Online Appendix to “Are Ideas Getting Harder to Find?”

Nicholas Bloom, Charles I. Jones, John Van Reenen, and Michael Webb

October 21, 2019

1. Overview

This document provides more details on robustness checks, the data used in our paper, and the programs that can be used to replicate our results.

The programs/data for each portion of the paper are stored in a separate subdirectory. In general, the basic data are contained in spreadsheet files, and matlab programs are used to conduct the analysis. We use Matlab 2018b.

The following program can be run in the main directory where “IdeaPFPrograms.zip” is unzipped:

- **MasterIdeaPF.m:** Master program for generating all the results in the paper (other than the Census results; see Section 8 below).

Note that you will need to edit this file to change to the main directory and to add the proper path to the “ChadMatlab” directory that is unzipped from IdeaPFPrograms.zip

2. Additional Robustness Results: Alternative Wage Series for Deflating R&D

A shortcoming of using the college earnings series as our deflator of nominal R&D expenditures that the increase in college participation may mean that less talented people are attending college over time. To the extent that this is true, our deflator may understate the rise in the wage for a constant-quality college graduate and hence overstate the rise in research productivity. As an alternative, we redid all our results using two alternative deflators: first by adding 1 percent per year to the high-skilled nominal wage growth as a coarse adjustment and second using nominal GDP per person to deflate R&D expenditures — which according to the discussion surrounding equation (12) in the paper is a valid way to proceed. The results are shown in Tables 1 and 2 and are broadly similar, in part because the decreases in research productivity that we document are so large.

3. Aggregate U.S. Evidence

The analysis for the aggregate data is contained in the "Aggregate" subdirectory.

- **AggregateBLSIPP.m:** This matlab program carries out the main calculations. The NIPA data on "intellectual property products" investment are from FRED; the download codes are reported in comments in the program.

- `mfp_tables_historical-2017-02-17.xls`:Contains the BLS data on private business sector TFP growth. The contribution from intellectual property products, which is netted out of TFP growth by the BLS, is added back in, in accordance with the model.

The idea output measure is TFP growth, by decade (and for 2000-2014 for the latest observation). For the years since 1950, this measure is the BLS Private Business Sector multifactor productivity growth series, adding back in the contributions from R&D and IPP. For the 1930s and 1940s, we use the measure from Robert Gordon (2016). The idea input measure is gross domestic investment in intellectual property products from the National Income and Product Accounts, deflated by a measure of the nominal wage for high-skilled workers.

Figure 1 shows alternative measures of aggregate research effort, confirming the statement in the main text that our results are robust to how we measure aggregate research. In particular, the "NIPA IPP" series, which is the baseline series we report in the main text, and the "U.S." measure of total researchers in full-time equivalents are very similar. The OECD and OECD+ series show that if we include broader measures of research effort, the decline in aggregate research productivity would be comparable in size or larger. These results are taken from the `AggregateBLS_SciEng.m` matlab program, and the underlying data are collected in `OECD-MSTI-TotalResearchers.xls` from OECD (2018).

Table 1: Robustness: Wage Deflator with +1% Annual Adjustment

Scope	Time Period	Average annual growth rate	Half-life (years)	Dynamic Diminishing Returns, β
Aggregate economy	1930–2015	-4.1%	13	2.5
Moore's law	1971–2014	-5.8%	12	0.2
Semiconductor TFP growth	1975–2011	-4.6%	15	0.3
Agriculture, US R&D	1970–2007	-2.7%	26	1.6
Agriculture, global R&D	1980–2010	-4.5%	15	2.7
Corn, version 1	1969–2009	-8.9%	8	6.4
Corn, version 2	1969–2009	-5.2%	13	3.8
Soybeans, version 1	1969–2009	-6.3%	11	5.5
Soybeans, version 2	1969–2009	-3.4%	20	2.9
Cotton, version 1	1969–2009	-2.4%	29	1.8
Cotton, version 2	1969–2009	+2.3%	-31	-1.7
Wheat, version 1	1969–2009	-5.1%	13	5.7
Wheat, version 2	1969–2009	-2.3%	30	2.6
New molecular entities	1970–2015	-2.5%	27	...
Compustat, sales	3 decades	-10.1%	7	1.0
Compustat, market cap	3 decades	-8.2%	8	0.8
Compustat, employment	3 decades	-13.5%	5	1.7
Compustat, sales/emp	3 decades	-3.4%	21	0.8

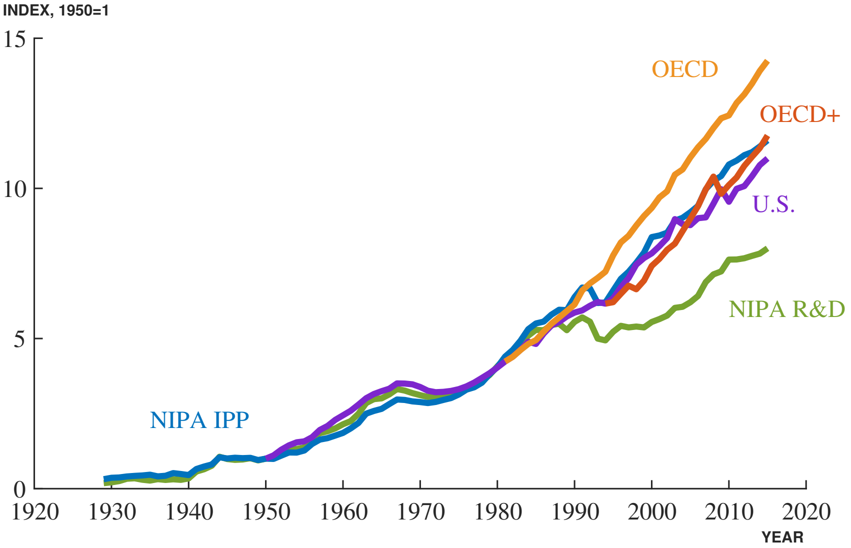
Note: This table shows robustness to using a wage deflator that grows 1 percent per year faster than the college wage series used in the main paper. See notes to Table 7 in the main paper.

