The long-run impact of new medical ideas on cancer survival and mortality

Frank R. Lichtenberg

frl1@columbia.edu

Columbia University, National Bureau of Economic Research, and CESifo

8 May 2017

I am extremely grateful to Pierre Azoulay and Bhaven Sampat for their sharing their complete version of the MEDLINE/PubMED database with me.

8 May 2017

The long-run impact of new medical ideas on cancer survival and mortality

Abstract

I perform tests of the hypothesis that the arrival of new medical ideas has played a major role in the long-run increase in U.S. cancer survival and decline in cancer mortality, by investigating whether the types of cancer (breast, colon, lung, etc.) subject to greater penetration of new ideas had larger subsequent survival gains and mortality reductions, controlling for changing incidence.

I use the MEDLINE/PubMED database, which contains more than 23 million references to journal articles published since 1946 in 5400 of the world's leading biomedical journals, to construct measures of the penetration of new medical ideas applied to most types of cancer over time. MEDLINE/PubMED records are indexed with Medical Subject Headings (MeSH), "one of the most highly sophisticated thesauri in existence today." Between 1955 and 2015, the number of MeSH descriptors increased from 15.8 thousand to 27.8 thousand. I show that newer descriptors are assigned to "research articles" than to "non-research" articles.

The estimates indicate that the 5-year survival rate is not related to the novelty of ideas in (i.e. the fraction of post-1975 MeSH descriptors of) articles published 0-9 years earlier, but strongly positively related to the novelty of ideas in articles published 12-24 years earlier. This finding is consistent with evidence from case studies that it takes a long time for research evidence to reach clinical practice. Between 1994 and 2008, the 5-year observed survival rate for all cancer sites combined increased from 52.1% to 61.2%. The estimates suggest that about 70% of this increase may have been due to the increase in the novelty of medical ideas 12-24 years earlier.

The number of years of potential life lost from cancer before ages 80 and 70 and the number of cancer deaths are inversely related to the novelty of ideas in articles published 12-24 years earlier, conditional on the number of patients diagnosed 1-10 years before and their mean age at time of diagnosis. The increase in descriptor novelty was estimated to have caused a 38% decline in the premature (before age 80) cancer mortality rate 12-24 years later. The fact that this is much larger than the actual (8%) reduction in the premature cancer mortality rate may be due, in part, to declining competing risk from cardiovascular disease.

Frank R. Lichtenberg
Columbia University
504 Uris Hall
3022 Broadway
New York, NY 10027
and NBER
frl1@columbia.edu

I. Introduction

Longevity increase is a very important part of economic growth, broadly defined. Nordhaus (2005) argued that "improvements in health status have been a major contributor to economic welfare over the twentieth century. 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." Murphy and Topel (2006) estimated that cumulative gains in life expectancy after 1900 were worth over $1.2 million to the representative American in 2000, whereas post-1970 gains added about $3.2 trillion per year to national wealth, equal to about half of GDP. The United Nations' Human Development Index, which is used to rank countries into four tiers of human development, is a composite statistic of life expectancy, income per capita, and education (United Nations (2017)).

There is a consensus among macroeconomists that technological progress is the principal source of GDP growth. Romer (1990) argued that "growth…is driven by technological change that arises from intentional investment decisions made by profit-maximizing agents" (S71). Jones argued that "long-run growth is driven by the discovery of new ideas throughout the world."[1,2] And Chien (2015) said that "it has been shown, both theoretically and empirically, that technological progress is the main driver of long-run growth."

Since technological progress, or the discovery of new ideas, is the fundamental source of one of the major components—GDP growth—of "human development," or economic growth, broadly defined, it is quite plausible that the discovery of new ideas has also played a major role in longevity growth. Some previous authors have suggested that this is the case. Fuchs (2010) said that "since World War II…biomedical innovations (new drugs, devices, and procedures) have been the primary source of increases in longevity," although he did not provide evidence to support this claim. Cutler, Deaton and Lleras-Muney (2006) performed a survey of a large and diverse literature on the determinants of mortality, and "tentatively identif[ied] the application of

[1] Sobel (1995) provides an excellent account of a specific innovation that had a substantial positive impact on economic growth: the development (by John Harrison, an 18th-century clockmaker) of the first clock (chronometer) sufficiently accurate to be used to determine longitude at sea—an important development in navigation.

[2] The discovery of new ideas could increase economic output for two different reasons. First, output could simply be positively related to the *quantity* (and variety) of ideas ever discovered. Second, output could be positively related to the (mean or maximum) *quality* of ideas ever discovered, and new ideas may be better (of higher quality), on average, than old ideas. As noted by Jovanovic and Yatsenko (2012), in "the Spence–Dixit–Stiglitz tradition…new goods [are] of higher quality than old goods."

scientific advance and technical progress (some of which is induced by income and facilitated by education) as the ultimate determinant of health." They concluded that "knowledge, science, and technology are the keys to any coherent explanation" of mortality.

In this paper, I will test the hypothesis that the arrival of new medical ideas has played a major role in the long-run increase in U.S. cancer survival and decline in cancer mortality.[3] A difference-in-differences research design will be used: I will investigate whether types of cancer (breast, colon, lung, etc.) subject to greater penetration of new ideas had larger survival gains and mortality reductions, controlling for changing incidence. I will allow for substantial lags in the relationship between new ideas and cancer patient outcomes; evidence from numerous case studies indicates that it takes a long time for research evidence to reach clinical practice. As noted by Morris, Wooding, and Grant (2011), Balas and Bohen (2000), Grant et al (2003) and Wratschko (2009) all estimated a time lag of 17 years measuring different points of the process.

Cancer provides a good opportunity to study the impact of new ideas on mortality, for two reasons. First, as shown in Figure 1, in 2015 cancer was the leading cause of an important mortality measure: the number of years of potential life lost before age 80 (YPLL80). It was the cause of 6.3 million YPLL80. It accounted for 22% of YPLL80 from all causes, and 39% more YPLL80 than the second highest cause, heart disease.[4,5] Second, due to the existence of cancer registries, I can control for changes in the number of people diagnosed and their characteristics (e.g. their mean age). Incidence data are not available for most other diseases.

I will use the MEDLINE/PubMED database to construct measures of the penetration of new medical ideas applied to most types of cancer over time. MEDLINE/PubMED is the U.S. National Library of Medicine's (NLM's) premier bibliographic database; it contains more than 23 million references to journal articles published since 1946 in 5400 of the world's leading biomedical journals. 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;

---

[3] Survival and mortality are not "mirror images" of each other. Survival measures (e.g. the 5-year survival rate) are conditional upon diagnosis, whereas mortality measures are not, although I will control, in an unrestrictive manner, for cancer incidence in the mortality analysis. Also, survival data are based on a sample (only people residing in SEER 9 cancer registry regions are included), whereas mortality data are derived from a census of all deaths.

[4] Cancer was the second highest cause of YPLL70, but it was a close second; it caused 99% as many YPLL70 as the highest cause, unintentional injury.

[5] In view of the importance of cancer as a cause of premature mortality, it is not surprising that Murphy and Topel estimated that "a 1 percent reduction in cancer mortality would be worth $500 billion."

the NLM says that MeSH "is one of the most highly sophisticated thesauri in existence today."[6] It consists of sets of terms ("descriptors") in a hierarchical structure that permits searching at various levels of specificity.[7] There were 27,883 descriptors in 2016 MeSH.[8] Figure 2 shows the distribution of MeSH descriptors by major branch of the MeSH tree.[9] Chemicals and Drugs (branch D) is the largest by far, accounting for 37.7% of descriptors; Analytical, Diagnostic and Therapeutic Techniques, and Equipment [branch E] is the fourth largest, accounting for 8.6% of descriptors.

The MeSH Section staff continually revise and update the MeSH vocabulary.[10] 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. As shown in Figure 3, between 1955 and 2015, the number of MeSH descriptors increased from 15.8 thousand to 27.8 thousand. On average, about 200 descriptors were added per year.

In Section II, I will describe the econometric models of cancer survival and mortality that I will estimate. The measurement of medical idea (MeSH descriptor) novelty will be discussed in Section III. Data sources and descriptive statistics will be presented in Section IV. Empirical results will be presented in Section V. The magnitude of the long-run impact of new ideas on cancer survival and mortality will be quantified in Section VI. Section VII provides a summary and conclusions.

---

[6] https://www.nlm.nih.gov/mesh/meshrels.html

[7] 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.

[8] In addition to these headings, there are more than 232,000 Supplementary Concept Records (SCRs) within a separate file. Generally SCR records contain specific examples of chemicals, diseases, and drug protocols. They are updated more frequently than descriptors. Each SCR is assigned to a related descriptor via the Heading Map (HM) field. The HM is used to rapidly identify the most specific descriptor class and include it in the citation.

[9] A given descriptor can appear multiple times in the MeSH tree. For example, "Drug Therapy, Computer-Assisted" occurs in both branch E (Analytical, Diagnostic and Therapeutic Techniques, and Equipment) and branch L (Information Science).

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

II. Econometric models of cancer survival and mortality

I will estimate models of the 5-year observed survival rate and of several mortality measures using longitudinal cancer-site-level data. The survival rate model to be estimated is:

$$\ln(SURV\_OBS_{st} / (1 - SURV\_OBS_{st})) = \beta_k \, NEW\_IDEAS_{s,t-k}$$
$$+ \gamma \ln(SURV\_EXP_{st} / (1 - SURV\_EXP_{st})) + \pi \ln(N\_DX_{st}) + \alpha_s + \delta_t + \varepsilon_{st} \qquad (1)$$

where

$SURV\_OBS_{st}$ = the observed 5-year survival rate of patients diagnosed in SEER 9 registries with cancer at site s in year t (t = 1994, 2008[11])

$NEW\_IDEAS_{s,t-k}$ = a measure of the novelty of ideas/descriptors in MEDLINE/PubMED articles about cancer at site s in year t-k (k = 0, 3, 6,…,24)

$SURV\_EXP_{st}$ = the expected 5-year survival rate of patients diagnosed in SEER 9 registries with cancer at site s in year t[12]

N_DX_{st} = the number of patients diagnosed in SEER 9 registries with cancer at site s in year t

The measures of $NEW\_IDEAS_{s,t-k}$ will be defined below. The mortality model to be estimated is:

$$\ln(MORT_{st}) = \beta_k \, NEW\_IDEAS_{s,t-k} + \gamma \ln(N\_DX\_10\_YEAR_{st})$$
$$+ \pi \, AGE\_DX\_10\_YEAR_{st} + \alpha_s + \delta_t + \varepsilon_{st} \qquad (2)$$

where $MORT_{st}$ is one of the following variables:

N_DEATHS_{st} = the number of deaths due to cancer at site s in year t (t = 1999, 2013)

$YPLL80_{st}$ = the number of years of potential life lost before age 80 due to cancer at site s in year t

$YPLL70_{st}$ = the number of years of potential life lost before age 70 due to cancer at site s in year t

[11] The 5-year survival rate is "forward-looking": the 5-year survival rate in 2008 is the fraction of patients diagnosed in 2008 who were alive at the end of 2013. 2008 is the most recent year for which data on the 5-year survival rate were available.

