NBER WORKING PAPER SERIES

DEMAND SHOCKS, PROCUREMENT POLICIES, AND THE NATURE OF MEDICAL INNOVATION: EVIDENCE FROM WARTIME PROSTHETIC DEVICE PATENTS

Jeffrey Clemens
Parker Rogers

Working Paper 26679
http://www.nber.org/papers/w26679

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
January 2020

We thank Joshua Chan and Yutong Wu for excellent research assistance. Many thanks to Guy Hasegawa for his generous assistance in sending us copies of archival materials used for his book "Mending Broken Soldiers." Thanks also to Rosemary Stevens and Rich Meckel for providing valuable perspective on the historical episodes we analyze. We also thank Dave Chan, Julie Cullen, Gordon Dahl, Michael Dickstein, Christian Dippel, Itzik Fadlon, Alex Gelber, Michela Giorcelli, Roger Gordon, Kate Ho, Neale Mahoney, Markus Nagler, Karthik Muralidharan, Elena Patel, Julian Reif, Kaspar Wuthrich, and seminar participants at the 2018 AEI Economists Roundtable, the 2019 Junior Health Economics Summit, the 2019 SIEPR Post-Doc Conference, the 2019 NTA Meetings, UNLV, and the Center for Economic Studies in Munich. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Demand Shocks, Procurement Policies, and the Nature of Medical Innovation: Evidence from Wartime Prosthetic Device Patents

Jeffrey Clemens and Parker Rogers

NBER Working Paper No. 26679
January 2020
JEL No. H57,I1,O31

ABSTRACT

We analyze wartime prosthetic device patents to investigate how procurement policy affects the cost, quality, and quantity of medical innovation. Analyzing whether inventions emphasize cost and/or quality requires generating new data. We do this by first hand-coding the economic traits emphasized in 1,200 patent documents. We then train a machine learning algorithm and apply the trained models to a century's worth of medical and mechanical patents that form our analysis sample. In our analysis of these new data, we find that the relatively stingy, fixed-price contracts of the Civil War era led inventors to focus broadly on reducing costs, while the less cost-conscious procurement contracts of World War I did not. We provide a conceptual framework that highlights the economic forces that drive this key finding. We also find that inventors emphasized dimensions of product quality (e.g., a prosthetic's appearance or comfort) that aligned with differences in buyers' preferences across wars. Finally, we find that the Civil War and World War I procurement shocks led to substantial increases in the quantity of prosthetic device patenting relative to patenting in other medical and mechanical technology classes. We conclude that procurement environments can significantly shape the scientific problems with which inventors engage, including the choice to innovate on quality or cost.

Jeffrey Clemens
Department of Economics
University of California, San Diego
9500 Gilman Drive #0508
La Jolla, CA 92093
and NBER
jeffclemens@ucsd.edu

Parker Rogers
Department of Economics
University of California, San Diego
9500 Gilman Drive #0508
La Jolla, CA 92093
United States
parogers@ucsd.edu

From 1960 to 2018, U.S. health spending rose from 5 to nearly 18 percent of GDP. Research documents that the advance of medical innovation underlies a substantial share of this cost growth (Smith, Newhouse, and Freeland, 2009; Cutler, 2004). A key question, then, is why medical innovation has tended to bring cost-increasing enhancements to quality rather than cost-reducing advances in productivity.

Expensive technologies pose dilemmas related to health system cost, access, and equity (Chandra and Skinner, 2012; Shepard, Baicker, and Skinner, 2019). Treatments including proton beam therapy, the cancer drug Avastin, and the hepatitis C drug Sovaldi, provide striking illustrations. Each treatment is used regularly in the United States. By contrast, the British National Health Service (NHS) has strongly limited their coverage. This reflects the NHS's assessment that these treatments are too costly to finance without severe restrictions.[1] This coverage dilemma prompts us to ask whether policy can shape the amount of costly versus cost-reducing innovation in which inventors engage.