Table 2: Robustness: Using Nominal GDP per Person as Wage Deflator

Scope	Time Period	Average annual growth rate	Half-life (years)	Dynamic Diminishing Returns, β
Aggregate economy	1930–2015	-4.4%	16	2.6
Moore's law	1971–2014	-5.6%	12	0.2
Semiconductor TFP growth	1975–2011	-4.4%	16	0.3
Agriculture, US R&D	1970–2007	-2.4%	28	1.5
Agriculture, global R&D	1980–2010	-4.7%	15	2.8
Corn, version 1	1969–2009	-8.7%	8	6.3
Corn, version 2	1969–2009	-5.0%	14	3.7
Soybeans, version 1	1969–2009	-6.2%	11	5.3
Soybeans, version 2	1969–2009	-3.2%	21	2.8
Cotton, version 1	1969–2009	-2.2%	32	1.6
Cotton, version 2	1969–2009	+2.4%	-28	-1.8
Wheat, version 1	1969–2009	-5.0%	14	5.6
Wheat, version	1969–2009	-2.1%	32	2.4
New molecular entities	1970–2015	-2.4%	29	...
Compustat, sales	3 decades	-10.3%	7	1.0
Compustat, market cap	3 decades	-8.4%	8	0.8
Compustat, employment	3 decades	-13.6%	5	1.7
Compustat, sales/emp	3 decades	-3.8%	18	0.9

Note: Note: This table shows robustness to using nominal GDP per person as the wage deflator instead of the college wage series. See notes to Table 7 in the main paper.

Figure 1: Alternative Measures of Aggregate Research Effort



Note: The figure shows alternative measures of aggregate research effort. The "NIPA IPP" series is the main one resported in the paper; the "NIPA R&D" series includes only U.S. R&D expenditures, also as measured by the NIPA. Both of these series are deflated by the high-skilled wage series, as described in the main text. The other three series show measures of "Total Researchers (FTE)" from the OECD Main Science and Technology Indicators, http://stats.oecd.org/ViewHTML.aspx?QueryId=58469#. The U.S. line reports researchers in the United States; data before 1981 are taken from Jones (2002). The "OECD" line plots total researchers in OECD countries since 1981, showing a 3.4-fold increase since that year. The "OECD+" line adds researchers from China and Russia to the OECD measure and reveals a 1.9-fold increase between 1994 and 2015. For visual clarity, the OECD and OECD+ lines are normalized to the U.S. value in their starting years.

4. Moore's Law

Our measurement of research spending related to Moore's Law draws primarily on two sources. First, we use the Compustat database to obtain R&D spending for more than 35 multinational companies; we are grateful to Unni Pillai for his advice and preliminary data on semiconductory R&D.[1] Second, we use the PATSTAT database (European Patent Office, 2016) to obtain the fraction of each company's patents that are in technology class "H01L" which is the class corresponding to semiconductors; we are grateful to Antoine Dechezlepretre for extensive help and computer code for extracting data from PATSTAT. Our various measures combine these data in different ways to create a measure of R&D relevant for Moore's Law. The PATSTAT database may be obtained by purchasing a copy of the bulk data from the European Patent Office at https://www.epo.org/searching-for-patents/business/patstat.html.

The spreadsheet "MooresLawRND-2018-01-08.xls" provides the basic background behind these calculations. The sheet labeled "Compustat" collates the Compustat R&D spending numbers with the (smoothed) patent shares. The sheet "PatentNarrow" provides our "narrow" measures — in which all firms research spending is weighted by their share of patents in the semiconductor class — while the sheet "PatentBroad" provides our "broad" measures — in which the research spending by focused companies like Intel or Fairchild is all included, while the research spending by conglomerates like AT&T or IBM or Toshiba is weighted according to their semiconductor patent shares.

TFP growth in the "semiconductor and related device manufacturing" industry (NAICS 334413) is taken from the NBER/CES Manufacturing Industry Database, variable "dtfp5"; see Bartelsman and Gray (1996). We smooth TFP growth using an HP filter with smoothing parameter 400 and lag R&D by 5 years in computing research productivity. In addition to the narrow/broad split, we also alternately include and exclude R&D from semiconductor equipment manufacturers: equipment is captured in a separate 6-digit industry — and therefore is perhaps most naturally excluded from our analysis. Alternatively, the pricing of the semiconductor equipment may not fully capture the benefits of that equipment, in which case the R&D from semiconductor equipment manufacturing spills over into TFP growth in the semiconductor manufacturing sector.

[1] These data are supplemented in a few cases — for example for Siemens (thanks to Dietmar Harhoff) and Samsung (thanks to Jihee Kim) — by data from company annual reports.

