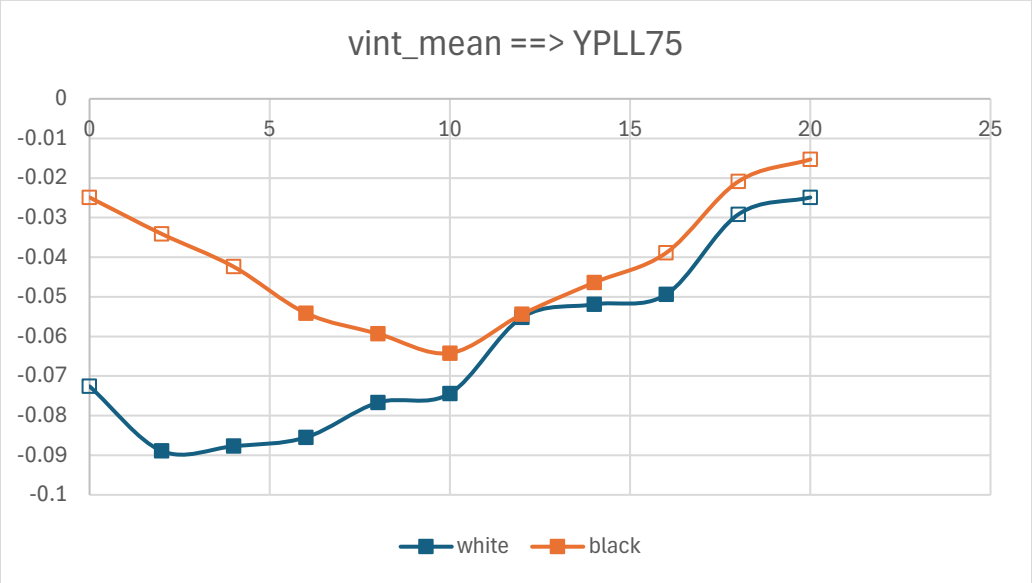


Each estimate is from a separate regression.
Disturbances were clustered within diseases.

Solid squares denote statistically significant estimates; the large solid squares denote the most significant estimates, and hollow squares denote statistically insignificant estimates.

Figure 4

Estimates of β_k from eq. (1): $\ln(\text{mortality}_{dt}) = \beta_k \text{vintage_measure}_{d,t-k} + \alpha_d + \delta_t + \varepsilon_{dt}$, by race



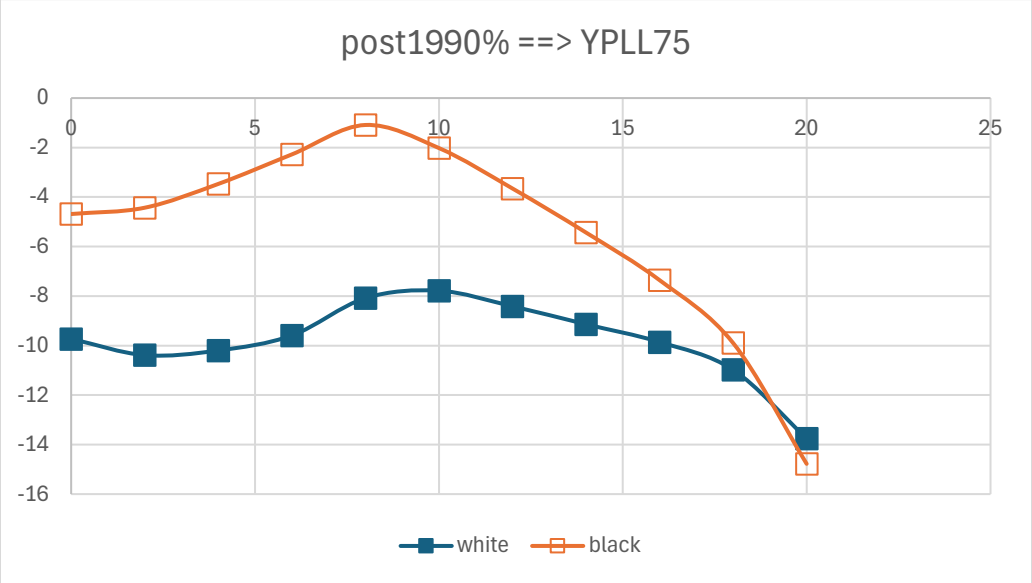


Figure 5

Estimated 1999-2019 changes in age-adjusted mortality rates from all diseases in the presence and absence of biomedical innovation