We analyze the effects of incentives on both the quantity and cost-consciousness of medical innovation. We first show that health care payment models can, at least in theory, play a role in steering inventor efforts between quality- and cost-oriented innovation. Our empirical analysis considers two important periods in the history of prosthetic device innovation, namely the U.S. Civil War and World War I. We show that both wars led to substantial increases in prosthetic device patenting. A key point of contrast is that the Civil War led to a much larger rise in cost-conscious innovation. To the best of our

[1]Early estimates implied that Avastin, for example, could extend life by several months at a cost exceeding $50,000 per year of treatment (Kolata and Pollack, 2008). This fails the NHS's cost-effectiveness thresholds. Avastin's coverage has thus come primarily through the Cancer Drugs Fund, which temporarily financed payments for a number of costly drugs in spite of the system's cost-effectiveness thresholds (Hawkes, 2015). While Sovaldi was recommended positively by the UK's National Institute for Health and Care Excellence, the NHS initially capped use to 10,000 patients per year, which was far below need (Boseley, 2016). With regards to proton beam therapy, Limb (2019) writes that "Since 2008, some 1,400 patients have been referred to hospitals in the US and Europe under an NHS overseas treatment programme that funds treatment, transport, and accommodation." The construction of major proton beam treatment centers in the UK itself is a relatively recent development.

knowledge, this analysis provides the first evidence that cost-conscious payment models can indeed steer medical innovation in a cost-conscious direction.

We begin by developing a framework to show how payment models can shape incentives for inventors to improve product quality or reduce costs.[2] The framework's first implication is straightforward: cost-based payments, which protect suppliers against cost overruns, provide far greater incentives for quality-enhancing innovation than for cost-conscious innovation. Fixed-price payments, which do not adjust with realized costs, are the standard alternative to cost-based payments. We explore the conditions under which fixed-price payments will tilt incentives in favor of cost-conscious innovation. Our framework's most novel insight is that the effects of fixed-price payment models depend crucially on market structure and on how high payments are set. When payments are *low*, fixed-price payment models will tend to steer innovation in a cost-conscious direction. For example, if a payment is set below a firm's baseline costs, then the firm must innovate to reduce costs before sales become profitable. By contrast, we show that fixed-price payment models with *high* payment rates can lead inventors to focus on quality-enhancing innovation: high payments make sales highly profitable, which can lead firms to compete for market share based on their products' quality.

Empirically assessing how incentives shape the emphases of inventors requires overcoming two primary challenges. First, existing data sources that categorize patents or clinical trials do not provide information on an invention's detailed economic attributes. Extracting this information requires going deeper into an invention's details. Second, linking incentives to the specific attributes on which inventors focus requires analyzing settings across which incentives vary meaningfully.

To gain insight into how inventors advanced the frontier of prosthetic device technology, we use machine learning tools to construct a novel data set. We begin by closely

[2]The framework applies insights from Rogerson (1989, 1994) and from Laffont and Tirole (1986).

reading 1,200 patents from the periods surrounding the U.S. Civil War and World War I. Our selection is comprised of prosthetic device patents and patents from other medical and mechanical technology classes. Based on these close readings, we code variables describing the economic traits emphasized in each patent. These variables include a set of four traits that we interpret as cost-conscious attributes and two traits that capture dimensions of product quality. We then use machine learning tools to extend our data set to include a much larger set of patents.

We analyze two striking episodes in the history of prosthetic device innovation, namely the U.S. Civil War and World War I. These episodes were associated with dramatic increases in demand, as amputations were remarkably common. Our empirical analysis takes a standard difference-in-differences structure. We compare pre-war patenting within out treatment and control groups to patenting during and after the wars. We use other medical and mechanical technology classes to establish control groups for our treated prosthetic device class.

Our first result quantifies the effects of the Civil War and World War I on the quantity of prosthetic device innovation. For several years during each episode, prosthetic device patenting was elevated by nearly 100 log points relative to patenting in our control groups. Despite analyzing only two "treatment events," the relative increases in prosthetic device patenting are quite strongly statistically distinguishable from zero.

Second, we find that the demand shock associated with the Civil War generated substantial effort to reduce the cost of producing prosthetic devices. During the Civil War period, the average prevalence of our four cost-oriented traits temporarily doubled in prosthetic device patents but was essentially flat within our control groups. This high degree of cost-oriented innovation was plausibly driven by the design of the U.S. government's procurement program, through which manufacturers received low, fixed-price payments.

Third, prosthetic device patents exhibit an increased emphasis on traits connected to the mass production of prosthetic devices during both wars. That is, wartime patents suggest shifts away from bespoke prosthetic limbs. This common shift in emphases is consistent with a role for economies of scale within the supply chain.