[12] The expected survival rate is the survival rate of cancer-free individuals of the same age, sex, and race as cancer patients.

The other variables in eq. (2) are

$N\_DX\_10\_YEAR_{st}$ = the average annual number of people diagnosed with cancer at site s in years t-10 to t-1

$AGE\_DX\_10\_YEAR_{st}$ = the mean age at time of diagnosis of people diagnosed with cancer at site s in years t-10 to t-1

Eqs. (1) and (2) will be estimated via weighted-least squares. The weight for eq. (1) will be N_DX_{st}; the weight for eq. (2) will be $(1 / T) \sum_t MORT_{st}$. Disturbances will be clustered within cancer sites.

From the fixed-effects models (eqs. (1) and (2)), we can derive "long-difference" models. For example, eq. (2) implies the following long-difference model of cancer mortality:

$$\Delta \ln(MORT_s) = \beta_k \, \Delta NEW\_IDEAS\_k_s + \gamma \, \Delta \ln(N\_DX\_10\_YEAR_s) + \pi \, \Delta AGE\_DX\_10\_YEAR_s + \delta' + \varepsilon'_s \quad (3)$$

where

$\Delta \ln(MORT_s) = \ln(MORT_{s,2013}) - \ln(MORT_{s,1999})$

$\Delta NEW\_IDEAS\_k_s = NEW\_IDEAS_{s,2013-k} - NEW\_IDEAS_{s,1999-k}$

$\Delta \ln(N\_DX\_10\_YEAR_s) = \ln(N\_DX\_10\_YEAR_{s,2013}) - \ln(N\_DX\_10\_YEAR_{s,1999})$

$\Delta AGE\_DX\_10\_YEAR_s = AGE\_DX\_10\_YEAR_{s,2013} - AGE\_DX\_10\_YEAR_{s,1999}$

$\delta' = \delta_{2013} - \delta_{1999}$

In Section V, I will present a graph (Figure 7) based on eq. (3).

III. Measurement of idea/descriptor novelty

One major branch (branch C) of the MeSH tree is the Diseases branch; it contains descriptors of thousands of diseases. By using these descriptors, we can identify all MEDLINE/PubMED articles that are about a particular disease, e.g. breast neoplasms. For each of those articles, we can determine the year of publication, and all of the descriptors assigned by

the MeSH Section staff. A total of 247 million descriptors are assigned to the 27 million articles in MEDLINE/PubMED, so the average number of descriptors per article is 9.2.

One potential measure of idea novelty is the (frequency-weighted) mean *vintage* of MeSH descriptors used in articles about cancer at site s published in year t:

$$VINTAGE_{st} = \frac{\Sigma_d \, FREQ_{dst} \, FIRST\_YEAR_d}{\Sigma_d \, FREQ_{dst}}$$

where

$VINTAGE_{st}$ = the (frequency-weighted) mean vintage of MeSH descriptors assigned to articles about cancer at site s published in year t

$FREQ_{dst}$ = the number of times MeSH descriptor d was assigned to articles about cancer at site s published in year t

$FIRST\_YEAR_d$ = the first year in which MeSH descriptor d was assigned to any article in MEDLINE/PubMED

Because the time coverage of MEDLINE/PubMED is generally 1946 to the present, with some older material, the variable $FIRST\_YEAR_d$ is left-censored.[13,14] Therefore, a measure like the following, based on a binary function of FIRST_YEAR, may be more appropriate than a measure based on FIRST_YEAR itself:

$$POST1975\%_{st} = \frac{\Sigma_d \, FREQ_{dst} \, POST1975_d}{\Sigma_d \, FREQ_{dst}}$$

where

$POST1975\%_{st}$ = the (frequency-weighted) fraction of MeSH descriptors assigned to articles about cancer at site s published in year t that were established after 1975

$POST1975_d$ = 1 if $FIRST\_YEAR_d > 1975$

[13] The number of descriptors increased from 297 in 1944 to 11,205 in 1947.

[14] I calculated the vintage of each descriptor ($FIRST\_YEAR_d$) by determining the first year in which the descriptor occurred in the file containing 247 million descriptor records merged with publication dates. Two alternative potential methods of calculating the vintage of each descriptor did not yield reliable results. One of these methods is to use the DA (DATE OF ENTRY) field in the ASCII MeSH file. The other is to use the Medline citation counts by descriptor and year from the Unified Medical Language System. See https://www.nlm.nih.gov/research/umls/2006AA_umls_documentation.pdf

$= 0$ if $FIRST\_YEAR_d \leq 1975$

The MeSH FTP Archive includes lists of new MeSH descriptors added each year since 1999. For example, one file shows MeSH descriptors added in 2016. Some new MeSH descriptors are not new ideas. For example, one MeSH descriptor added in 2016 was "Alcohol Drinking in College" (UI D000067292); this "idea" undoubtedly occurred to someone well before 2016. Hence, the relative frequency of new MeSH descriptors is a noisy, or imperfect, measure of the penetration of new medical ideas. If this measurement error is random, it will bias the coefficients on our measures of new ideas towards zero.

The MEDLINE/PubMED data provide us with an opportunity to assess the validity of MeSH descriptor novelty as an indicator of new ideas, or technological progress. Most scholars agree with Jones' (1998, pp.89-90) statement that "technological progress is driven by research and development (R&D) in the advanced world." It is possible to distinguish between MEDLINE/PubMED articles that resulted from research funding (when that financial support is mentioned in the articles) and MEDLINE/PubMED articles that did not result from (or mention) research funding.[15] One would expect newer descriptors to be assigned to "research-based articles" than to "non-research-based" articles. Figure 4 shows the % of descriptors in 2010 publications established after 1980, by type of research support. 4.5% of the descriptors in 2010 "non-research" publications were established after 1980. 8.1% of the descriptors in 2010 publications that mentioned non-government (but not government) research support were established after 1980. 8.8% of the descriptors in 2010 publications that mentioned government (but not non-government) research support were established after 1980. And 10.5% of the descriptors in 2010 publications that mentioned both non-government and government research support were established after 1980. This evidence supports the hypothesis that newer descriptors are assigned to "research articles" than to "non-research" articles.

The vintage measure described above will be constructed using data on one MeSH record type: MeSH descriptors. Another MeSH record type is Supplementary Concept Records (SCRs).

---

[15] MeSH includes Publication Types to identify financial support of the research that resulted in the published papers when that support is mentioned in the articles. Three types of research support (Non-U.S. Gov't, U.S. Gov't--Non-P.H.S., and U.S. Gov't--P.H.S.) have been coded since 1975; two types (N.I.H.--Extramural and N.I.H.--Intramural) have been coded since 2005; and one type (American Recovery & Reinvestment Act) has been coded since 2010. Funding Support (Grant) Information in MEDLINE/PubMed.

SCRs are used to index chemicals, drugs, and other concepts such as rare diseases. The vintage measure described above (POST1975%$_{st}$) can be constructed using both SCRs and descriptor records combined instead of using descriptor records alone.[16]

There are several reasons to believe that vintage measures based solely on descriptor records are more reliable than vintage measures based on both SCRs and descriptor records combined. The NLM says that descriptors "play a central role in MeSH vocabulary as a unit of indexing and retrieval"[17]; they don't say that about SCRs. Descriptors seem to be more carefully curated than SCRs: descriptors are generally updated on an annual basis (but may, on occasion, be updated more frequently), while SCRs are created daily. Most importantly, vintage measures based on both SCRs and descriptor records combined may be subject to double-counting, since "each SCR is linked to one or more Descriptors by the Heading Mapped To (HM) field in the SCR."[18]

IV. Data sources and descriptive statistics

Survival rate data for SEER 9 registries were obtained from SEER*Stat Software (Version 8.3.4).

Mortality data were obtained from the raw datafiles of the WHO Mortality Database, which is a compilation of mortality data by age, sex and cause of death, as reported annually by Member States from their civil registration systems.

MEDLINE/PubMed data were provided by Pierre Azoulay and Bhaven Sampat, who obtained and reformatted data from the NLM Bulk Download FTP site.

Cancer incidence data from SEER 9 registries were obtained from SEER research data, which include SEER incidence and population data associated by age, sex, race, year of diagnosis, and geographic areas (including SEER registry and county).

Population by age data were obtained from CDC Wonder Bridged-Race Population Estimates 1990-2013.

---

[16] Although the number of SCR terms (> 230,000) is much greater than the number of descriptors (about 28,000), the number of descriptor records in PubMed cancer publications is more than 5 times as great as the number of SCRs.

[17] MeSH Record Types.

[18] 9 of the 10 (and 19 of the 31) largest-selling cancer drugs (e.g. RITUXIMAB, BEVACIZUMAB) have descriptor records rather than SCRs. Other cancer drugs with SCRs are mapped to descriptor records. For example, NILOTINIB is mapped to Pyrimidines, and IPILIMUMAB is mapped to Antibodies, Monoclonal.

Disease classification. The disease (cancer site) classification used in the analysis was based on the SEER Cause of Death Recode 1969+. Appendix Table 1 shows the MeSH descriptors linked to SEER causes of death.

Observed and expected 5-year survival rates in 1994 and 2008 of patients diagnosed in SEER 9 registries, by cancer site ranked by descending number of patients diagnosed in 1994, are shown in Appendix Table 2.

Mortality and incidence data for 1999 and 2013, by cancer site ranked by descending YPLL80 in 1999, are shown in Appendix Table 3.

Figure 5 shows the percent of descriptors in 2013 articles that were established after 1980, by cancer site, for cancer sites with at least 10,000 descriptors in 2013. The percent of post-1980 descriptors in 2013 was almost twice as great for the top two cancer sites as it was for the bottom two.

V. Empirical results

a. Survival model (eq. (1)) estimates

Estimates of β_k parameters of models of the 5-year observed survival rate (eq. (1)) are shown in Table 1. Each estimate is from a separate model. All models included $\ln(SURV\_EXP_{st} / (1 - SURV\_EXP_{st}))$, $\ln(N\_DX_{st})$, cancer-site and year fixed effects; coefficients on these variables are not shown to conserve space.[19] I estimated 6 sets of models; these sets varied with respect to (1) whether all cancer sites were included; (2) whether the NEW_IDEAS measure was based on descriptor records only or descriptor + supplementary concept records; and (3) the year threshold (e.g. 1975) for distinguishing between "new ideas" and "old ideas." For each set, I estimated the model for 9 different assumed lags (0, 3, 6,…,24 years) of NEW_IDEAS.