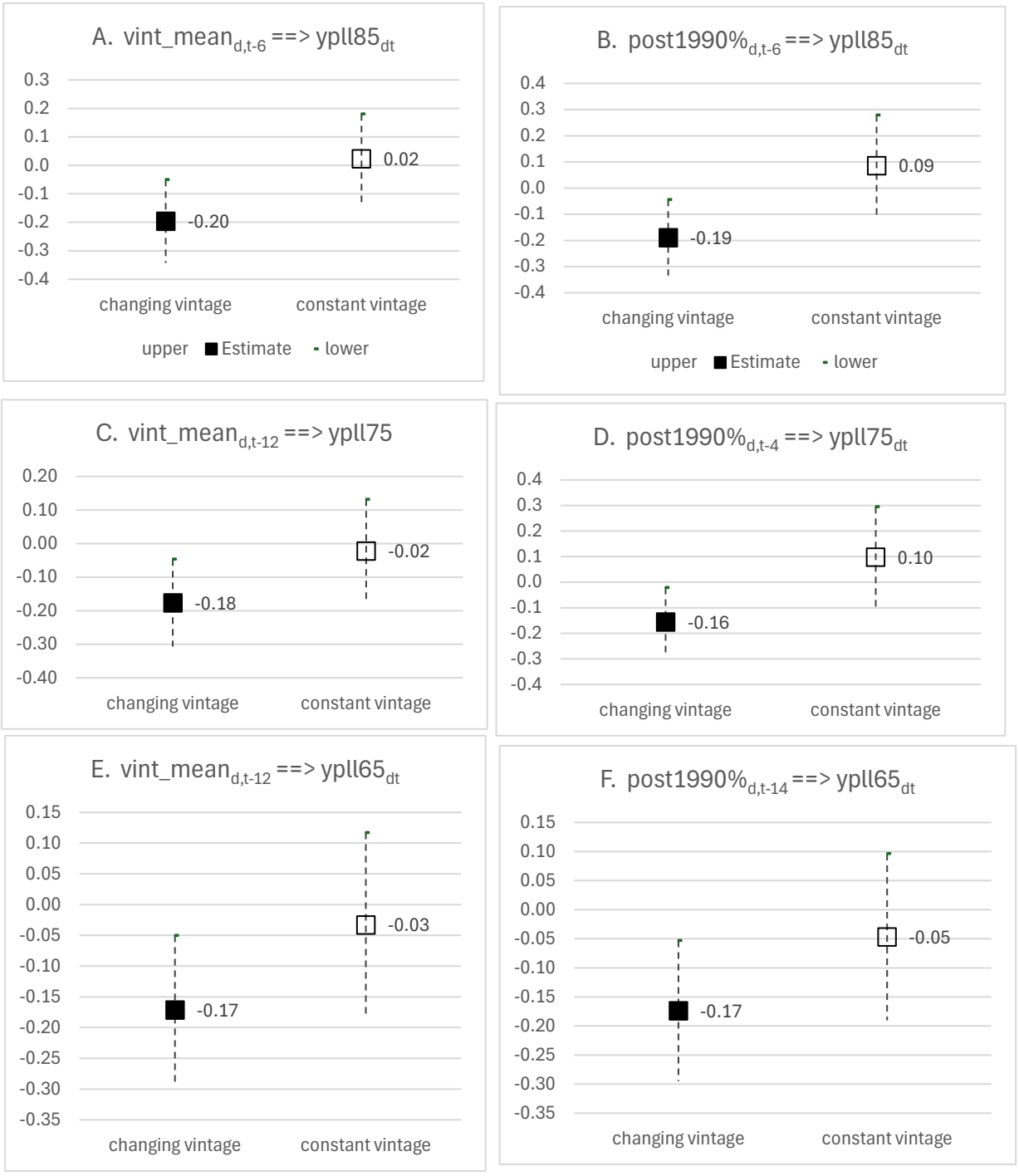


Figure 6

Estimated 1975-2019 change in age-adjusted cancer mortality rate in the presence and absence of changes in cancer incidence and biomedical innovation