The following matlab programs are used:

- **MooresLaw/IntelGraph.m:** This program produces the main results for Moore's Law reported in the paper.

- **MooresLaw/SemiconductorTFP.m:** This program produces the main results for TFP growth in the semiconductor industry.

- **Patents/ReadPatentData.m:** This program reads the PATSTAT patent data and constructs the smoothed share of each firm's patenting that is in the semiconductor class. These numbers show up in the Compustat tab of the main MooresLawRND*.xls spreadsheet.

5. Agricultural Innovations

The key files are contained in the "Seeds" subdirectory:

- **Seed data v6.xlsx:** This file contains the details of the data on seed yields and research spending.

- **SeedYields.m:** This matlab program carries out the main calculations that are reported in the paper.

- **AgIdeaPF.m:** TFP growth and research productivity for the agriculture sector as a whole. Both TFP growth and U.S. R&D spending for the agriculture sector as a whole are taken from the U.S. Department of Agricultures Economic Research Service. The TFP series is smoothed with an HP filter. Global R&D spending for agriculture is taken from Fuglie, Heisey, King, Day-Rubenstein, Schimmelpfennig, Wang, Pray and Karmarkar-Deshmukh (2011), Beintema, Stads, Fuglie and Heisey (2012), and Pardey, Chan-Kang, Beddow and Dehmer (2016). Nominal R&D spending is deflated by the wage for college graduates, as described earlier. The data are collected in the spreadsheet files "AgTFP v1.xlsx" , `USDA-ERS-ag_all_research.xls`, and `GlobalRND-Agriculture.xls`.

We calculate idea input and output measures for agricultural crop yields in the United States for each of four crops: corn, soybeans, cotton, and wheat.

5.1. Idea output measure: crop yields

For each crop, we use realized average yields in the United States, measured in bushels or pounds harvested per acre planted. These data are provided by the U.S. Department of Agriculture (USDA) National Agricultural Statistics Service (NASS) (U.S. Department of Agriculture National Agricultural Statistics Service (2016)). We use yield figures provided annually at the national level. The URL to access the data is:
https://quickstats.nass.usda.gov/#C57CA751-B131-3065-9F7C-E7DE08D92F87

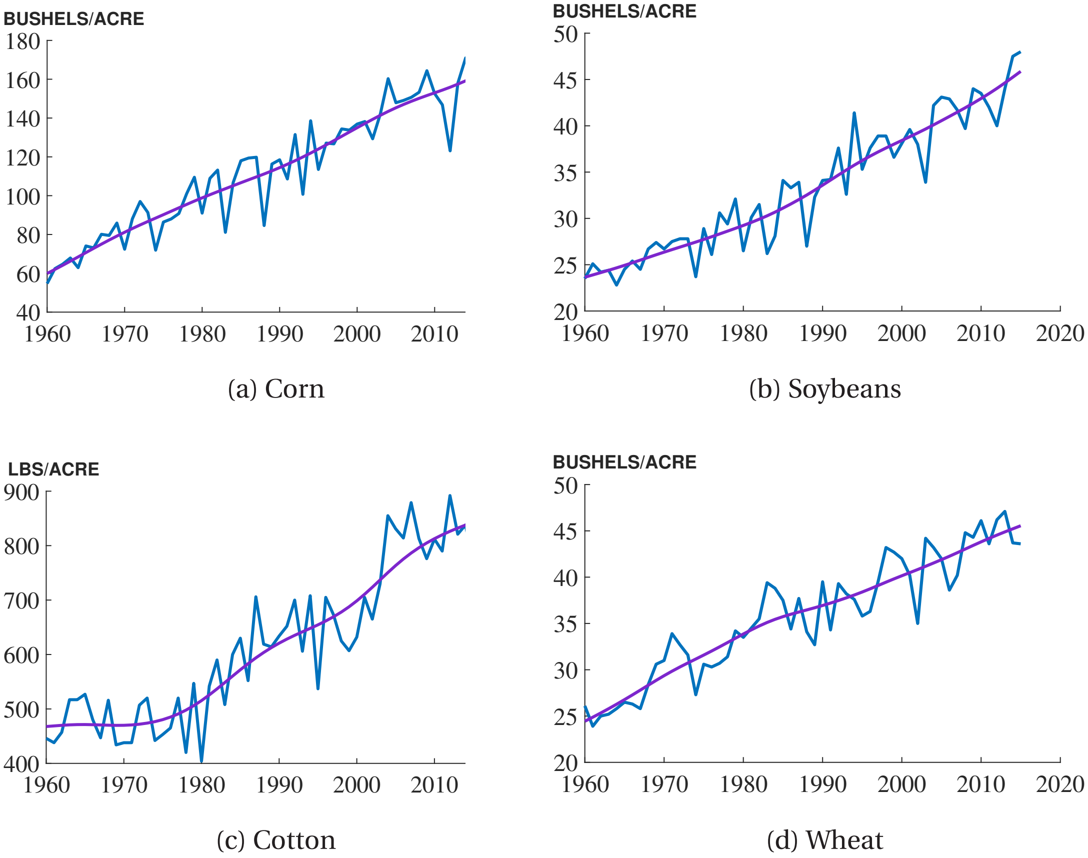
To obtain our final measure, we compute smoothed yields using an HP filter with a smoothing parameter of 400, then take an annualized average 5-year growth rate. Figure 2 shows these yields for our four crops back to the 1960s, measured in bushels or pounds harvested per acre planted. These correspond to average yields realized on U.S. farms. They are therefore subject to many influences, including choice of inputs and random shocks. These shocks, especially adverse weather and pest events, tend to have asymmetric effects: adverse events cause much larger reductions in yields than favorable events increase them, as indicated by the many large one-year reductions followed by recoveries in the figure (see Huffman, Jin and Xu (2018)). Nevertheless, yields across these four crops roughly doubled between 1960 and 2015.