In Panel A of Table 5, all cancer sites are included; the NEW_IDEAS measure is based on descriptor records only; and the new-idea year threshold is 1975. When the lag length $k \leq 9$, the estimates of β_k are not statistically significant. However, when $k \geq 12$, the estimates of β_k are

[19] The coefficient on $\ln(SURV\_EXP_{st} / (1 - SURV\_EXP_{st}))$ was positive and significant in some, but not all, models. The coefficient on $\ln(N\_DX_{st})$ was insignificant in all models.

positive and highly statistically significant. This indicates that the 5-year survival rate is not related to the novelty of ideas in (i.e. the fraction of post-1975 descriptors of) articles published 0-9 years earlier, but strongly related to the novelty of ideas in articles published 12-24 years earlier. This finding is consistent with the evidence from case studies cited above that it takes a long time for research evidence to reach clinical practice.

The estimates in Table 1 are weighted by the number of patients diagnosed, and as shown in Appendix Table 2, the cancer site with the largest number of patients diagnosed is prostate cancer. Although prostate cancer does not appear to be an outlier (e.g. with respect to NEW_IDEAS, as shown in Figure 5), some estimates suggested that it might be an influential observation. Panel B of Table 1 shows estimates of eq. (1) when prostate cancer is excluded. The point estimates of the parameters are somewhat smaller, but as in Panel A, when the lag length $k \leq 9$, the estimates of β_k are not statistically significant, and when $k \geq 12$, the estimates of β_k are positive and highly statistically significant.

In Panel C of Table 1, all cancer sites are included, and the NEW_IDEAS measure is based on descriptor + supplementary concept records. The same pattern emerges: estimates of β_k are positive and highly statistically significant only when $k \geq 12$.

In Panel D, prostate cancer is excluded, and the NEW_IDEAS measure is based on descriptor + supplementary concept records. In this case, the β_{15} and β_{18} coefficients are statistically significant (the β_{12} coefficient is marginally significant), but the β_{21} and β_{24} coefficients are insignificant. However, it may be inappropriate to exclude prostate cancer,[20] and as discussed above, vintage measures based solely on descriptor records are probably more reliable than vintage measures based on both SCRs and descriptor records combined.

Panels E and F of Table 1 examine the sensitivity of the estimates to the choice of the year threshold for distinguishing between "new ideas" and "old ideas." In Panel E, the year threshold is 1970; in Panel F, it is 1980. In both cases, estimates of β_k are positive and highly statistically significant only when $k \geq 12$.

[20] The fact that an observation is influential does not necessarily mean that it should be excluded.

b. Mortality model (eq. (2)) estimates

Estimates of β_k parameters of models of mortality measures (eq. (2)) are shown in Table 2. Once again, each estimate is from a separate model. All models included $\ln(N\_DX\_10\_YEAR_{st})$, $AGE\_DX\_10\_YEAR_{st}$, cancer-site and year fixed effects; coefficients on these variables are not shown to conserve space.[21] I estimated 6 sets of models; these sets varied with respect to (1) the dependent variable (mortality measure); and (2) whether the NEW_IDEAS measure was based on descriptor records only or descriptor + supplementary concept records. In all models in Table 2, the year threshold for distinguishing between "new ideas" and "old ideas" was 1980. For each set, I estimated the model for 9 different assumed lags (0, 3, 6,…,24 years).

In Panel A of Table 2, the dependent variable is ln(YPLL80)—the log of the number of years of potential life lost before age 80. The estimated coefficient on $POST1980\%_{s,t-k}$ is not significant for $k \leq 6$, but is negative and highly significant for $k \geq 9$. This indicates that premature (before age 80) mortality is inversely related to the fraction of descriptors established after 1980 9-24 years earlier. The estimates in Panel A are plotted in Figure 6. The relationship across cancer sites between the 1981-1995 change in POST1980% and the 1999-2013 change in ln(YPLL80), controlling for changes in incidence, is shown in Figure 7.

The estimates in Panel A are weighted by mean YPLL80. As shown in Appendix Table 3, cancer of the lung and bronchus is the largest cause of YPLL80 by far, accounting for more than a quarter of the total, and almost three times as much as the second highest cause, breast cancer. But excluding cancer of the lung and bronchus has very little effect on the estimates. For example, estimates of β_{18} when cancer of the lung and bronchus are included (as in Panel A) and excluded are -10.92 (Z = 4.26) and -10.83 (Z = 4.41), respectively. Changing the year threshold for distinguishing between "new ideas" and "old ideas" to either 1975 or 1985 also has little effect on the estimates.

In Panel B of Table 2, POST1980% is based on descriptor + supplementary concept records. The estimates of β_k are negative and highly statistically significant only when $k \geq 12$.

[21] The coefficient on $\ln(N\_DX\_10\_YEAR_{st})$ was positive and highly significant in all models. The coefficient on $AGE\_DX\_10\_YEAR_{st}$ was negative and highly significant in the YPLL80 and YPLL70 models, and insignificant in the DEATHS models.

In panels C and D of Table 2, a lower age cutoff—age 70—is used for measuring premature mortality. The results are similar to those in Panels A and B, where the higher (age 80) age cutoff was used. In Panels E and F of Table 2, the dependent variable is the log of the number of deaths. The estimates indicate that the number of deaths is inversely related to the novelty of ideas published in articles 12-24 years earlier, conditional on the number of patients diagnosed 1-10 years before and their mean age at time of diagnosis.

VI. Discussion

The estimates in Tables 1 and 2 are highly consistent with the hypothesis that the application of new ideas has increased cancer survival and reduced cancer mortality, with a significant lag. Now I will quantify the magnitude of these impacts.

The 1994-2008 change in the weighted (by number of patients diagnosed) mean value of the log-odds of the observed 5-year survival rate ($\ln(SURV\_OBS_{st} / (1 - SURV\_OBS_{st}))$) was 0.527.[22] The increase in the log-odds of survival attributable to the increasing share of post-1975 descriptors may be estimated by multiplying the estimated coefficients in Table 1 by the weighted mean change in the lagged share of post-1975 descriptors. Panel A of Table 3 shows these calculations using the statistically significant estimates in Panel A of Table 1.

The estimate of β_{12} suggests that the increase in descriptor novelty 12 years earlier accounted for almost all (94% = .496 / .527) of the increase in survival. The estimate of β_{24} suggests that the increase in descriptor novelty 24 years earlier accounted for about half as much (48% = .251 / .527) of the increase in survival. The average of the estimates for $12 \leq k \leq 24$ suggests that the increase in descriptor novelty 12-24 years earlier accounted for 71% (= .372 / .527) of the increase in survival.

The 1999-2013 change in the weighted (by mean YPLL80) mean value of ln(YPLL80) was 0.045.[23] During that period, the population below age 80 increased by 12.7%, from 270.0 to 304.3 million, so that YPLL80 from cancer per person below age 80 declined by about 8.0%. As shown in Panel B of Table 3, the average of the estimates for $9 \leq k \leq 24$ in Panel A of Table 2

[22] This is larger than the 0.371 increase in the log-odds of the survival rate for all cancer sites combined. That survival rate increased from 52.1% in 1994 to 61.2% in 2008.

[23] This is similar to the 0.040 log increase (from 6.06 to 6.31 million) indicated by Years of Potential Life Lost (YPLL) Reports, 1999 – 2015.

suggests that the increase in descriptor novelty 9-24 years earlier resulted in a 26% decline in YPLL80, and a 38% decline in YPLL80 per person below age 80.

This is much larger than the actual (8%) reduction in YPLL80 from cancer per person below age 80. This suggests that, if new ideas had not been applied to cancer, the premature (before age 80) cancer mortality rate would have increased significantly (by about 30%) between 1999 and 2013. A possible explanation for this is declining competing risk from cardiovascular disease (Honoré and Lleras-Muney (2006)), although inclusion of cancer incidence measures in the mortality model may have controlled for this to some extent.

VII. Summary and conclusions

I performed tests of the hypothesis that the arrival of new medical ideas has played a major role in the long-run increase in U.S. cancer survival and decline in cancer mortality, by investigating whether the types of cancer (breast, colon, lung, etc.) subject to greater penetration of new ideas had larger subsequent survival gains and mortality reductions, controlling for changing incidence.

I used the MEDLINE/PubMED database, which contains more than 23 million references to journal articles published since 1946 in 5400 of the world's leading biomedical journals, to construct measures of the penetration of new medical ideas applied to most types of cancer over time. MEDLINE/PubMED records are indexed with Medical Subject Headings (MeSH), “one of the most highly sophisticated thesauri in existence today.” Between 1955 and 2015, the number of MeSH descriptors increased from 15.8 thousand to 27.8 thousand. Newer descriptors are assigned to “research articles” than to “non-research” articles.

The estimates indicated that the 5-year survival rate is not related to the novelty of ideas in (i.e. the fraction of post-1975 MeSH descriptors of) articles published 0-9 years earlier, but strongly positively related to the novelty of ideas in articles published 12-24 years earlier. This finding is consistent with evidence from case studies that it takes a long time for research evidence to reach clinical practice. Between 1994 and 2008, the 5-year observed survival rate for all cancer sites combined increased from 52.1% to 61.2%. The estimates suggest that about 70% of this increase may have been due to the increase in the novelty of medical ideas 12-24 years earlier.

The number of years of potential life lost from cancer before ages 80 and 70 and the number of cancer deaths are inversely related to the novelty of ideas in articles published 12-24 years earlier, conditional on the number of patients diagnosed 1-10 years before and their mean age at time of diagnosis. The increase in descriptor novelty was estimated to have caused a 38% decline in the premature (before age 80) cancer mortality rate 12-24 years later. The fact that this is much larger than the actual (8%) reduction in the premature cancer mortality rate may be due, in part, to declining competing risk from cardiovascular disease.

References

Balas EA, Boren SA (2000). *Yearbook of Medical Informatics: Managing Clinical Knowledge for Health Care Improvement*. Stuttgart, Germany: Schattauer Verlagsgesellschaft mbH.

Chien YL (2015). What Drives Long-Run Economic Growth?, Federal Reserve Bank of St. Louis.

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

Fuchs, VR (2010), New Priorities for Future Biomedical Innovations. *N Engl J Med* 363:704-706, August 19,

Grant J, Green L, Mason B (2003). Basic research and health: a reassessment of the scientific basis for the support of biomedical science. Res Eval 12:217–24 Google Scholar

Greenwood, J., Z. Hercowitz, and P. Krusell. 1997. "Long-Run Implications of Investment-Specific Technological Change." *The American Economic Review*: 342–362.

Grossman GM, Helpman E (1991). "Quality Ladders in the Theory of Growth." *The Review of Economic Studies*, 58(1): 43–61.

Hanney SR, Castle-Clarke S, Grant J, et al (2015). How long does biomedical research take? Studying the time taken between biomedical and health research and its translation into products, policy, and practice. *Health Research Policy and Systems*. 13:1. doi:10.1186/1478-4505-13-1.

Honoré BE, Lleras-Muney A (2006). Bounds in Competing Risks Models and the War on Cancer. *Econometrica* 74(6): 1675-1698, November.

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.

Morris ZS, Wooding S, Grant J. (2011). The answer is 17 years, what is the question: understanding time lags in translational research. *J R Soc Med.* 104 (12): 510-20, December.