Finally, the prosthetic device patents of the Civil War and World War I episodes diverged with respect to dimensions of product quality. Civil War-era prosthetic device patents exhibit a substantial increase in emphasis on comfort. By contrast, World War I-era prosthetic device patents de-emphasize comfort and exhibit an increase in emphasis on appearance. These differences are plausibly, though not definitively, linked to a World War I-era shift in choice away from veterans and towards medical professionals. This shift was accompanied by a heightened emphasis (by both the government and medical professionals) on the re-employment and social re-integration of amputee veterans.

Our analysis contributes to several lines of research. First, an important line of research studies the effects of market size on the pace of new drug development and, to a lesser extent, the development of medical devices (Finkelstein, 2004; Acemoglu and Linn, 2004; Budish, Roin, and Williams, 2015).[3] This research has not previously spoken to the question of why new medical technologies have tended to focus on quality rather than cost. We make progress on this question by developing the requisite data and by identifying historical settings in which it can be addressed. Our findings suggest that cost-conscious payment models can steer innovation in a cost-conscious direction.

Second, our findings add to broader lines of research on innovation's determinants. Our analysis complements existing research on demand-induced innovation, within

[3]Additional papers include Acemoglu, Cutler, Finkelstein, and Linn (2006), who find that the introduction of Medicare had no effect on the development of drugs for the elderly, Clemens (2013), who finds that U.S.-based medical equipment and device patenting rose following the introduction of Medicare, Blume-Kohout and Sood (2013), who find that research on drugs with high Medicare market shares rose following the introduction of Medicare Part D, and Dubois, De Mouzon, Scott-Morton, and Seabright (2015), who find that potential profits affect the number of new molecular entities that come to market.

which the environmental literature is extensive.[4] A study of particular relevance to our work comes from Newell, Jaffe, and Stavins (1999), who analyzed the effects of energy prices on the detailed energy efficiency attributes of A/C technology. We also add to a body of research that analyzes innovation in the context of shocks connected to wars.[5]

Finally, we add to an expanding set of papers that use natural language processing, or text analysis, in economics research. Analyses of patent texts have become increasingly common in the innovation literature.[6] Our application shares similarities with recent analyses of "sentiment" and "partisanship," where the objective is to construct new variables describing a text's economic content (Shapiro, Sudhof, and Wilson, 2018; Gentzkow, Shapiro, and Taddy, 2019). We develop several practical insights into best practice methods for this class of machine learning applications.

The paper proceeds as follows. Section 1 develops an analytic framework that connects payment models to the emphasis of inventors on quality and cost. Section 2 provides background on the historical episodes we analyze. Section 3 discusses our novel data set and section 4 our empirical strategy. Section 5 presents our results and section 6 concludes.

---

[4]See Popp (2010) and Popp (2019) for reviews. Acemoglu, Aghion, Bursztyn, and Hemous (2012) present a theoretical framework for analyzing the dynamic effects of environmental policy on innovation, while papers by Aghion, Dechezleprtre, Hemous, Martin, and Van Reenen (2016), Howell (2017), Johnstone, Hascic, and Popp (2008), Ito and Sallee (2018), Knittel (2011) and Newell, Jaffe, and Stavins (1999) provide empirical evidence.

[5]Hanlon (2015), for example, analyzes innovation in the British textile industry as it responded to the supply chain shock connected to declines in access to imported cotton during the U.S. Civil War. Moser and Voena (2012) and Baten, Bianchi, and Moser (2017) use the U.S.'s World War I era "Trading with the Enemy" act to analyze the effects of compulsory licensing. Moser, Voena, and Waldinger (2014) analyze how innovation was shaped by Jewish migration during World War II, while Waldinger (2010) and Waldinger (2011) analyze how innovation was shaped by the Nazi expulsion of professors and scientists. Iaria, Schwarz, and Waldinger (2018a) study the effects of the collapse in scientific communication associated with the onset of World War I. Finally, Khan (2009) and Khan (2015) focus on the Civil War's effects on the trajectories of entrepreneurial inventors.

[6]See, for example, Khoury and Bekkerman (2016); Bergeaud, Potiron, and Raimbault (2017); Iaria, Schwarz, and Waldinger (2018b); Watzinger and Schnitzer (2019); Arts, Cassiman, and Gomez (2018); Cockburn, Henderson, and Stern (2018).