5.2. Idea input measure: seed R&D

For each crop, we calculate annual R&D expenditure in the United States directed at improving that crops yields. Our data sources have three relevant dimensions: crop (corn, soybeans, cotton, and wheat), sector (public and private), and research area (biological efficiency, and crop protection and maintenance). For the private sector, we have measures of expenditure by research area that come aggregated over crops, so combine these with data on the share of a given research area devoted to a particular crop to produce an annual series of research area spending by crop. For the public sector, we have measures of total R&D by crop that come aggregated over research areas, so combine these with data on the share of a given crops total R&D devoted to a particular research area to produce an annual series of research area spending by crop. We sum across the two sectors to get an estimate for each crop-research area-year cell.

For the private sector, our measures of expenditure by research area were provided by Keith Fuglie of the U.S. Department of Agricultures Economic Research Service.

Figure 2: U.S. Crop Yields



(a) Corn

(b) Soybeans

(c) Cotton

(d) Wheat

Note: Smoothed yields are computed using an HP filter with a smoothing parameter of 400.

These series are an updated version of the series in Fuglie et al. (2011) as discussed in Fuglie (2016), and are annual from 1960 to 2015, in nominal dollars.[2] The distribution of expenditure for seed efficiency by crop are taken from Perrin et al. (1983), Fernandez-Cornejo et al. (2004), Traxler et al. (2005), and Huffman and Evenson (2006). These provide shares for the years 1960, 1965, 1975, 1979, 1982, 1989, 1994, and 2001. No direct data are available on distribution of expenditure for crop protection by crop. As such, we took 3 related measures (crop protection sales shares from University of York (2016), the public crop protection shares described below, and the private seed efficiency shares described above) and took the average. We use linear interpolation where required to fill in missing years.

For the public sector, we begin with two comparable raw series of R&D by crop, covering different time periods, that each come aggregated over research areas. One is taken from annual versions of Table C from the U.S. Department of Agriculture National Institute of Food and Agriculture Current Research Information System (CRIS) Funding Summaries, and covers the years 1993-2015. The other is from Huffman and Evenson (2006), covering 1969, 1984, and 1997. The year 1997 is thus an overlapping year. CRIS figures are in nominal dollars; H&E figures are deflated by a price index. When we use this index to un-deflate the H&E series to get back nominal figures, the CRIS amounts for the overlapping year (1997) are close to 60% of the un-deflated H&E figures for all four crops. As such, we use the CRIS numbers for all the years available (1993-2015), and multiply the un-deflated H&E figures for the years 1969 and 1984 by this 'splicing factor' to get a consistent nominal series. The distribution of expenditure by research area for each crop is taken from Huffman and Evenson (2006). This provides shares for the years 1969, 1984, and 1997. We use linear interpolation where required to fill in missing years. We use the output from this methodology for spending on biological efficiency; for spending on biological efficiency and crop protection and maintenance combined, we use a new series provided by Huffman of absolute productivity-directed public research by crop for the years 1960-2009. This series is, as expected, very close to the equivalent series generated using the methodology just outlined.

To obtain our final measure of idea inputs, we deflate the summed private and public annual series using a measure of the average annual earnings for people with

[2] Private communication with Michael Webb, July 25, 2016.

4 or more years of college, for reasons explained in section 3.

6. Medical Innovations

Programs and data for the disease measures are in the “Mortality” subdirectory, while those for the new molecular entities are in the “Pharma” subdirectory.

- **Cancer.m:, BreastCancer.m:, HeartDisease.m:** These are the main programs that carry out the calculations for the three diseases.
- **mortality.m:** This function is called to do the heavy lifting.
- **LifeExpectancy.m:** Create the life expectancy graph in Figure 7.
- **BasicLifeTable.m:** Reads the basic life tables from Mortality.org for all people.
- **BasicLifeTableWomen.m:** Reads the basic life tables from Mortality.org for women.
- **NMEGraph.m:** The basic program for generating the results for new molecular entities.
- **NME-Since1938.xls:** Data on new molecular entities since 1938, from http://www.fda.gov/AboutFDA/WhatWeDo/History/ProductRegulation/SummaryofNDAApprovals Also the R&D data from various issues of “Pharmaceutical Industry Profile”; see http://www.phrma.org/sites/default/files/pdf/PhRMA%20Profile%202013.pdf

Our measures of life expectancy and mortality from all sources by age come from the Human Mortality Database at http://mortality.org. To measure the percentage declines in mortality rates from cancer, we use the age-adjusted mortality rates for people ages 50 and over computed from 5-year survival rates, taken from the National Cancer Institutes Surveillance, Epidemiology, and End Results program (2015) at http://seer.cancer.gov/. For heart disease, we report the crude death rate in each year for people aged 55–64.