Murphy KM, Topel RH (2006). The Value of Health and Longevity, *Journal of Political Economy* 114 (5).

Romer P (1990). Endogenous Technological Change. *Journal of Political Economy* 98 (5), Part 2: S71-S102.

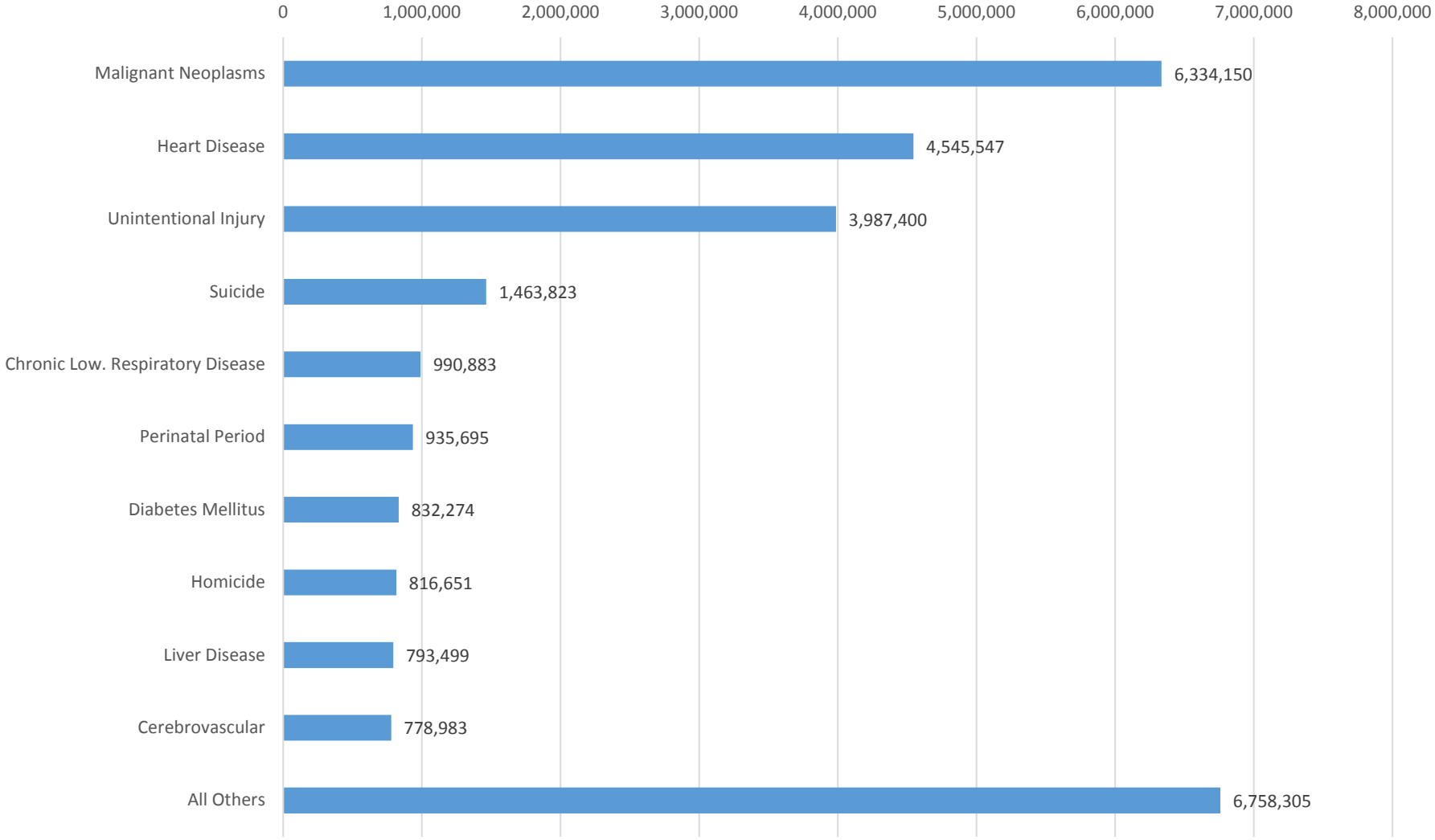
Sobel D (1995). *Longitude: The True Story of a Lone Genius Who Solved the Greatest Scientific Problem of His Time.* New York: Bloomsbury.

Solow R. 1960. "Investment and Technological Progress." In *Mathematical methods in the social sciences, 1959*, edited by K. Arrow, S. Karlin, and P. Suppes. Stanford, Calif: Stanford University Press.

Westfall J, Mold J, Fagnan L. Practice-based research – "Blue Highways" on the NIH roadmap. JAMA 2007;297:403–6.

Wratschko K (2009). Empirical Setting: The pharmaceutical industry. Strategic Orientation and Alliance Portfolio Configuration. New York, NY: Springer.

Figure 1
Number of years of potential life lost before age 80, 2015



Source: CDC, Years of Potential Life Lost (YPLL) Reports, 1999 - 2015, https://webappa.cdc.gov/sasweb/ncipc/ypll10.html