1 Conceptual Framework

Our analysis is concerned with two aspects of the relationship between market forces and medical innovation. The first is the effect of potential profits on the volume of innovation in a particular technological space. The second is whether incentives shape the extent to which an inventor allocates effort to improve production processes or particular dimensions of treatment quality. This second choice has not previously received attention in the literature on medical innovation.

Inventors make meaningful economic choices regarding the time and resources they devote to improving each of a product's attributes. They presumably do this, at least in part, to maximize their effort's impact on the product's value. Cancer treatments provide a natural illustration. Key dimensions of innovation in the cancer treatment setting are life extension, quality of life improvements (e.g., reduction of side effects), and cost reduction. Similarly, innovation in the production of coronary stents could involve streamlining production, the use of lower-cost materials, the use of materials with greater durability, and improvements to the mechanisms through which drug-eluting stents store and release medication. Importantly, advancing a particular dimension of a treatment's frontier can involve solving a distinct scientific problem with unique costs and payoffs. An effort to advance the frontier can thus involve important choices regarding which problems to solve.

In a well-functioning innovation space, returns would connect cleanly to the social value an inventor's efforts are expected to create. There are multiple reasons, however, why markets for new technologies might fail to meet this standard. These include the classic public goods problem and the "business-stealing" effect, which can apply broadly across industries (Romer, 1986; Aghion and Howitt, 1992; Tirole, 1988). Incentives for medical innovation are also shaped by the institutional characteristics of health care markets. These include regulatory approval and reimbursement systems designed

by governments or other third-party payers. Because many medical innovations require regulatory approval to reach the market, attributes that influence approval will weigh heavily in inventors' objective functions. In the U.S., this will tend to reward life extension and safety over cost and dimensions of well-being that are difficult to measure (Budish, Roin, and Williams, 2015).

The following framework connects market size, market structure, and the structure of reimbursements to inventors' decisions to engage in quality-enhancing and cost-reducing innovation. Many of the considerations we highlight are featured in analyses by Rogerson (1989, 1994). A distinction of interest is that Rogerson focuses on procurement from a single source, which is typical in the settings he analyzes. Our setting involves a moderate number of mid-sized manufacturers, which makes competition over market share an ongoing consideration.

Suppose, as in our empirical applications, that a government needs to procure Q prosthetic devices from a market with N potential manufacturers indexed by j. While the government determines how firms are reimbursed, demand for a given firm's product may be driven by either the government or the final consumers. Let firm j's market share, $m_j(v_j, \boldsymbol{v}_{-j}, p_j^c, \boldsymbol{p}_{-j}^c)$, be an increasing function of its own quality ($\frac{dm_j}{dv_j} > 0$) and a decreasing function of other firms' quality ($\frac{dm_j}{d\boldsymbol{v}_{-j}} < 0$). In general, market shares will also be functions of consumer prices (p_j^c and $\boldsymbol{p}_{-j}^c$). In our contexts, however, consumers do not pay directly for a manufacturer's output, such that $p_j^c = 0$ and $\boldsymbol{p}_{-j}^c = \mathbf{0}$. Quality can be increased through innovative effort $e_{j,v}$, while cost can be reduced through innovative effort $e_{j,c}$. Firm j's per-unit production costs, $c_j(e_{j,c}, e_{j,v})$, are increasing in innovation on quality ($\frac{dc_j}{de_{j,v}} > 0$) and decreasing in innovation on cost ($\frac{dc_j}{de_{j,c}} < 0$). The cost of innovative effort itself is $b(e_j) = b(e_{j,c} + e_{j,v})$, with $b(0) = 0$, $b'(0) = 0$, $b' \geq 0$, and $b'' > 0$.

Both overall profitability and the relationship between profit and innovation depend on the government's reimbursement schedule. Our description of the reimbursement

schedule nests classic "cost-plus" and "fixed-price" reimbursement schemes. That is, reimbursements can contain both a fixed component and a cost-based component: $r_j = \underline{r} + \beta c_j(e_{j,c}, e_{j,v})$. Cost-plus reimbursement, for example, ensures that firms make a profit regardless of c_j by either setting $\beta > 1$ and $\underline{r} = 0$ or $\beta = 1$ and $\underline{r} > 0$.[7] A fixed-price reimbursement, by contrast, sets $\beta = 0$ and pays exclusively through $\underline{r}$.