For our research input, we measure the number of scientific publications in PUBMED (2016) that have “Neoplasms” or “Breast Cancers” or “Heart Diseases”, as a MESH (Medical Subject Heading) term. MESH is the National Library of Medicine’s controlled vocabulary thesaurus. For more information on MESH, see https://www.nlm.nih.gov/mesh/. Our queries of the PUBMED data use the webtool MEDSUM (2016) created by the

Institute for Biostatistics and Medical Informatics (IBMI) Medical Faculty, University of Ljubljana, Slovenia available at http://webtools.mf.uni-lj.si/.

6.1. New Molecular Entities

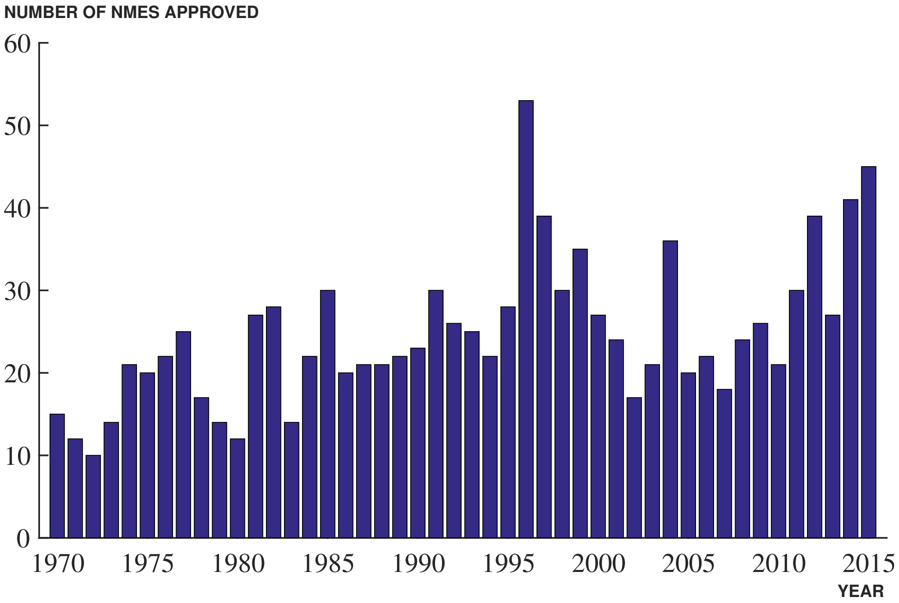
Our first example in the text of the paper from the medical sector is a fact that is well-known in the literature, recast in terms of research productivity. Here we report the details.

Figure 3 shows the number of new molecular entities ("NMEs") approved by the Food and Drug Administration (FDA). NMEs are compounds that emerge from the process of medicine discovery, that are not a version or derivative of an existing, previously investigated/approved substance. They are new drugs that include both chemical and biological products and virtually all pharmaceutical advances in the last 50 years show up in these counts (Zambrowicz and Sands, 2003). Famous examples that became commercial blockbuster drugs are Zocor (for cholesterol), Prilosec (for gastroesophagal reflux), Claritin (for allergies), Celebrex (for arthritis), and Taxol (for treating various types of cancer). Only two or three of the NMEs in any given year become commercial successes. Among famous drugs, only morphine and aspirin do not show up in these counts, because their discovery pre-dates the FDA. The flow of NMEs is well-known to show very little trend, although 2014 and 2015 are two of the years with the most approvals. Based on this fact, we proceed conservatively and measure idea output as the flow of NMEs rather than as the percentage change.

We obtain data on pharmaceutical R&D spending from the Pharmaceutical Research and Manufacturers of America (Phrma), which has conducted an annual survey of its members back to 1970 and includes R&D performed both domestically and abroad by these companies.[3] Using the procedures described earlier, we get the research productivity and effective research numbers shown in Figure 4. Research effort rises by a factor of 9, while research productivity falls by a factor of 11 by 2007 before rising in recent years so that the overall decline by 2014 is a factor of 5. Over the entire period, research effort rises at an annual rate of 6.0 percent, while research productivity falls at an annual rate of 3.5 percent. It is well documented that the number of NMEs per

[3] A limitation is that it does not include R&D done by foreign companies that is performed abroad. However, Figure 1 of Congressional Budget Office (2006) suggests that this is still a very useful measure.

Figure 3: New Molecular Entities Approved by the FDA



Note: Historical data on NME approvals are from Food and Administration (2013). Data for recent years are taken from Pharmaceutical Research and Manufacturers of America (2016).

dollar of R&D is declining; our statement is different in that, importantly, our measure of research input is deflated by the wage of college graduates. Akcigit and Liu (2016) examine this case in more detail and suggest that the rising replication of dead-ends in pharmaceutical research (and elsewhere) could be part of the story.