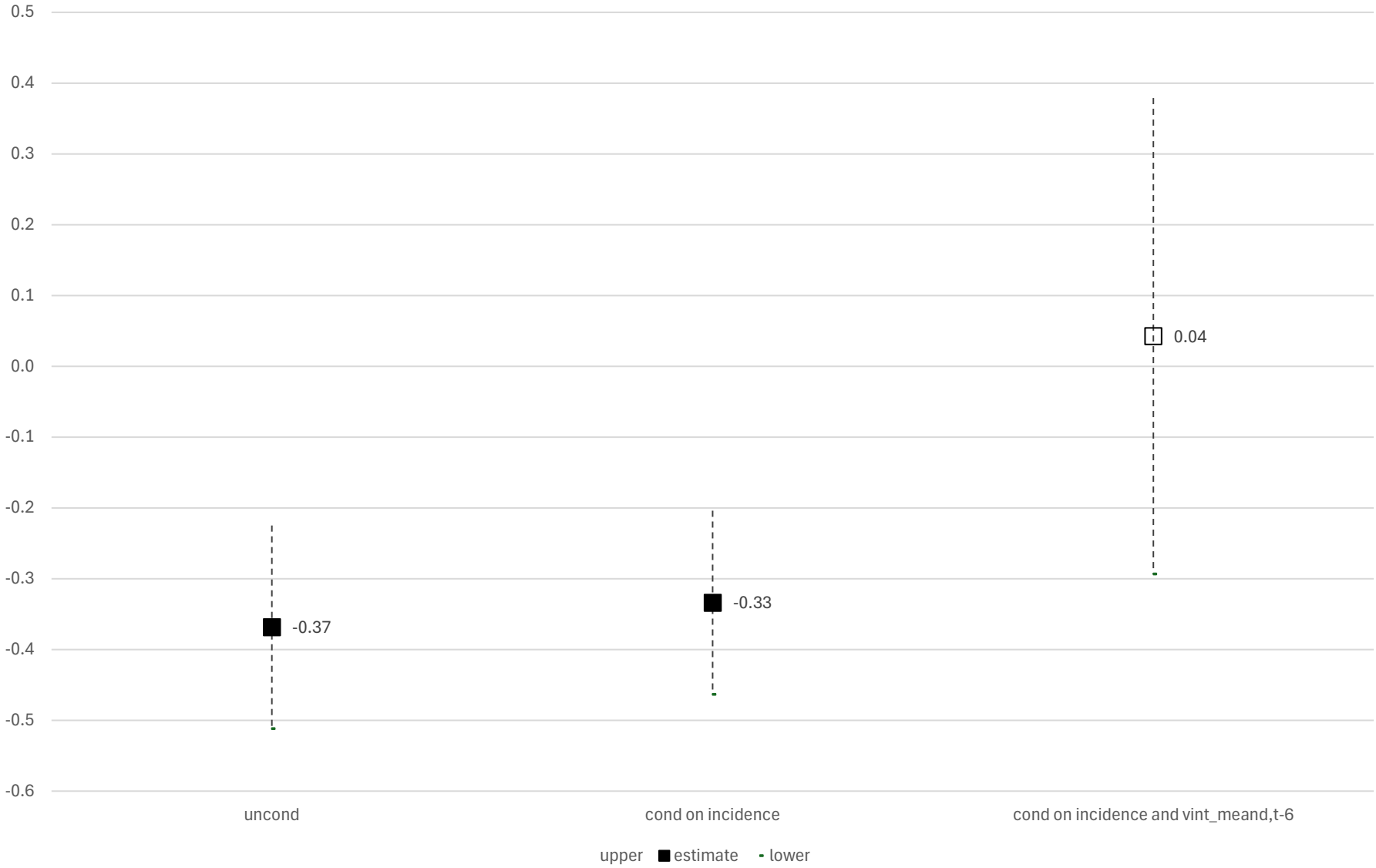


Table 1

Top 25 (out of 1028) 3-digit ICD10 diseases, ranked by number of PubMed articles

Rank	3-digit ICD10 disease	Number of PubMed articles
1	Z33 Pregnant state, incidental	968,121
2	C80 Malignant neoplasm, without specification of site	573,408
3	D36 Benign neoplasm of other and unspecified sites	486,572
4	D49 Neoplasms of unspecified behavior	486,572
5	C50 Malignant neoplasm of breast	323,322
6	B20 Human immunodeficiency virus [HIV] disease resulting in infectious and parasitic diseases	282,371
7	R41 Other symptoms and signs involving cognitive functions and awareness	280,589
8	I10 Essential (primary) hypertension	253,494
9	C34 Malignant neoplasm of bronchus and lung	250,714
10	R54 Senility	247,361
11	E66 Obesity	223,599
12	K76 Other diseases of liver	195,545
13	I21 Acute myocardial infarction	185,429
14	I22 Subsequent myocardial infarction	185,429
15	C22 Malignant neoplasm of liver and intrahepatic bile ducts	184,084
16	F91 Conduct disorders	181,054
17	T14 Injury of unspecified body region	178,993
18	F99 Mental disorder, not otherwise specified	175,350
19	A15 Respiratory tuberculosis, bacteriologically and histologically confirmed	165,146
20	F03 Unspecified dementia	164,569
21	E11 Type 2 diabetes mellitus	163,852
22	R52 Pain, not elsewhere classified	151,130
23	I67 Other cerebrovascular diseases	151,049
24	F20 Schizophrenia	150,860
25	N28 Other disorders of kidney and ureter, not elsewhere classified	148,999

Table 2

Estimates of ρ_k from eq. (2), based on data on about 600 drugs:

$$n_rx_{st} = \rho_k \, n_descriptors_{s,t-k} + \alpha_s + \delta_t + \varepsilon_{st}$$

k (lag)	Estimate	Std. Err.	95% Lower Conf. Limit	95% Upper Conf. Limit	Z	Pr > |Z|
0	0.0000	0.0000	0.0000	0.0000	0.75	0.4542
1	0.0000	0.0001	-0.0001	0.0001	0.85	0.3973
2	0.0014	0.0013	-0.0012	0.0040	1.06	0.2908
3	0.0018	0.0015	-0.0012	0.0047	1.16	0.2455
4	0.0023	0.0017	-0.0011	0.0057	1.34	0.1808
5	0.0031	0.0019	-0.0005	0.0068	1.67	0.0958
6	**0.0039**	**0.0019**	**0.0001**	**0.0076**	**2.03**	**0.0427**
7	**0.0039**	**0.0018**	**0.0004**	**0.0074**	**2.16**	**0.0306**
8	**0.0037**	**0.0017**	**0.0005**	**0.0069**	**2.24**	**0.0248**
9	**0.0037**	**0.0016**	**0.0006**	**0.0068**	**2.35**	**0.0186**
10	**0.0036**	**0.0015**	**0.0007**	**0.0066**	**2.39**	**0.0170**
11	**0.0036**	**0.0015**	**0.0006**	**0.0066**	**2.37**	**0.0176**
12	**0.0035**	**0.0015**	**0.0005**	**0.0065**	**2.27**	**0.0232**
13	**0.0033**	**0.0015**	**0.0003**	**0.0063**	**2.14**	**0.0323**
14	**0.0031**	**0.0016**	**0.0000**	**0.0063**	**1.97**	**0.0491**
15	0.0029	0.0016	-0.0003	0.0061	1.79	0.0741
16	0.0027	0.0017	-0.0006	0.0060	1.61	0.1083
17	0.0025	0.0017	-0.0009	0.0058	1.45	0.1457
18	0.0021	0.0017	-0.0012	0.0055	1.27	0.2058
19	0.0019	0.0017	-0.0014	0.0052	1.12	0.2634
20	0.0016	0.0016	-0.0015	0.0047	1.02	0.3096

Each estimate is from a separate regression.
Disturbances were clustered within drugs.
Estimates in bold are statistically significant (p-value < .05).

n_rx_{st} = the estimated number (in millions) of U.S. outpatient prescriptions for drug (chemical substance) s in year t (t = 1996, 1997, …, 2021)

$n_descriptors_{s,t-k}$ = the number of times the descriptor of drug s occurred in PubMed in year t-k (k = 0, 1, 2, …, 20)