We make several simplifying assumptions about the environment that are worth stating explicitly. First, we treat the problem as static rather than separating the periods during which innovation choices are made from the periods during which sales occur.[8] Second, we abstract from the possibility that the government may separate the innovation and manufacturing functions by directly financing, or even producing, innovation itself.[9] Third, we characterize how innovative effort affects a firm's profitability while holding other firms' effort levels fixed. Fourth, our characterization of cost-based reimbursements abstracts from the fact that the procurer's estimates of cost might be averaged across firms and might be updated with lags.[10] These assumptions do not affect the qualitative insights we emphasize but allow for simplified exposition.

This set-up yields three expressions of interest. First, profit for firm j, π_j, is

$$\pi_j(e_{j,v}, e_{j,c}) = Qm_j(v_j, \boldsymbol{v}_{-\boldsymbol{j}}, p_j^c, \boldsymbol{p}_{-\boldsymbol{j}}^{\boldsymbol{c}})[r_j - c_j(e_{j,c}, e_{j,v})] - b(e_j). \tag{1}$$

Second, the effect of an increase in quality-oriented innovation on profit is

[7]Rogerson (1994) points out that setting $\beta > 1$ can be attractive when production-phase profits are needed to encourage innovation on quality and when the procurer desires for the magnitude of that incentive to rise with overall project costs.

[8]Canonical models have effectively captured key features of the problem of contracting to induce effort to reduce production costs in one-period frameworks (Shleifer, 1985; Laffont and Tirole, 1986; Rogerson, 2003). While incentives for innovation on product quality are best captured by models with distinct "product development" and "production" phases, our framework nonetheless captures the forces emphasized by Rogerson (1989, 1994) that are most relevant to our setting.

[9]While direct public financing for research and development was absent in the context of our Civil War application, it was a factor in the context of our World War I application.

[10]Rogerson (1994) points out that lags can be used purposefully to make cost-conscious innovation profitable within an ostensibly cost-based reimbursement structure.

$$\frac{d\pi_j}{de_{j,v}} = Q\frac{dm_j}{de_{j,v}}[r_j - c_j(e_{j,c}, e_{j,v})] + Qm_j(\cdot)[\frac{dr_j}{de_{j,v}} - \frac{dc_j}{de_{j,v}}] - \frac{db}{de_{j,v}}. \quad (2)$$

Third, the effect of an increase in cost-oriented innovation on profit is

$$\frac{d\pi_j}{de_{j,c}} = Qm_j(\cdot)[\frac{dr_j}{de_{j,c}} - \frac{dc_j}{de_{j,c}}] - \frac{db}{de_{j,c}}. \quad (3)$$

These expressions have implications for the effects of the size of the market, market structure, and the structure of reimbursements on innovation. First, so long as profit is increasing in either quality-enhancing or cost-reducing innovation, the magnitude of the incentive to innovate is strictly increasing in the size of the market, Q. Second, the relative returns to quality-enhancing and cost-reducing innovation depend, among other things, on market structure and the structure of reimbursements.

Under the most basic form of cost-plus reimbursement, the government sets $\underline{r} > 0$ and $\beta = 1$, which implies that $\frac{dr_j}{dc_j} = 1$. This has three direct implications: 1) the first term in equation (2), which describes increases in profit from increases in units sold, will be positive, 2) the second term in equation (2), which captures changes in profit per unit, is equal to 0, and 3) the first term in equation (3), which again captures changes in profit per unit, is 0. Together, these implications push innovation towards quality enhancement and away from cost reduction under a cost-plus reimbursement regime. A positive return is initially (i.e., starting from $e_j = 0$) guaranteed for quality-enhancing innovation, while there is no benefit to innovating to reduce cost.

Under a fixed-price regime, we have $\beta = 0$ and $r_j = \underline{r}$. Under this regime, note that the initial return to cost-saving innovation is guaranteed to be positive, since $\frac{dr_j}{de_{j,c}} = 0$, $\frac{dc_j}{de_{j,c}} < 0$, and $\frac{db(0)}{de_{j,c}} = 0$. Note that the initial return to quality enhancing innovation depends crucially on the level at which the payment is set. Since $\frac{dr_j}{de_{j,v}} = 0$, the change in profit per unit sold (the second term in equation (2)) is negative. A positive return thus requires the first term, which describes profit linked to increases in the number

of units sold, to be positive. Now note that if $\underline{r} < c_j(0,0)$, the first term will also be negative. A fixed payment regime with a reimbursement rate set below baseline cost thus guarantees that cost-reducing innovation will occur before quality enhancing innovation. Note, however, that if the payment is set too low firms will neither innovate nor be willing to make sales, since all sales would generate losses.