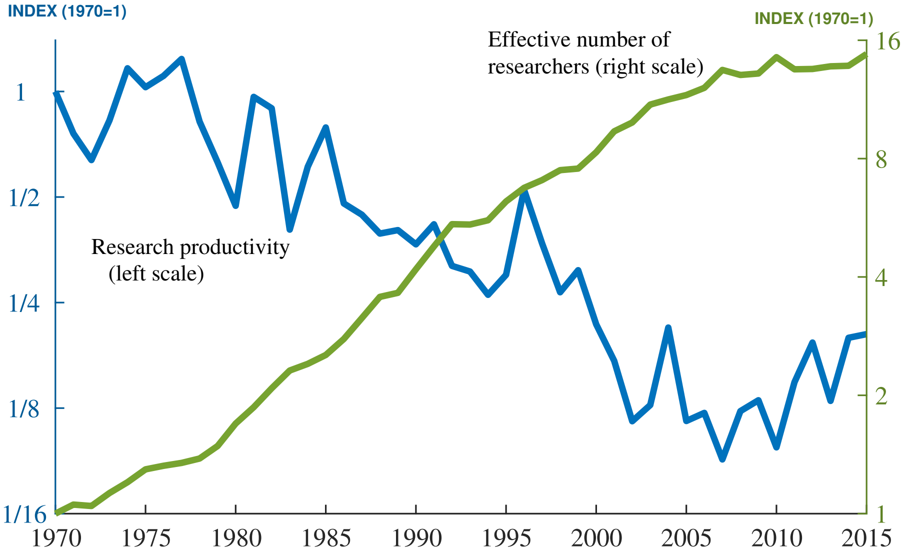
Of course, it is far from obvious that simple counts of NMEs appropriately measure the output of ideas; we would really like to know how important each innovation is.[4] In addition, the NMEs still suffer from an important aggregation issue, adding up across a wide range of health conditions. These limitations motivate the main approach in the paper, in which we focus on the productivity of medical research in specific diseases.

7. Compustat Firm-Level Results

These programs and files are in the "Compustat" subdirectory.

[4] An alternative source of information on pharmaceuticals would be clinical trial data. These are available in the rich Cortellis dataset used in Krieger (2017). Unfortunately, the data is only reliable after the mid 1990s, so it is not suitable to use over long periods of time, which is our purpose in this paper.

Figure 4: Research Productivity for New Molecular Entities



Note: Research productivity is the ratio of idea output, NME approvals, to the effective number of researchers, measured as R&D expenditures deflated by the nominal wage for high-skilled workers. Historical data on NME approvals are from Food and Administration (2013). Data on research spending by the pharmaceutical industry are from the 2010, 2013, and 2016 editions of Pharmaceutical Research and Manufacturers of America (2016). See the online data appendix for more details.

- **Compustat-WRDS-2016-06-13.xlsx, Compustat-WRDS-2016-06-13.csv:** Basic data file downloaded from Compustat via WRDS.

- **CompustatRead.m:** Reads the downloaded data from Compustat-WRDS-2016-06-13.csv.

- **MasterCompustat.m:** The master program for the Compustat results, including robustness.

- **CompustatIdeaPF.m:** The basic program that does the heavy lifting, given a set of parameters and assumptions.

- **SetParameters.m:** Sets the baseline parameter values.

- **ShowParameters.m:** Reports the parameter values.

- **GDPDeflator.m:** Loads and saves the basic GDP Deflator used to deflates sales revenue and market cap.

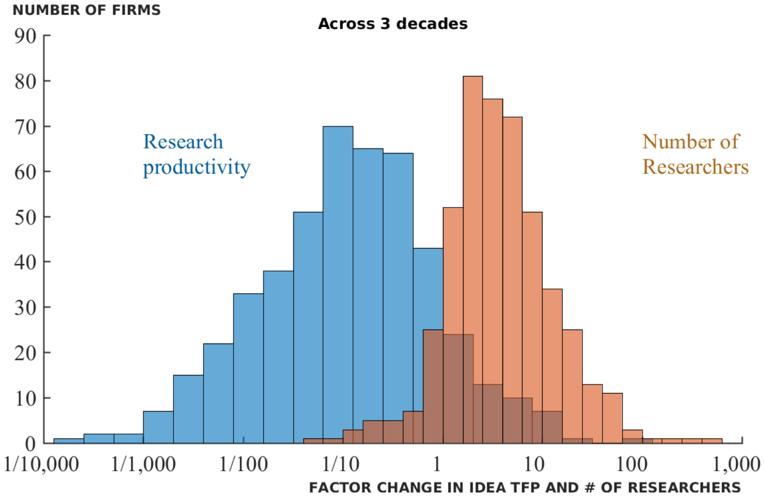
- **compugrowthrate.m:** A function for computing various growth rates.

As a measure of the output of the idea production function, we use decadal averages of annual growth in sales revenue, market capitalization, employment, and revenue labor productivity within each firm. Sales revenue and market cap are deflated by the GDP implicit price deflator. We take the decade as our period of observation to smooth out fluctuations.

To measure the research input, we use a firm's spending on research and development from Compustat. This means we are restricted to publicly-listed firms that report formal R&D, and such firms are well-known to be a select sample (e.g. disproportionately in manufacturing and large). We look at firms since 1980 that report non-zero R&D, and this restricts us to an initial sample of 15,128 firms. Our additional requirements for sample selection in our baseline sample are

1. We observe at least 3 annual growth observations for the firm in a given decade. These growth rates are averaged to form the idea output growth measure for that firm in that decade.

Figure 5: Compustat Distributions, Sales Revenue (3 Decades)



Note: Based on 469 firms. 11.9% of firms have increasing research productivity. Only 4.3% have research productivity that is roughly constant, defined as a growth rate whose absolute value is less than 1% per year.