Table 3

Estimates of γ_j from eq. (3):

$$\ln(\text{aa_mort_rate}_{dt}) = \gamma_j \ln(\text{aa_inc_rate}_{d,t-j}) + \alpha_d + \delta_t + \varepsilon_{dt}$$

j (lag)	Estimate	Std. Err.	Z	Pr > |Z|
0	**0.578**	**0.100**	**5.78**	**<.0001**
2	**0.564**	**0.108**	**5.23**	**<.0001**
4	**0.521**	**0.124**	**4.19**	**<.0001**
6	**0.459**	**0.145**	**3.15**	**0.0016**
8	**0.382**	**0.167**	**2.29**	**0.0219**
10	0.306	0.183	1.68	0.0932
12	0.223	0.193	1.16	0.2476
14	0.160	0.196	0.82	0.4125
16	0.101	0.191	0.53	0.5952
18	0.066	0.181	0.36	0.7157
20	0.026	0.168	0.16	0.8766

aa_mort_rate_{dt} = the age-adjusted mortality rate from disease (cancer site) d in year t (t = 1975, 1976, …, 2019)
$\text{aa_inc_rate}_{d,t-j}$ = the age-adjusted incidence rate of disease (cancer site) d in year t

Each estimate is from a separate regression.

Eq. (3) was estimated by weighted least squares, weighting by ($\sum_t \text{aa_mort_rate}_{dt}$).

Disturbances were clustered within cancert sites.

Estimates in bold are statistically significant (p-value < .05).

Table 4

Estimates of the effect of descriptor vintage on age adjusted cancer mortality rate, not controlling and controlling for incidence

	$\ln(aa_mort_rate_{dt}) = \beta_k\ vint_mean_{d,t-k} + \alpha_d + \delta_t + \varepsilon_{dt}$					$\ln(aa_mort_rate_{dt}) = \beta_k\ vint_mean_{d,t-k} + \gamma_j \ln(aa_inc_rate_{dt}) + \alpha_d + \delta_t + \varepsilon_{dt}$			
	not controlling for incidence					controlling for incidence			
k (lag)	β_k Estimate	Std. Err.	Z	Pr > |Z|		β_k Estimate	Std. Err.	Z	Pr > |Z|
0	**-0.065**	**0.028**	**-2.33**	**0.0196**		**-0.056**	**0.028**	**-1.99**	**0.0462**
2	**-0.064**	**0.031**	**-2.07**	**0.0383**		**-0.056**	**0.028**	**-2.04**	**0.0412**
4	-0.064	0.033	-1.95	0.0511		**-0.059**	**0.027**	**-2.19**	**0.0286**
6	-0.061	0.034	-1.80	0.0717		**-0.058**	**0.026**	**-2.27**	**0.0234**
8	-0.058	0.035	-1.64	0.1002		**-0.055**	**0.026**	**-2.17**	**0.0303**
10	-0.051	0.036	-1.42	0.1552		**-0.051**	**0.025**	**-2.02**	**0.0431**
12	-0.041	0.036	-1.12	0.2623		-0.045	0.025	-1.78	0.0749
14	-0.037	0.031	-1.18	0.2395		**-0.046**	**0.022**	**-2.12**	**0.0338**
16	-0.030	0.027	-1.11	0.2661		**-0.043**	**0.020**	**-2.18**	**0.0291**
18	-0.030	0.023	-1.27	0.203		**-0.044**	**0.018**	**-2.51**	**0.0122**
20	-0.023	0.020	-1.16	0.2461		**-0.041**	**0.017**	**-2.45**	**0.0142**

$aa_mort_rate_{dt}$ = the age-adjusted mortality rate from disease (cancer site) d in year t (t = 1975, 1976, …, 2019)

$aa_inc_rate_{dt}$ = the age-adjusted incidence rate of disease (cancer site) d in year t

$vint_mean_{d,t-k}$ = the mean vintage (year of first appearance in PubMed) of descriptors of articles about disease d published in year t-k (k = 0, 2, …, 20)

Each estimate is from a separate regression.

Eq. (5) was estimated by weighted least squares, weighting by ($\sum_t aa_mort_rate_{dt}$).

Disturbances were clustered within cancer sites.

Estimates in bold are statistically significant (p-value < .05).

Table 5

Years of potential life lost before ages 85, 75, and 65 in 1999 and 2019

	1999	2019	% change	ln change
	Before age 85			
YPLL				
all causes	33,530,448	38,561,965	15.0%	0.140
non-external causes[1]	27,986,901	30,429,676	8.7%	0.084
age adjusted rate				
all causes	12,321	10,655	-13.5%	-0.145
non-external causes[1]	10,316	8,074	-21.7%	-0.245
age-adjusted population (100,000s)[2]				
all causes	2,721	3,619	33.0%	0.285
non-external causes[1]	2,713	3,769	38.9%	0.329
	Before age 75			
YPLL				
all causes	19,309,613	21,939,705	13.6%	0.128
non-external causes[1]	15,009,430	15,841,878	5.5%	0.054
age adjusted rate				
all causes	7,418	6,691	-9.8%	-0.103
non-external causes[1]	5,791	4,628	-20.1%	-0.224
age-adjusted population (100,000s)[2]				
all causes	2,603	3,279	26.0%	0.231
non-external causes[1]	2,592	3,423	32.1%	0.278
	Before age 65			
YPLL				
all causes	11,145,856	11,638,857	4.4%	0.043
non-external causes[1]	7,955,848	7,360,735	-7.5%	-0.078
age adjusted rate				
all causes	4,593	4,237	-7.8%	-0.081
non-external causes[1]	3,298	2,644	-19.8%	-0.221
age-adjusted population (100,000s)[2]				
all causes	2,427	2,747	13.2%	0.124
non-external causes[1]	2,413	2,784	15.4%	0.143