Figure 2
Distribution of MeSH descriptors, by major branch of MeSH Tree

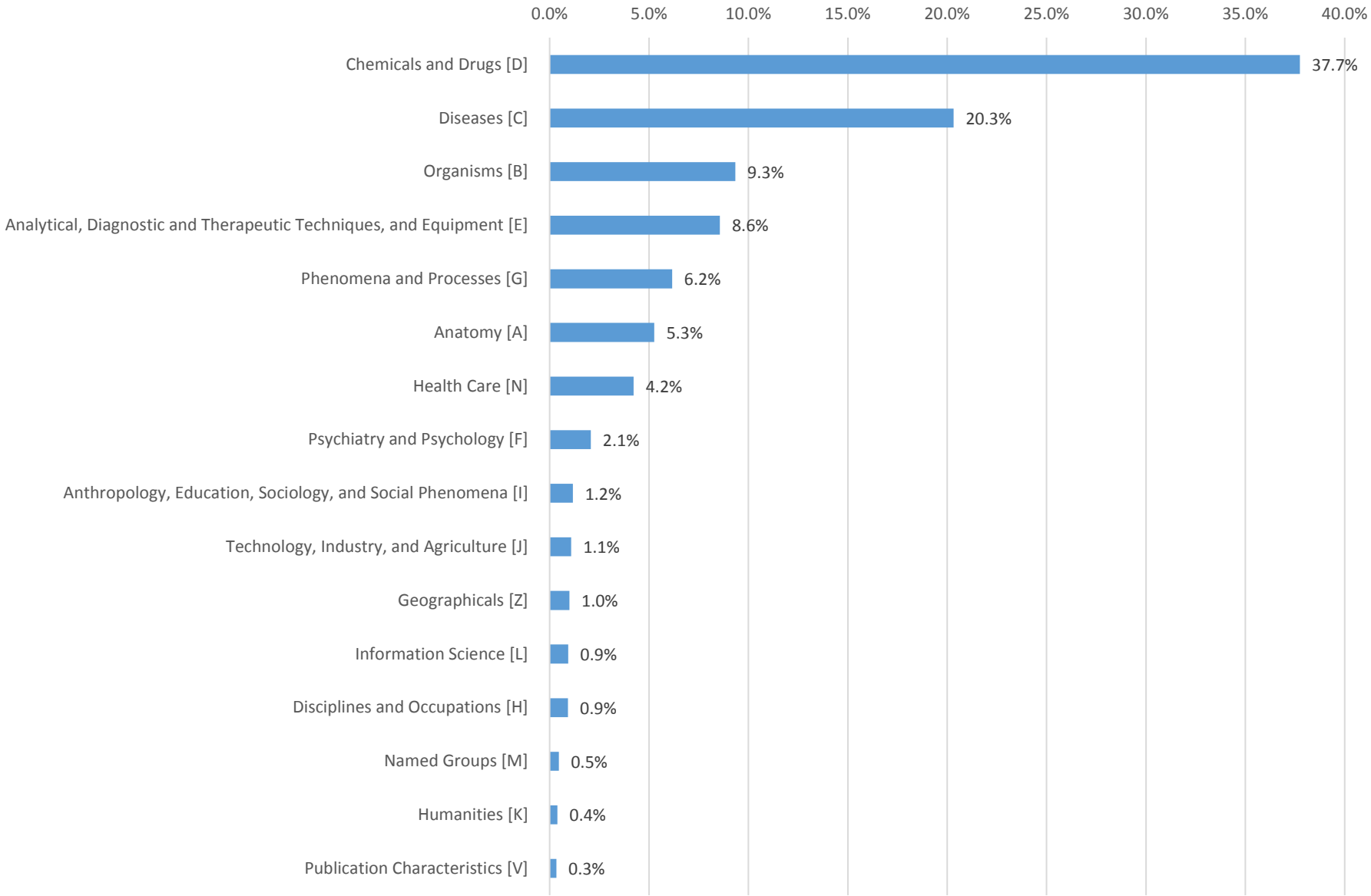


Figure 3
Number of MeSH descriptors ever established, 1955-2015

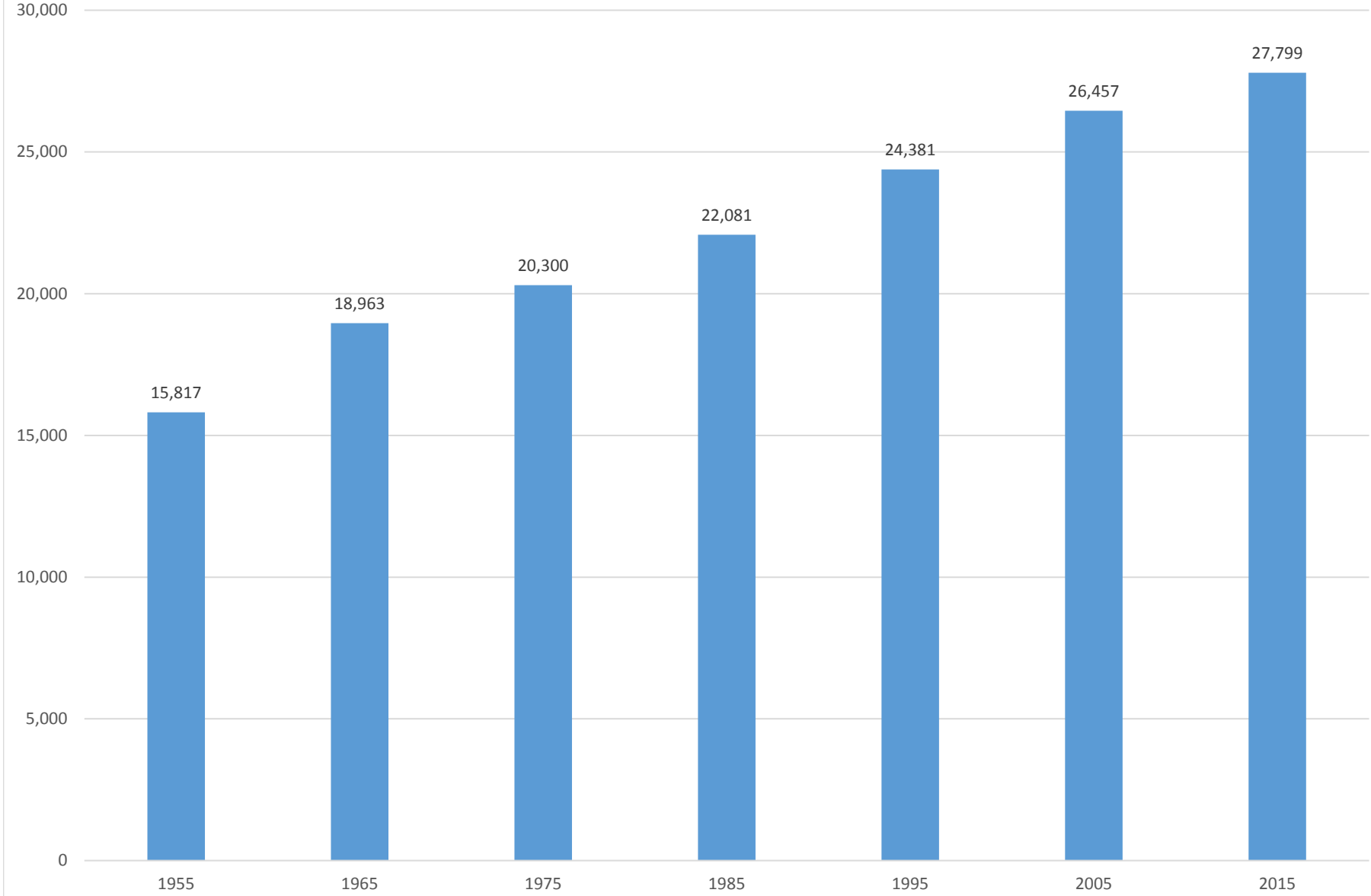


Figure 4
% of descriptors in 2010 publications established after 1980, by type of research support

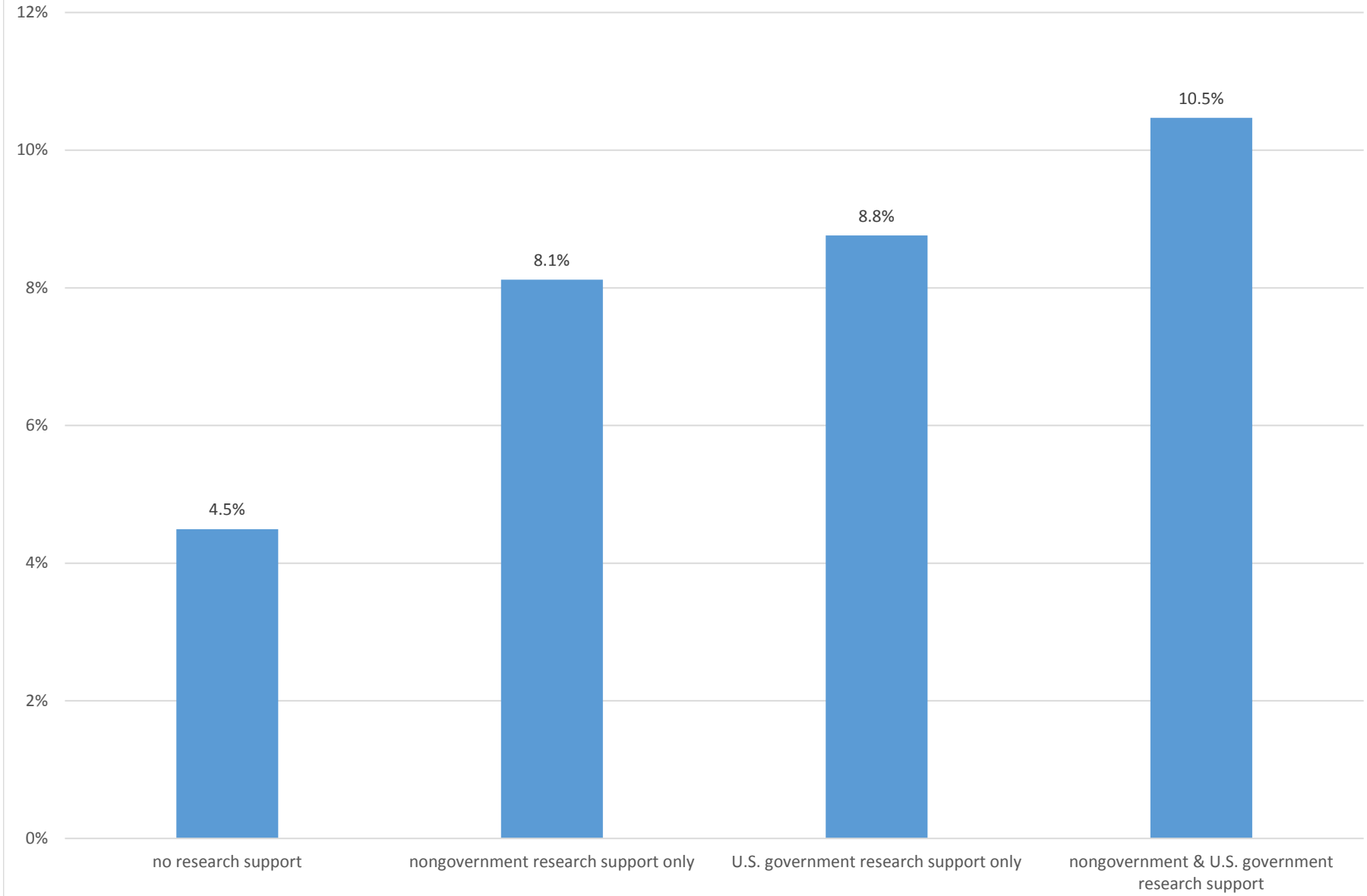


Figure 5
% of descriptors in 2013 articles that were established after 1980, by cancer site
(cancer sites with at least 10,000 descriptors in 2013)

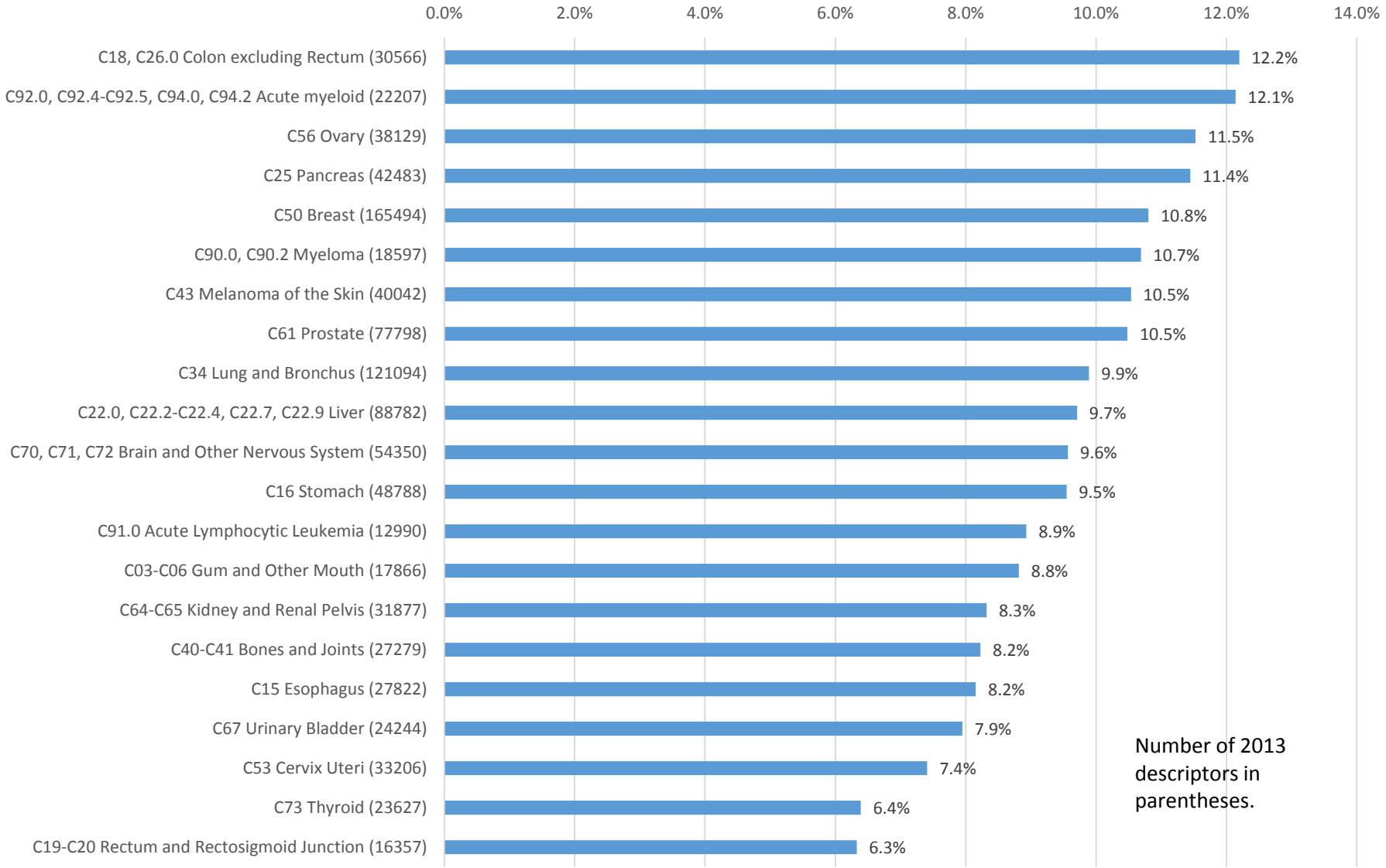


Figure 6

Estimates of β_k from the equation

$$\ln(YPLL80_{st}) = \beta_k POST1980\%_{s,t-k} + \gamma \ln(N\_DX\_10\_YEAR_{st}) + \pi AGE\_DX\_10\_YEAR_{st} + \alpha_s + \delta_t + \varepsilon_{st}$$