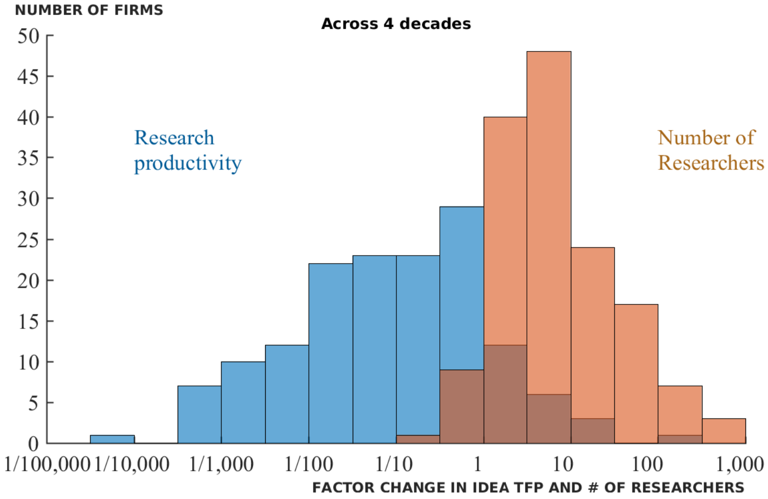
2. We only consider decades in which our idea output growth measure for the firm is positive (negative growth is clearly not the result of the firm innovating, and our framework cannot make sense of negative research productivity).

3. We require the firm to be observed (for both the output growth measure and the research input measure) for two consecutive decades. Our decades are the 1980s, the 1990s, the 2000s (which refers to the 2000-2007 period), and the 2010s (which refers to the 2010-2015 period); we drop the years 2008 and 2009 because of the financial crisis.

We relax many of these conditions in our robustness checks.

Figures 5 and 6 demonstrate the heterogeneity across firms in our Compustat sample by showing the distribution of the factor changes in effective research and research productivity across all the firms Figures 5 and 6 shows the distributions for the firms observed for three and four decades; the distribution for firms observed for two decades is shown in the main paper.

Figure 6: Compustat Distributions, Sales Revenue (4 Decades)



Note: Based on 149 firms. 14.8% of firms have increasing research productivity. 4.7% firms in this sample have research productivity that is roughly constant, defined as a growth rate whose absolute value is less than 1% per year.

8. Firm-Level Results with U.S. Census Data

We used all firms that reported manufacturing shipments (sales) in the Economic Censuses of Manufacturers (CMF) of 1982, 1992, 2002 and 2012, as well as positive R&D expenditure in the BRDIS (or SIRD before 2008) surveys of R&D in at least one year in each decade. The Census of Manufacturers surveys around 250,000 manufacturing establishments (distinct geographical locations), covering what the Census believes is the population of US manufacturing establishments. The Census itself builds this population file from a combination of prior years surveys, the survey of business organization (which asks firms to report all current production locations) and IRS tax returns.

The Business Research and Development and Innovation Survey (BRDIS) is a firm-level survey that is sent to the population of firms the Census believes undertakes R&D (based from prior years data, patenting records etc) plus a random sample of all other firms. Given the skewed nature of R&D expenditure — most R&D is carried out by a few well identified large firms — this should capture the large majority of R&D expenditure. This data spanned 1,300 firms over 2,700 firm-year observations (where numbers have been rounded for disclosure purposes). All Census data was accessed in the Stanford Census Research Data Center (RDC) by Nicholas Bloom working with generous research assistance from Brian Lucking.

As discussed in the text there are a number of differences between US Census data and US Compustat date. First, the Census uses the responses to official (mandatory) government surveys, rather than audited financial accounts. So this could potentially induce some measurement error (if the auditing process helps to eliminate recording errors for example), but it also helps to address concerns over potential bias in reported accounting data for publicly-listed firms (e.g. if firms manipulate their reported R&D activity to influence their stock market valuation). Second, the Census covers the activity of all firms operating in the U.S. public and private including the subsidiaries of foreign multinationals. Thus, smaller firms, start-ups and subsidiaries of overseas firms are included so long as they are covered by the BRDIS survey. Third, the Census data excludes the R&D and sales activities of U.S. firms abroad, which for large manufacturing firms is often substantial. By contrast, Compustat reports the global consolidated accounts, so overseas sales and R&D will be included in the totals. Fourth, the Census also collects data on the number of scientists and engineers engaged in R&D activity, providing a quantity measure of innovation inputs. Finally, the Census compares the figures for large firms against administrative data e.g. IRS tax returns and social security filings helping to ensure data accuracy. This means, for example, figures in the CMF or BRDIS that appear implausible given IRS tax returns or other Census data will be checked and edited.

Full replication of our Census results can be carried out using the **"itfp.do"** file, which runs in Stata 16 on the RDC Census data. To run this requires having an eligible Census project which grants access to the CMF and BRDIS. Those interested in applying for an RDC Project should visit https://www.census.gov/fsrdc.

9. The Wage Series for Deflating R&D Spending

The program and data for this series are stored in the subdirectory "WageSci."

- **WageEducation.m:** Reads the wage data and creates WageScientistData.mat, which is used in many other programs.

- **WageEducationPlus1.m:** Adds 1 percent per year to the wage growth in WageEducation.m for robustness.
- **WageNominalGDP.m:** Uses nominal GDP per person as the wage deflator for R&D expenditures, for robustness.

As our benchmark measure of the nominal wage in our empirical applications, we use mean personal income from the Current Population Survey (2016) for males with a Bachelor's degree or more of education. These data are from Census Tables P18 and P19, available at http://www.census.gov/topics/income-poverty/income/data/tables.html. Prior to 1991, we use the series for "4 or more years of college." For years between 1939 and 1967, we use the series Bc845 from the Historical Statistics for the U.S. Economy, Millennial Edition (2006). Finally, for the aggregate research productivity calculations, we require a deflator from the 1930s. We extrapolate the college earnings series backward into the 1930s using nominal GDP per person for this purpose. As an alternative, we have redone our results using nominal GDP per person as the deflator; this yielded broadly similar results; see Tables 1 and 2 earlier in this document.