1. Deaths from unintentional injury, suicide, and homicide are excluded
2. Age-adjusted population = YPLL / age-adjusted rate

Source: CDC, WISQARS Years of Potential Life Lost (YPLL) Report, 1981 - 2020, https://wisqars.cdc.gov/ypll

Table 6

Estimates of β_k from eq. (1): $\ln(\text{mortality}_{dt}) = \beta_k \text{ vintage_measure}_{d,t-k} + \alpha_d + \delta_t + \varepsilon_{dt}$

		vintage_measure = vint_mean					vintage_measure = post1990%			
row	k (lag)	Estimate	Std. Err.	Z	Pr > |Z|		Estimate	Std. Err.	Z	Pr > |Z|
		$\text{mortality}_{dt} = \text{YPLL85}_{dt}$								
1	0	-0.075	0.043	-1.75	0.0793		**-8.806**	**3.697**	**-2.38**	**0.0172**
2	2	**-0.091**	**0.038**	**-2.40**	**0.0165**		**-9.809**	**3.522**	**-2.79**	**0.0053**
3	4	**-0.091**	**0.034**	**-2.68**	**0.0074**		**-9.374**	**3.359**	**-2.79**	**0.0053**
4	6	**-0.086**	**0.029**	**-3.02**	**0.0025**		**-7.905**	**3.098**	**-2.55**	**0.0107**
5	8	**-0.074**	**0.026**	**-2.81**	**0.0049**		**-5.775**	**2.917**	**-1.98**	**0.0477**
6	10	**-0.067**	**0.024**	**-2.87**	**0.0041**		-5.099	3.082	-1.65	0.0980
7	12	**-0.051**	**0.018**	**-2.81**	**0.0050**		-5.040	3.403	-1.48	0.1385
8	14	**-0.046**	**0.019**	**-2.42**	**0.0153**		-5.257	3.723	-1.41	0.1579
9	16	**-0.041**	**0.019**	**-2.14**	**0.0322**		-5.625	4.125	-1.36	0.1727
10	18	-0.017	0.020	-0.85	0.3967		-6.365	4.581	-1.39	0.1647
11	20	-0.014	0.019	-0.74	0.4588		-8.554	5.717	-1.50	0.1346
		$\text{mortality}_{dt} = \text{YPLL75}_{dt}$								
12	0	-0.061	0.042	-1.48	0.1401		**-8.493**	**3.258**	**-2.61**	**0.0091**
13	2	**-0.077**	**0.037**	**-2.06**	**0.0399**		**-9.002**	**3.103**	**-2.90**	**0.0037**
14	4	**-0.078**	**0.032**	**-2.44**	**0.0149**		**-8.721**	**2.939**	**-2.97**	**0.0030**
15	6	**-0.078**	**0.027**	**-2.87**	**0.0041**		**-7.991**	**2.760**	**-2.90**	**0.0038**
16	8	**-0.072**	**0.025**	**-2.85**	**0.0044**		**-6.563**	**2.637**	**-2.49**	**0.0128**
17	10	**-0.071**	**0.023**	**-3.04**	**0.0024**		**-6.509**	**2.782**	**-2.34**	**0.0193**
18	12	**-0.055**	**0.016**	**-3.39**	**0.0007**		**-7.344**	**2.958**	**-2.48**	**0.0130**
19	14	**-0.051**	**0.018**	**-2.89**	**0.0038**		**-8.267**	**3.203**	**-2.58**	**0.0098**
20	16	**-0.048**	**0.018**	**-2.62**	**0.0088**		**-9.164**	**3.561**	**-2.57**	**0.0101**
21	18	-0.028	0.021	-1.29	0.1960		**-10.504**	**4.101**	**-2.56**	**0.0104**
22	20	-0.024	0.020	-1.18	0.2387		**-13.510**	**5.417**	**-2.49**	**0.0126**
		$\text{mortality}_{dt} = \text{YPLL65}_{dt}$								
23	0	-0.043	0.042	-1.03	0.3032		**-7.358**	**3.095**	**-2.38**	**0.0174**
24	2	-0.056	0.038	-1.46	0.1443		**-7.465**	**3.033**	**-2.46**	**0.0138**
25	4	-0.058	0.031	-1.89	0.0593		**-7.247**	**2.881**	**-2.52**	**0.0119**
26	6	**-0.063**	**0.026**	**-2.48**	**0.0133**		**-7.117**	**2.748**	**-2.59**	**0.0096**
27	8	**-0.062**	**0.025**	**-2.52**	**0.0118**		**-6.148**	**2.670**	**-2.30**	**0.0213**
28	10	**-0.065**	**0.024**	**-2.72**	**0.0066**		**-6.607**	**2.834**	**-2.33**	**0.0197**
29	12	**-0.049**	**0.018**	**-2.75**	**0.0060**		**-7.901**	**2.972**	**-2.66**	**0.0079**
30	14	**-0.046**	**0.019**	**-2.48**	**0.0132**		**-9.291**	**3.264**	**-2.85**	**0.0044**
31	16	**-0.045**	**0.019**	**-2.36**	**0.0185**		**-10.383**	**3.656**	**-2.84**	**0.0045**
32	18	-0.032	0.021	-1.47	0.1402		**-12.119**	**4.303**	**-2.82**	**0.0049**
33	20	-0.029	0.020	-1.48	0.1398		**-15.285**	**5.806**	**-2.63**	**0.0085**

Each estimate is from a separate equation. Estimates in bold are statistically significant (p-value < .05).