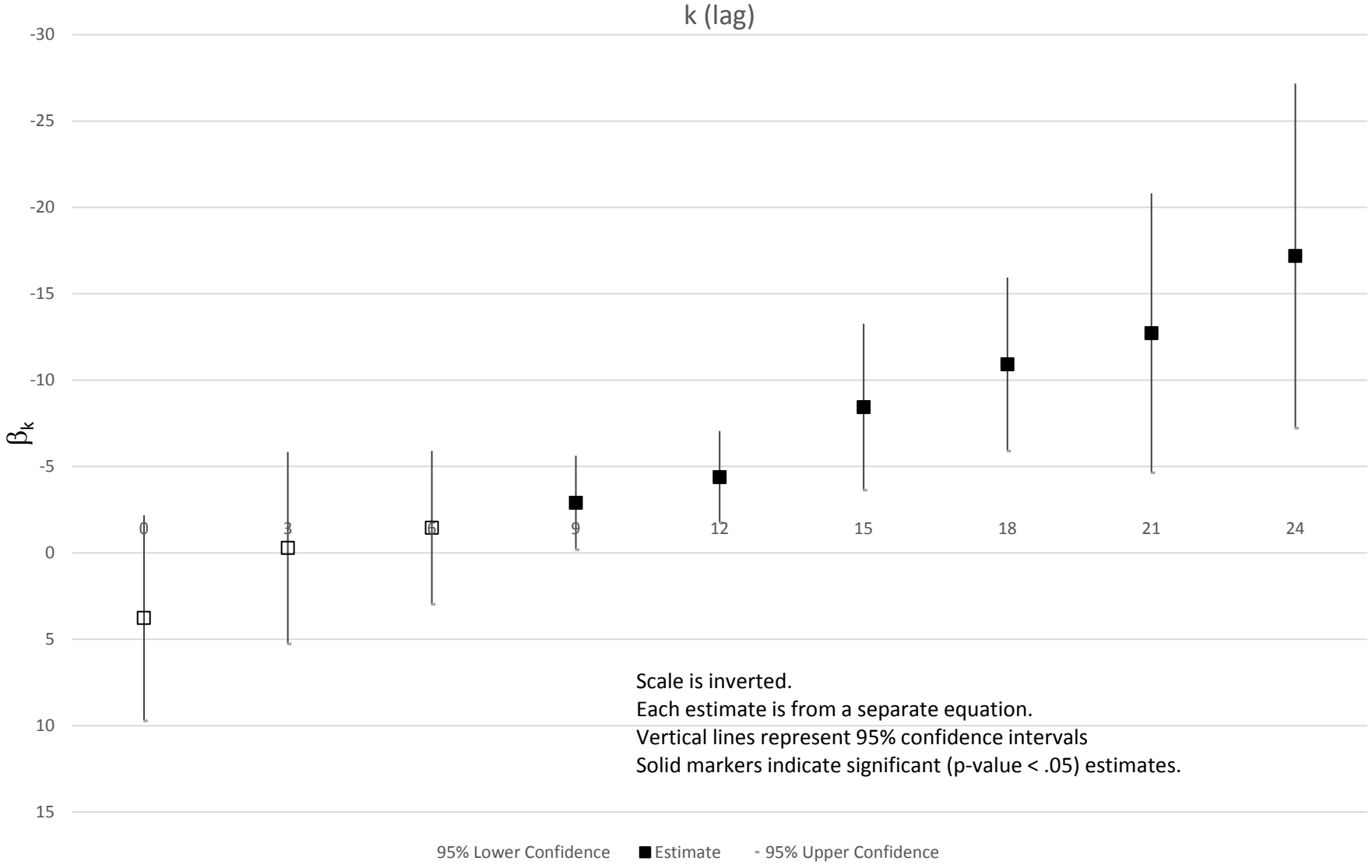
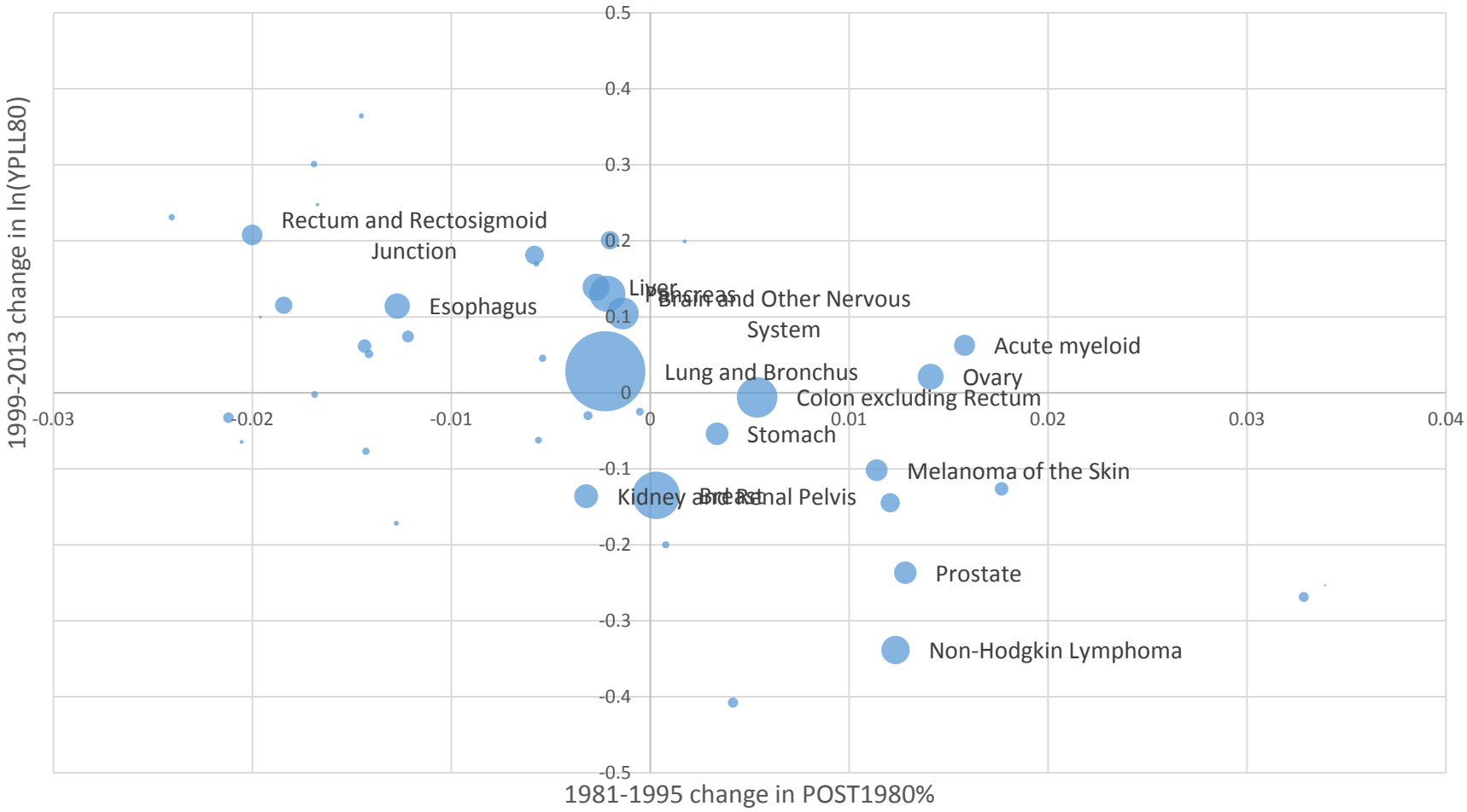


Figure 7

Relationship across cancer sites between 1981-1995 change in POST1980% and 1999-2013 change in ln(YPLL80) , controlling for changes in incidence



To improve legibility, only the largest (top 15), in terms of mean YPLL80, are labeled. Bubble size is proportional to mean(YPLL80). The chart plots the residuals from the regression of $\Delta\ln(YPLL80_{st})$ on $\Delta\ln(N\_DX\_10\_YEAR_{st})$ and $\Delta AGE\_DX\_10\_YEAR_{st}$ against the residuals from the regression of $\Delta POST1980\%_{s,t-18}$ on $\Delta\ln(N\_DX\_10\_YEAR_{st})$ and $\Delta AGE\_DX\_10\_YEAR_{st}$.

Table 1

Estimates of β_k parameters of models of the 5-year observed survival rate (eq. (1))

lag	Estimate	Std. Err.	Z	Pr > |Z|		lag	Estimate	Std. Err.	Z	Pr > |Z|
A. POST1975%; descriptor records only; all cancer sites						**B. POST1975%; descriptor records only; prostate cancer excluded**				
0	-0.275	3.512	-0.08	0.9377		0	0.602	2.908	0.21	0.836
3	2.765	3.229	0.86	0.3918		3	1.565	2.649	0.59	0.5545
6	4.433	2.932	1.51	0.1305		6	3.668	2.462	1.49	0.1362
9	4.190	2.618	1.60	0.1094		9	3.242	2.370	1.37	0.1714
12	**8.960**	**2.720**	**3.29**	**0.001**		**12**	**7.287**	**2.793**	**2.61**	**0.0091**
15	**8.979**	**2.747**	**3.27**	**0.0011**		**15**	**7.249**	**2.644**	**2.74**	**0.0061**
18	**11.563**	**3.363**	**3.44**	**0.0006**		**18**	**9.036**	**3.354**	**2.69**	**0.0071**
21	**20.192**	**4.708**	**4.29**	**<.0001**		**21**	**15.487**	**5.445**	**2.84**	**0.0045**
24	**21.569**	**6.878**	**3.14**	**0.0017**		**24**	**15.561**	**7.332**	**2.12**	**0.0338**
C. POST1975%; descriptor + supp. concept records; all cancer sites						**D. POST1975%; descriptor + supp. concept records; prostate cancer excluded**				
0	-0.251	2.628	-0.10	0.9241		0	0.291	2.199	0.13	0.8947
3	1.176	2.024	0.58	0.5614		3	0.055	1.640	0.03	0.9731
6	1.958	1.951	1.00	0.3158		6	1.558	1.535	1.02	0.3099
9	2.881	1.673	1.72	0.085		9	2.011	1.456	1.38	0.1672
12	**4.907**	**1.706**	**2.88**	**0.004**		12	3.456	1.834	1.88	0.0594
15	**5.718**	**1.697**	**3.37**	**0.0008**		**15**	**4.160**	**1.747**	**2.38**	**0.0172**
18	**7.399**	**2.195**	**3.37**	**0.0007**		**18**	**5.425**	**2.301**	**2.36**	**0.0184**
21	**10.178**	**2.931**	**3.47**	**0.0005**		21	6.138	3.771	1.63	0.1035
24	**13.893**	**4.761**	**2.92**	**0.0035**		24	8.997	6.620	1.36	0.1741
E. POST1970%; descriptor records only; all cancer sites						**F. POST1980%; descriptor records only; all cancer sites**				
0	-0.744	2.934	-0.25	0.7997		0	0.171	3.022	0.06	0.955
3	1.447	2.748	0.53	0.5985		3	3.143	3.163	0.99	0.3204
6	2.099	2.623	0.80	0.4235		6	4.893	3.130	1.56	0.1179
9	3.601	2.357	1.53	0.1267		9	5.965	3.436	1.74	0.0825
12	**6.884**	**2.851**	**2.41**	**0.0158**		**12**	**12.561**	**3.825**	**3.28**	**0.001**
15	**5.840**	**2.668**	**2.19**	**0.0286**		**15**	**15.346**	**4.613**	**3.33**	**0.0009**
18	**5.680**	**2.786**	**2.04**	**0.0414**		**18**	**23.223**	**7.814**	**2.97**	**0.003**
21	**13.919**	**2.818**	**4.94**	**<.0001**		**21**	**56.802**	**13.795**	**4.12**	**<.0001**
24	**12.608**	**3.855**	**3.27**	**0.0011**		**24**	**88.315**	**37.237**	**2.37**	**0.0177**

Each estimate is from a separate model. All models include $\ln(SURV\_EXP_{st} / (1 - SURV\_EXP_{st}))$, $\ln(N\_DX_{st})$, cancer-site and year fixed effects; coefficients on these variables are not shown to conserve space. Models were estimated via weighted-least squares, weighting by N_DX_{st}. Disturbances were clustered within cancer sites. Estimates in bold are statistically significant (p-value < .05).