References

Akcigit, Ufuk and Qingmin Liu, "The Role Of Information in Innovation and Competition," *Journal of the European Economic Association*, August 2016, *14* (4), 828–870.

Bartelsman, Eric J. and Wayne Gray, "The NBER Manufacturing Productivity Database," Working Paper 205, National Bureau of Economic Research October 1996.

Beintema, Nienke, Gert-Jan Stads, Keith Fuglie, and Paul Heisey, "ASTI Global Assessment of Agricultural R&D Spending," Technical Report, International Food Policy Research Institute 2012.

Carter, Susan B., Scott S. Gartner, Michael R. Haines, Alan L. Olmstead, Richard Sutch, and Gavin Wright, *Historical Statistics of the United States: Millennial Edition*, Cambridge University Press, 2006. https://hsus.cambridge.org/HSUSWeb/toc/hsusHome.do (accessed June 8, 2016).

Congressional Budget Office, "Research and Development in the Pharmaceutical Industry," Technical Report October 2006.

European Patent Office, “EPO Worldwide Patent Statistical Database (PATSTAT),” 2016. https://www.epo.org/searching-for-patents/business/patstat.html.

Fernandez-Cornejo, Jorge et al., “The seed industry in US agriculture: An exploration of data and information on crop seed markets, regulation, industry structure, and research and development,” Technical Report, United States Department of Agriculture, Economic Research Service 2004.

Food and Drug Administration, “Summary of NDA Approvals & Receipts, 1938 to the present,” 2013. Online data report.

Fuglie, Keith, “The Growing Role of the Private Sector in Agricultural Research and Development World-wide,” *Global Food Security*, 2016, *10*, 29–38.

_ , Paul Heisey, John L King, Kelly Day-Rubenstein, David Schimmelpfennig, Sun Ling Wang, Carl E Pray, and Rupa Karmarkar-Deshmukh, “Research investments and market structure in the food processing, agricultural input, and biofuel industries worldwide,” *USDA-ERS Economic Research Report*, 2011, (130).

Gordon, Robert J., *The Rise and Fall of American Growth: The US Standard of Living since the Civil War*, Princeton University Press, 2016.

Huffman, Wallace E and Robert E Evenson, *Science for agriculture: A long-term perspective*, John Wiley & Sons, 2006.

Huffman, Wallace E., Yu Jin, and Zheng Xu, “The economic impacts of technology and climate change: new evidence from U.S. corn yields,” January 2018. Working Paper.

Jones, Charles I., “Sources of U.S. Economic Growth in a World of Ideas,” *American Economic Review*, March 2002, *92* (1), 220–239.

Krieger, Joshua, “Trials and Terminations: Learning from Competitors’ R&D Failures,” 2017. M.I.T. unpublished manuscript.

MEDSUM, 2016. An online MEDLINE summary tool by Galsworthy, MJ. Hosted by the Institute of Biomedical Informatics (IBMI), Faculty of Medicine, University of Ljubljana, Slovenia. www.medsum.info (accessed August 18, 2016).

National Cancer Institute, “Surveillance, Epidemiology, and End Results Program,” 2015. https://seer.cancer.gov/ (accessed November 18, 2015).

OECD, “Main Science and Technology Indicators,” 2018. http://stats.oecd.org/ViewHTML.aspx?QueryId=58469# (accessed February 19, 2018).

Pardey, Philip G., Connie Chan-Kang, Jason M. Beddow, and Steven P. Dehmer, "Shifting Ground: Food and Agricultural R&D Spending Worldwide, 1960-2011," Technical Report, International Science and Technology Practice and Policy (InSTePP) Center, University of Minnesota 2016.

Perrin, Richard K, KA Kunnings et al., "Some effects of the US Plant Variety Protection Act of 1970.," *North Carolina State University. Dept. of Economics and Business. Economics research report (USA). no. 46.*, 1983.

Pharmaceutical Research and Manufacturers of America, *2016 Biopharmaceutical Research Industry Profile*, Washington, DC: PhRMA, 2016.

PUBMED, 2016. US National Library of Medicine National Institutes of Health, https://www.ncbi.nlm.nih.gov/pubmed/ (accessed August 18, 2016).

Traxler, Greg, Albert KA Acquaye, Kenneth Frey, and Ann Marie Thro, "Public sector plant breeding resources in the US: Study results for the year 2001," *USDA Cooperative State Research, Education and Extension Service*, 2005, pp. 1–7.

University of York, "The Essential Chemical Industry," 2016. http://www.essentialchemicalindustry.org/ (accessed September 8, 2016).

U.S. Bureau of the Census, "Current Population Survey," 2016. https://www.census.gov/programs-surveys/cps.html (accessed June 8, 2016).

U.S. Department of Agriculture National Agricultural Statistics Service, 2016. https://quickstats.nass.usda.gov/#C57CA751-B131-3065-9F7C-E7DE08D92F87 (accessed September 8, 2016).

Zambrowicz, Brian P. and Arthur T. Sands, "Knockouts Model the 100 Best-Selling Drugs — Will they model the next 100?," *Nature Reviews Drug Discoveries*, January 2003, *2* (1), 38–51.