Table 7

Estimates of β_k from eq. (1): $\ln(\text{mortality}_{dt}) = \beta_k \text{ vintage_measure}_{d,t-k} + \alpha_d + \delta_t + \varepsilon_{dt}$, by race

		$\text{mortality}_{dt} = \text{YPLL75}_{dt}$								
		vintage_measure = vint_mean								
		White					Black			
row	lag	Estimate	Std. Err.	Z	Pr > |Z|		Estimate	Std. Err.	Z	Pr > |Z|
1	0	-0.073	0.041	-1.79	0.0737		-0.025	0.062	-0.40	0.6866
2	2	**-0.089**	**0.036**	**-2.44**	**0.0147**		-0.034	0.056	-0.61	0.5450
3	4	**-0.088**	**0.032**	**-2.73**	**0.0063**		-0.042	0.043	-0.99	0.3220
4	6	**-0.086**	**0.029**	**-2.99**	**0.0028**		**-0.054**	**0.027**	**-2.00**	**0.0454**
5	8	**-0.077**	**0.027**	**-2.89**	**0.0039**		**-0.059**	**0.025**	**-2.36**	**0.0184**
6	10	**-0.074**	**0.024**	**-3.06**	**0.0022**		**-0.064**	**0.026**	**-2.46**	**0.0140**
7	12	**-0.055**	**0.016**	**-3.57**	**0.0004**		**-0.054**	**0.025**	**-2.15**	**0.0314**
8	14	**-0.052**	**0.018**	**-2.93**	**0.0034**		**-0.046**	**0.024**	**-1.97**	**0.0490**
9	16	**-0.049**	**0.019**	**-2.62**	**0.0088**		-0.039	0.022	-1.74	0.0813
10	18	-0.029	0.022	-1.30	0.1921		-0.021	0.021	-1.00	0.3194
11	20	-0.025	0.021	-1.18	0.2368		-0.015	0.019	-0.79	0.4271
		vintage_measure = post1990%								
		White					Black			
12	0	**-9.729**	**3.317**	**-2.93**	**0.0034**		-4.684	4.080	-1.15	0.2508
13	2	**-10.385**	**3.094**	**-3.36**	**0.0008**		-4.430	4.144	-1.07	0.2851
14	4	**-10.198**	**2.927**	**-3.48**	**0.0005**		-3.469	3.837	-0.90	0.3659
15	6	**-9.592**	**2.783**	**-3.45**	**0.0006**		-2.277	3.362	-0.68	0.4982
16	8	**-8.084**	**2.745**	**-2.95**	**0.0032**		-1.087	2.805	-0.39	0.6982
17	10	**-7.783**	**2.945**	**-2.64**	**0.0082**		-2.019	2.695	-0.75	0.4538
18	12	**-8.412**	**3.138**	**-2.68**	**0.0073**		-3.653	2.918	-1.25	0.2105
19	14	**-9.139**	**3.373**	**-2.71**	**0.0067**		-5.433	3.437	-1.58	0.1139
20	16	**-9.857**	**3.681**	**-2.68**	**0.0074**		-7.356	4.451	-1.65	0.0984
21	18	**-10.979**	**4.148**	**-2.65**	**0.0081**		-9.895	5.860	-1.69	0.0913
22	20	**-13.752**	**5.271**	**-2.61**	**0.0091**		-14.777	9.134	-1.62	0.1057

YPLL75_{dt} = the number of years of potential life lost before age 75 due to disease d in year t (t = 1999, 2000, …, 2019)

$\text{vint_mean}_{d,t-k}$ = the mean vintage (year of first appearance in PubMed) of descriptors of articles about disease d published in year t-k (k = 0, 2, …, 20)

Each estimate is from a separate regression.

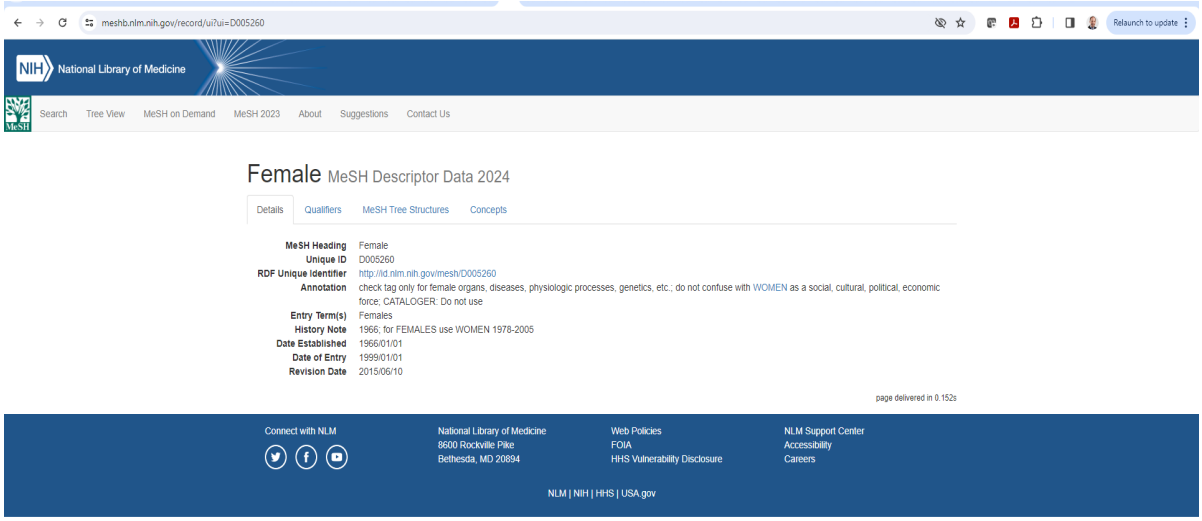
Eq. (1) was estimated by weighted least squares, weighting by ($\sum_t \text{YPLL75}_{dt}$).