Table 2

Estimates of β_k parameters of models of mortality measures (eq. (2))

lag	Estimate	Std. Err.	Z	Pr > |Z|		lag	Estimate	Std. Err.	Z	Pr > |Z|
A. YPLL80; descriptor records only						**B. YPLL80; descriptor + supp. concept records**				
0	3.770	3.039	1.24	0.2148		0	5.028	3.781	1.33	0.1835
3	-0.283	2.840	-0.10	0.9205		3	-0.281	2.114	-0.13	0.8941
6	-1.456	2.266	-0.64	0.5206		6	-1.638	1.488	-1.10	0.2709
9	**-2.904**	**1.392**	**-2.09**	**0.037**		9	-1.586	0.994	-1.60	0.1104
12	**-4.391**	**1.357**	**-3.24**	**0.0012**		**12**	**-2.404**	**0.991**	**-2.43**	**0.0153**
15	**-8.445**	**2.462**	**-3.43**	**0.0006**		**15**	**-5.152**	**1.512**	**-3.41**	**0.0007**
18	**-10.916**	**2.564**	**-4.26**	**<.0001**		**18**	**-6.960**	**1.627**	**-4.28**	**<.0001**
21	**-12.720**	**4.130**	**-3.08**	**0.0021**		**21**	**-7.907**	**2.569**	**-3.08**	**0.0021**
24	**-17.193**	**5.091**	**-3.38**	**0.0007**		**24**	**-12.500**	**3.870**	**-3.23**	**0.0012**
C. YPLL70; descriptor records only						**D. YPLL70; descriptor + supp. concept records**				
0	3.637	3.042	1.20	0.2318		0	6.893	3.986	1.73	0.0837
3	-0.840	2.946	-0.28	0.7757		3	-0.678	2.128	-0.32	0.7499
6	-1.873	2.281	-0.82	0.4116		6	-1.911	1.519	-1.26	0.2083
9	-2.729	1.477	-1.85	0.0646		9	-1.366	1.031	-1.32	0.1853
12	**-4.365**	**1.573**	**-2.78**	**0.0055**		**12**	**-2.258**	**1.099**	**-2.05**	**0.0399**
15	**-8.246**	**2.527**	**-3.26**	**0.0011**		**15**	**-4.898**	**1.679**	**-2.92**	**0.0035**
18	**-10.871**	**2.614**	**-4.16**	**<.0001**		**18**	**-6.852**	**1.754**	**-3.91**	**<.0001**
21	**-12.785**	**4.335**	**-2.95**	**0.0032**		**21**	**-7.658**	**2.745**	**-2.79**	**0.0053**
24	**-17.922**	**5.326**	**-3.37**	**0.0008**		**24**	**-12.733**	**4.136**	**-3.08**	**0.0021**
E. DEATHS; descriptor records only						**F. DEATHS; descriptor + supp. concept records**				
0	3.058	2.920	1.05	0.2949		0	-0.313	1.594	-0.20	0.8445
3	-0.008	2.589	0.00	0.9974		3	0.018	1.997	0.01	0.9927
6	-1.225	2.338	-0.52	0.6004		6	-1.370	1.578	-0.87	0.3854
9	-2.759	1.409	-1.96	0.0502		9	-1.625	1.030	-1.58	0.1146
12	**-3.787**	**1.127**	**-3.36**	**0.0008**		**12**	**-2.192**	**0.880**	**-2.49**	**0.0127**
15	**-7.239**	**2.327**	**-3.11**	**0.0019**		**15**	**-4.638**	**1.412**	**-3.29**	**0.001**
18	**-8.778**	**2.371**	**-3.70**	**0.0002**		**18**	**-5.731**	**1.461**	**-3.92**	**<.0001**
21	**-10.089**	**3.687**	**-2.74**	**0.0062**		**21**	**-7.043**	**2.467**	**-2.86**	**0.0043**
24	**-12.320**	**4.544**	**-2.71**	**0.0067**		**24**	**-9.518**	**3.354**	**-2.84**	**0.0045**

Each estimate is from a separate model. All models include $\ln(N\_DX\_10\_YEAR_{st})$, $AGE\_DX\_10\_YEAR_{st}$, cancer-site and year fixed effects; coefficients on these variables are not shown to conserve space. Models were estimated via weighted-least squares, weighting by $(1 / T) \sum_t MORT_{st}$. Disturbances were clustered within cancer sites. Estimates in bold are statistically significant (p-value < .05).

Table 3

Quantification of the long-run impact of new medical ideas on cancer survival and mortality

A. Impact on 5-year observed survival rate

k	β_k	mean(ΔPOST1975%$_{t-k}$)	β_k * mean(ΔPOST1975%$_{t-k}$)
12	9.0	5.5%	0.496
15	9.0	4.4%	0.394
18	11.6	3.0%	0.346
21	20.2	1.8%	0.373
24	21.6	1.2%	0.251
average			0.372

β_k estimates are from Panel A of Table 1.

B. Impact on number of years of potential life lost before age 80

k	β_k	mean(ΔPOST1980%$_{t-k}$)	β_k * mean(ΔPOST1980%$_{t-k}$)	(β_k * mean(ΔPOST1980%$_{t-k}$)) - pop. growth
9	-2.9	5.5%	-0.158	-0.278
12	-4.4	5.9%	-0.258	-0.377
15	-8.4	4.5%	-0.380	-0.499
18	-10.9	3.3%	-0.359	-0.479
21	-12.7	2.1%	-0.269	-0.388
24	-17.2	0.9%	-0.162	-0.282
average			-0.264	-0.384

β_k estimates are from Panel A of Table 2.

Appendix Table 1

MeSH Descriptors linked to Cancer Causes of Death

Cancer Cause of Death	ICD-10	MeSH Descriptor	MeSH Unique ID
Lip	C00	Lip Neoplasms	D008048
Tongue	C01-C02	Tongue Neoplasms	D014062
Floor of Mouth, Gum and Other Mouth	C03-C06	Mouth Neoplasms	D009062
Salivary Gland	C07-C08	Salivary Gland Neoplasms	D012468
Tonsil	C09	Tonsillar Neoplasms	D014067
Oropharynx	C10	Oropharyngeal Neoplasms	D009959
Nasopharynx	C11	Nasopharyngeal Neoplasms	D009303
Esophagus	C15	Esophageal Neoplasms	D004938
Stomach	C16	Stomach Neoplasms	D013274
Small Intestine	C17	Intestinal Neoplasms	D007414
Colon excluding Rectum	C18, C26.0	Colonic Neoplasms	D003110
Rectum and Rectosigmoid Junction	C19-C20	Rectal Neoplasms	D012004
Anus, Anal Canal and Anorectum	C21	Anus Neoplasms	D001005
Liver	C22.0, C22.2-C22.4, C22.7, C22.9	Liver Neoplasms	D008113
Intrahepatic Bile Duct	C22.1	Bile Duct Neoplasms	D001650
Gallbladder	C23	Gallbladder Neoplasms	D005706
Pancreas	C25	Pancreatic Neoplasms	D010190
Nose, Nasal Cavity and Middle Ear	C30-C31	Nose Neoplasms	D009669
Larynx	C32	Laryngeal Neoplasms	D007822
Lung and Bronchus	C34	Lung Neoplasms	D008175
Bones and Joints	C40-C41	Bone Neoplasms	D001859
Melanoma of the Skin	C43	Melanoma	D008545
Mesothelioma (ICD-10 only)+	C45+	Mesothelioma	D008654
Kaposi Sarcoma (ICD-10 only)+	C46+	Sarcoma, Kaposi	D012514
Soft Tissue including Heart$	C47, C49, C38.0, C45.2+	Soft Tissue Neoplasms	D012983
Breast	C50	Breast Neoplasms	D001943
Vulva	C51	Vulvar Neoplasms	D014846

Appendix Table 1

MeSH Descriptors linked to Cancer Causes of Death

Cancer Cause of Death	ICD-10	MeSH Descriptor	MeSH Unique ID
Vagina	C52	Vaginal Neoplasms	D014625
Cervix Uteri	C53	Uterine Cervical Neoplasms	D002583
Ovary	C56	Ovarian Neoplasms	D010051
Penis	C60	Penile Neoplasms	D010412
Prostate	C61	Prostatic Neoplasms	D011471
Testis	C62	Testicular Neoplasms	D013736
Kidney and Renal Pelvis	C64-C65	Kidney Neoplasms	D007680
Ureter	C66	Ureteral Neoplasms	D014516
Urinary Bladder	C67	Urinary Bladder Neoplasms	D001749
Eye and Orbit	C69	Eye Neoplasms	D005134
Brain and Other Nervous System	C70, C71, C72	Brain Neoplasms	D001932
Thyroid	C73	Thyroid Neoplasms	D013964
Hodgkin Lymphoma	C81	Hodgkin Disease	D006689
Non-Hodgkin Lymphoma	C82-C85, C96.3	Lymphoma, Non-Hodgkin	D008228
Myeloma	C90.0, C90.2	Multiple Myeloma	D009101
Acute Lymphocytic Leukemia	C91.0	Precursor Cell Lymphoblastic Leukemia-Lymphoma	D054198
Chronic Lymphocytic Leukemia	C91.1	Leukemia, Lymphocytic, Chronic, B-Cell	D015451
Acute myeloid	C92.0, C92.4-C92.5, C94.0, C94.2	Leukemia, Myeloid, Acute	D015470
Chronic Myeloid Leukemia	C92.1	Leukemia, Myeloid	D007951
Acute Monocytic Leukemia	C93.0	Leukemia, Monocytic, Acute	D007948

SEER Cause of Death Recode 1969+

Appendix Table 2

Observed and expected 5-year survival rates, patients diagnosed in SEER 9 registries, by cancer site, 1994