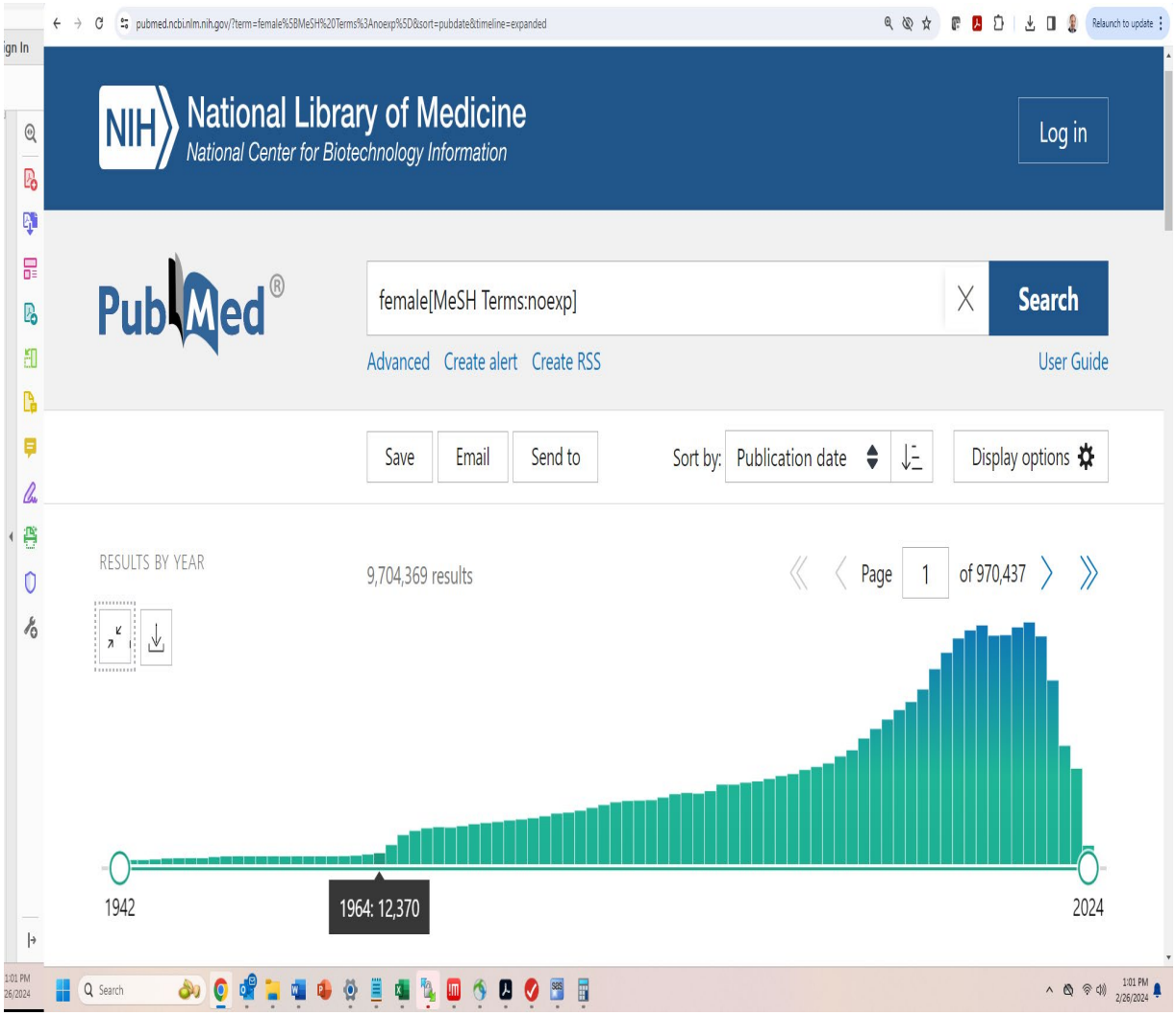
Disturbances were clustered within diseases.

Estimates in bold are statistically significant (p-value < .05).

Appendix Figure 1

Year in which MeSH descriptor 'female" was established, and timeline of PubMed occurences of MeSH descriptor "female"





Appendix Table 1

Estimates of γ_j from eq. (4):

$$vint_mean_{d,t} = \gamma_j \ln(aa_inc_rate_{d,t-j}) + \alpha_d + \delta_t + \varepsilon_{dt}$$

j (lag)	Estimate	Std. Err.	Z	Pr > |Z|
0	-0.378	0.240	-1.58	0.1146
2	-0.361	0.259	-1.39	0.1638
4	-0.346	0.277	-1.25	0.2119
6	-0.294	0.279	-1.05	0.2918
8	-0.204	0.280	-0.73	0.4673
10	-0.092	0.285	-0.32	0.7463
12	-0.011	0.292	-0.04	0.9691
14	0.018	0.305	0.06	0.9535
16	0.021	0.281	0.08	0.9396
18	-0.011	0.270	-0.04	0.9662
20	-0.006	0.268	-0.02	0.9826

$vint_mean_{d,t}$ = the mean vintage (year of first appearance in PubMed) of descriptors of articles about disease d published in year t

$aa_inc_rate_{d,t-j}$= the age-adjusted incidence rate of disease (cancer site) d in year t-j (j = 0, 2, …, 20)

Each estimate is from a separate regression.