ICD10CM	Number diagnosed		Observed 5-year survival rate		Expected 5 -year survival rate	
Year	1994	2008	1994	2008	1994	2008
C61 Prostate	15,945	19,241	74.9%	86.1%	79.1%	86.6%
C50 Breast	14,020	16,359	77.3%	82.7%	89.3%	91.4%
C34 Lung and Bronchus	12,449	12,617	12.2%	16.2%	86.0%	86.3%
C18, C26.0 Colon excluding Rectum	7,266	7,289	48.0%	56.1%	79.6%	84.0%
C82-C85, C96.3 Non-Hodgkin Lymphoma	3,943	4,813	46.5%	64.3%	88.0%	88.6%
C67 Urinary Bladder	3,856	4,439	64.5%	62.7%	80.5%	81.0%
C43 Melanoma of the Skin	3,135	5,163	80.8%	84.2%	91.0%	90.2%
C19-C20 Rectum and Rectosigmoid Junction	2,812	2,942	49.9%	60.9%	83.4%	89.0%
C90.0, C90.2 Myeloma	2,673	3,174	45.5%	62.5%	88.1%	88.9%
C25 Pancreas	2,112	2,846	4.0%	6.7%	85.1%	88.4%
C64-C65 Kidney and Renal Pelvis	2,080	3,588	54.2%	66.5%	87.4%	89.5%
C16 Stomach	1,757	1,652	17.9%	27.2%	81.2%	86.0%
C56 Ovary	1,563	1,714	40.5%	42.9%	91.8%	93.0%
C70, C71, C72 Brain and Other Nervous System	1,432	1,687	31.3%	34.7%	97.2%	97.1%
C73 Thyroid	1,360	3,321	92.0%	94.2%	96.1%	96.3%
C53 Cervix Uteri	1,097	917	68.6%	66.7%	95.4%	96.2%
C46+ Kaposi Sarcoma (ICD-10 only)+	933	171	16.9%	67.9%	95.9%	92.2%
C91.1 Chronic Lymphocytic Leukemia	908	1,237	62.6%	68.7%	81.8%	81.9%
C32 Larynx	873	720	53.6%	56.7%	86.3%	88.4%
C15 Esophagus	843	1,059	10.8%	16.5%	85.6%	88.2%
C62 Testis	704	815	94.4%	95.4%	98.3%	98.5%
C22.0, C22.2-C22.4, C22.7, C22.9 Liver	688	1,719	6.2%	18.1%	89.1%	91.7%
C92.0, C92.4-C92.5, C94.0, C94.2 Acute myeloid	663	827	13.9%	22.9%	95.2%	95.5%
C81 Hodgkin Lymphoma	661	762	80.6%	87.5%	96.9%	96.8%
C03-C06 Gum and Other Mouth	549	471	50.6%	56.6%	86.6%	89.1%
C47, C49, C38.0, C45.2+ Soft Tissue including Heart$	546	759	57.9%	59.7%	90.7%	91.3%
C01-C02 Tongue	465	755	48.6%	61.8%	89.1%	91.9%
C92.1 Chronic Myeloid Leukemia	372	427	32.5%	58.2%	88.8%	89.9%
C91.0 Acute Lymphocytic Leukemia	288	394	57.6%	71.0%	98.7%	99.0%
C23 Gallbladder	253	242	13.6%	17.8%	83.6%	81.5%
C17 Small Intestine	249	480	44.0%	63.6%	86.8%	88.5%
C00 Lip	229	138	79.0%	77.5%	79.7%	83.0%
C07-C08 Salivary Gland	223	297	61.1%	67.2%	85.5%	90.4%
C45+ Mesothelioma (ICD-10 only)+	223	198	7.2%	8.1%	86.6%	82.2%
C09 Tonsil	220	459	48.2%	66.7%	92.4%	93.2%
C51 Vulva	219	269	62.1%	57.0%	81.6%	84.0%

Appendix Table 2

Observed and expected 5-year survival rates, patients diagnosed in SEER 9 registries, by cancer site, 1994

ICD10CM	Number diagnosed		Observed 5-year survival rate		Expected 5 -year survival rate	
Year	1994	2008	1994	2008	1994	2008
C40-C41 Bones and Joints	198	238	61.6%	69.0%	97.4%	96.9%
C21 Anus, Anal Canal and Anorectum	181	404	55.8%	59.6%	85.4%	92.0%
C69 Eye and Orbit	170	171	78.2%	80.9%	88.4%	91.6%
C22.1 Intrahepatic Bile Duct	167	186	3.6%	6.5%	83.9%	90.8%
C30-C31 Nose, Nasal Cavity and Middle Ear	143	169	51.7%	52.5%	86.8%	87.8%
C11 Nasopharynx	139	174	58.7%	61.7%	94.5%	95.5%
C66 Ureter	77	96	49.2%	40.6%	76.5%	82.2%
C60 Penis	74	90	63.5%	57.1%	77.3%	79.9%
C52 Vagina	61	68	36.1%	36.8%	86.7%	86.1%
C10 Oropharynx	56	90	33.9%	41.1%	91.8%	91.4%
C93.0 Acute Monocytic Leukemia	39	45	12.8%	28.1%	97.6%	95.5%

Appendix Table 3
Mortality and incidence data, by cancer site, 1999 and 2013

Cancer site	YPLL80		YPLL70		Number of deaths		Mean no. dx, previous 10 years		Mean age at dx, previous 10 years	
Year	1999	2013	1999	2013	1999	2013	1999	2013	1999	2013
TOTAL	**5,214,705**	**5,449,779**	**2,439,838**	**2,490,520**	**483,127**	**517,327**	**102,739**	**126,815**		
C34 Lung and Bronchus	1,594,468	1,520,657	644,698	598,650	152,061	156,176	15,446	17,118	68.1	69.9
C50 Breast	562,581	540,537	303,366	279,565	41,528	41,324	16,715	20,310	62.4	61.6
C18, C26.0 Colon excluding Rectum	415,537	416,003	180,102	195,683	48,962	41,963	9,394	9,235	71.3	69.7
C25 Pancreas	280,470	373,805	119,735	155,075	29,081	38,996	2,557	3,604	70.6	70.4
C70, C71, C72 Brain and Other Nervous System	248,026	271,081	152,121	158,149	12,765	15,343	1,605	1,912	51.3	52.8
C82-C85, C96.3 Non-Hodgkin Lymphoma	237,663	171,634	115,958	77,844	22,802	20,113	4,424	5,935	62.1	64.5
C56 Ovary	159,123	161,010	76,830	74,445	13,627	14,276	1,880	2,043	62.7	62.8
C15 Esophagus	141,068	170,148	63,325	74,103	11,917	14,689	1,048	1,353	67.5	68.0
C61 Prostate	134,275	131,640	32,727	39,330	31,728	27,681	17,637	20,550	70.9	66.9
C16 Stomach	132,130	129,683	63,773	64,283	12,711	11,261	2,033	2,131	69.9	68.5
C64-C65 Kidney and Renal Pelvis	130,606	151,219	62,611	68,717	11,116	13,906	2,503	4,344	63.7	63.5
C22.0, C22.2-C22.4, C22.7, C22.9 Liver	125,424	244,486	62,917	114,351	9,830	18,394	870	2,031	65.4	63.5
C43 Melanoma of the Skin	113,872	117,513	64,712	61,348	7,215	9,394	3,709	6,490	56.6	59.8
C92.0, C92.4-C92.5, C94.0, C94.2 Acute myeloid	98,445	112,558	56,202	58,278	6,932	9,712	796	1,084	61.5	62.8
C90.0, C90.2 Myeloma	92,173	93,958	36,740	35,560	10,508	11,801	1,316	1,821	69.2	68.6
C53 Cervix Uteri	89,690	90,628	56,623	55,885	4,204	4,217	1,228	1,011	50.4	50.4
C19-C20 Rectum and Rectosigmoid Junction	87,573	123,325	41,075	62,143	8,260	9,850	3,539	3,689	68.1	64.5
C67 Urinary Bladder	75,485	94,128	27,053	34,208	11,910	15,757	4,774	6,069	70.0	71.5
C47, C49, C38.0, C45.2+ Soft Tissue including Heart$	70,699	82,803	45,109	51,403	3,684	4,564	624	903	53.9	56.6
C32 Larynx	46,608	43,383	20,603	18,473	3,815	3,729	1,039	951	64.7	65.3
C91.0 Acute Lymphocytic Leukemia	46,237	40,736	35,065	29,453	1,361	1,425	346	434	23.9	25.6
C40-C41 Bones and Joints	33,968	35,551	25,115	25,453	1,224	1,453	241	303	40.4	42.8
C81 Hodgkin Lymphoma	31,889	19,658	21,494	12,433	1,403	1,090	730	824	39.8	41.9
C92.1 Chronic Myeloid Leukemia	29,191	9,825	17,744	5,358	1,788	989	333	367	60.5	58.9
C91.1 Chronic Lymphocytic Leukemia	26,765	24,428	9,005	7,975	4,476	4,657	1,025	1,546	70.3	70.5

Appendix Table 3
Mortality and incidence data, by cancer site, 1999 and 2013

Cancer site	YPLL80		YPLL70		Number of deaths		Mean no. dx, previous 10 years		Mean age at dx, previous 10 years	
Year	1999	2013	1999	2013	1999	2013	1999	2013	1999	2013
C22.1 Intrahepatic Bile Duct	26,293	63,498	11,978	28,423	2,552	5,638	180	254	70.8	69.0
C01-C02 Tongue	24,785	31,005	12,953	15,288	1,738	2,208	582	948	62.4	61.8
C45+ Mesothelioma (ICD-10 only)+	21,193	18,150	8,033	6,478	2,338	2,493	255	279	69.6	72.4
C23 Gallbladder	17,680	20,150	7,253	8,365	2,059	2,160	289	327	72.5	71.2
C03-C06 Gum and Other Mouth	15,250	13,800	7,495	6,330	1,395	1,332	688	641	65.0	66.1
C62 Testis	14,230	13,408	10,670	9,838	378	383	693	831	34.6	35.4
C73 Thyroid	12,620	17,585	5,958	7,723	1,241	1,850	1,479	3,664	47.2	49.9
C17 Small Intestine	12,403	13,310	6,218	6,085	1,036	1,270	372	645	65.5	65.5
C11 Nasopharynx	11,954	11,335	7,077	6,323	638	643	170	197	54.7	55.1
C09 Tonsil	8,780	12,815	4,590	6,170	543	839	285	517	60.3	58.5
C10 Oropharynx	7,730	12,458	3,690	5,900	600	906	71	110	63.7	62.7
C21 Anus, Anal Canal and Anorectum	6,838	13,605	3,818	7,020	462	900	269	496	63.3	61.4
C07-C08 Salivary Gland	6,708	8,923	3,293	4,243	656	886	276	375	61.3	61.5
C30-C31 Nose, Nasal Cavity and Middle Ear	6,503	6,072	3,615	3,187	456	443	167	200	63.2	62.1
C51 Vulva	4,805	8,118	2,060	3,915	762	1,003	296	387	68.6	67.3
C52 Vagina	3,505	3,740	1,760	1,658	403	437	92	114	68.0	66.7
C69 Eye and Orbit	3,321	4,005	1,983	2,100	227	319	199	241	53.9	55.5
C60 Penis	2,163	3,195	1,025	1,553	202	270	74	100	69.0	68.5
C66 Ureter	2,138	2,363	733	763	345	434	131	167	71.8	73.7
C93.0 Acute Monocytic Leukemia	1,467	1,252	815	700	136	94	55	69	58.9	58.5
C00 Lip	385	595	130	293	52	59	305	197	68.7	68.5