Eq. (4) was estimated by weighted least squares, weighting by $n_descriptors_{d,t}$ = the total number of descriptors of articles about disease (cancer site) d in year t

Disturbances were clustered within cancer sites.

Estimates in bold are statistically significant (p-value < .05).

Appendix Table 2

Weighted mean PubMed descriptor vintage, 1979-2019

year	vint_mean	post1990%
1979	1943.1	0.0%
1989	1944.3	0.0%
1999	1946.1	0.9%
2009	1946.9	1.8%
2019	1947.5	3.5%

Diseases are weighted by total YPLL75 during 1999-2019

Appendix Table 3

20 most frequently occuring MeSH descriptors that first appeared in PubMed after 1990

RANK	FREQUENCY	HEADING	first year of PubMed appearance
1	195,584	COVID-19	2019
2	162,322	Oxidative Stress	1992
3	154,372	Reverse Transcriptase Polymerase Chain Reaction	1994
4	142,552	SARS-CoV-2	2019
5	135,213	Polymorphism, Single Nucleotide	1997
6	92,688	Nanoparticles	2000
7	67,087	HEK293 Cells	2001
8	66,247	Proteomics	1999
9	63,546	Real-Time Polymerase Chain Reaction	2002
10	62,884	Tandem Mass Spectrometry	2001
11	60,947	Transcriptome	2002
12	55,791	RNA Interference	2000
13	49,822	Green Fluorescent Proteins	1992
14	48,295	Anti-HIV Agents	1991
15	46,057	Metal Nanoparticles	2001
16	45,147	Tomography, Optical Coherence	1994
17	44,500	Xenograft Model Antitumor Assays	1994
18	44,424	High-Throughput Nucleotide Sequencing	2008
19	42,902	Nanostructures	2000
20	42,706	Molecular Dynamics Simulation	1999

Appendix Table 4

Data on mortality and descriptor attributes for the top 25 diseases, ranked by total ypll75 during 1999-2019

	YPLL75			no. of descriptors		vint_mean		post1990%	
3-digit ICD10 disease	**1999-2019**	**1999**	**2019**	**1999**	**2019**	**1999**	**2019**	**1999**	**2019**
I25 Chronic ischaemic heart disease	26,463,709	1,193,754	#######	8,542	81,530	1946.2	1948.7	0.4%	4.4%
C34 Malignant neoplasm of bronchus and lung	23,423,058	1,087,688	811,290	47,996	141,276	1946.7	1948.6	1.1%	4.3%
I21 Acute myocardial infarction	16,937,069	932,303	610,883	40,769	55,557	1946.7	1948.7	1.0%	4.5%
J44 Other chronic obstructive pulmonary disease	9,660,016	307,373	511,028	419	38,407	1946.2	1947.1	0.7%	2.2%
C50 Malignant neoplasm of breast	9,627,251	432,312	396,333	75,003	164,137	1945.1	1947.1	1.6%	5.2%
K70 Alcoholic liver disease	7,859,390	241,047	444,014	13,996	20,945	1945.3	1946.7	0.7%	2.8%
C18 Malignant neoplasm of colon	6,773,165	284,574	298,696	22,039	31,658	1948.1	1949.6	1.8%	5.9%
C25 Malignant neoplasm of pancreas	5,665,382	194,367	290,714	17,374	50,213	1947.3	1949.7	1.3%	4.7%
A41 Other sepsis	5,104,245	160,817	240,551	10,081	35,403	1945.9	1947.6	0.6%	3.2%
I42 Cardiomyopathy	4,944,219	246,647	185,673	14,459	26,159	1947.0	1948.1	1.1%	3.7%
C71 Malignant neoplasm of brain	4,716,946	194,836	211,735	29,002	63,062	1946.1	1948.2	1.1%	4.5%
K74 Fibrosis and cirrhosis of liver	4,600,938	167,741	212,661	15,266	42,188	1946.0	1948.1	0.6%	3.9%
C80 Malignant neoplasm, without specification of site	4,304,555	207,319	183,201	77,310	216,918	1944.3	1946.5	1.2%	5.4%
C22 Malignant neoplasm of liver and intrahepatic bile ducts	4,256,428	111,365	215,620	37,336	108,606	1947.3	1949.2	1.2%	4.1%
J18 Pneumonia, organism unspecified	3,980,368	157,327	165,789	399	280	1944.4	1946.0	0.0%	1.8%
F10 Mental and behavioural disorders due to use of alcohol	3,782,428	135,467	211,629	16,235	20,966	1944.6	1945.4	0.4%	1.9%
I50 Heart failure	3,198,534	95,506	200,362	22,647	77,982	1946.7	1946.9	0.9%	2.6%
N18 Chronic kidney disease	2,806,031	81,765	169,219	22,306	76,222	1945.6	1946.4	0.7%	2.2%
C56 Malignant neoplasm of ovary	2,713,803	116,141	106,071	20,835	38,888	1947.3	1949.3	2.4%	5.9%
C15 Malignant neoplasm of oesophagus	2,690,895	100,916	122,814	12,645	29,080	1946.2	1948.0	1.3%	4.2%
I51 Complications and ill-defined descriptions of heart disease	2,668,533	80,014	140,902	23,026	48,432	1946.3	1947.8	1.1%	3.9%
E66 Obesity	2,554,340	54,444	141,239	30,974	143,243	1944.6	1945.9	0.6%	2.6%
E11 Type 2 diabetes mellitus	2,528,098	41,218	187,822	27,980	116,215	1946.1	1947.0	0.6%	2.7%
I46 Cardiac arrest	2,467,697	108,896	113,435	5,154	10,827	1944.8	1945.6	0.3%	1.9%