Fixed-price regimes will tend to generate innovation on both cost and quality. When $\underline{r} < c_j(0,0)$, participating firms will initially focus on cost-conscious innovation, but may ultimately choose positive levels of both cost-conscious and quality-enhancing innovation.[11] Interestingly, it is possible for a fixed-price regime to generate predominantly quality enhancing innovation. This outcome will be relatively likely when the fixed reimbursement is very high ($r_j >> c_j(e_{j,c}, e_{j,v})$) and when market share is highly sensitive to quality ($\frac{dm_j}{de_{j,v}} >> 0$). These conditions can lead the first term of equation (2) to exceed the sum of the second term of equation (2) and the first term of equation (3). Market structure and the level of reimbursement within fixed payment regimes can thus determine the focus of innovative efforts.

Market structure influences several aspects of the returns to innovating on both cost and quality. First, as noted above, the returns to innovation on quality are increasing with the effect of quality on market-share ($\frac{dm_j}{de_{j,v}}$). Markets in which consumers are highly sensitive to variations in quality will thus tend to generate intensive effort to innovate on quality relative to cost. A related point is that contracts over fixed quantities reduce firms' incentives to innovate on quality by shutting down (or at least blunting) the market share channel. Second, market structure may, for practical reasons, either facilitate or inhibit the administration of cost-based reimbursements. The cost structure for a monopolist, for example, describes the cost structure for an entire market. The procurer

[11] A firm makes sales and innovates if its profit-maximizing innovation choices result in per-unit production costs that are below the reimbursement rate, or $\underline{r} > c_j(e^*_{j,c}, e^*_{j,v})$ (which can occur even when $\underline{r} < c_j(0,0)$).

may thus adjust reimbursement rates quickly in response to a monopolist's innovation on cost, which blunts the incentive for cost-conscious innovation. In a market with many small players, by contrast, reimbursements may be set to align with the procurer's estimates of cost, perhaps as averaged across participating firms. A single firm's innovation on cost might then have very little effect on payments, making cost-conscious innovation profitable.

Finally, the procurer may, in some cases, contract directly on dimensions of innovative effort. During World War I, for example, the British government played a direct role in identifying the "kinds of devices" manufacturers should produce (Guyatt, 2001, p. 312). Further, both the U.S. and British governments of World War I took the step of directly employing researchers for the production of new materials and prosthetic device designs (Guyatt, 2001; Linker, 2011).

Our basic framework is useful for analyzing the incentives generated by a rich set of reimbursement systems. Systems of potential interest include traditional Medicare's fee-for-service model and the widely used Prospective Payment System for hospital reimbursement. In what follows, we analyze the innovation that occurred under the Civil War and World War I era systems for procuring and reimbursing prosthetic devices.

2 Background on Wartime Prosthetic Device Procurement

Both the U.S. Civil War and World War I were associated with dramatic increases in demand for prosthetic devices. In this section, we describe the size of these demand shocks, then provide background on U.S. and foreign systems for rehabilitating amputee veterans and procuring their artificial limbs. Because the histories connected to each conflict are dense, our brief discussion will inevitably miss many nuances.

2.1 Background on Wartime Demand Shocks

The U.S. Civil War was contested between the armies of the Union and the Confederacy from April 1861 to May 1865. An estimated 35,000 amputees survived the war (Linker, 2011, p. 98). Because the government had not formed a permanent bureaucracy for addressing veteran health care needs prior to the war, both the Union and Confederacy implemented ad hoc artificial limb procurement systems as the scope of need became clear. The Union army's program for procuring artificial limbs, which was administered within its broader pension system, was initiated through a Congressional appropriation dating to July 21st, 1862 (Hasegawa, 2012, p. 21). In a communication to Congress, Barnes and Stanton (1866) report that as of May 1866 the Union program had delivered 6,075 artificial limbs, including 3,798 legs and 2,204 arms, at a cost of just under $360,000 (roughly $6 million in 2018). Hasegawa (2012) documents the delivery of just under 750 prosthetic devices by the Confederacy over a similar period.

World War I was contested from July 1914 to November 1918, with U.S. involvement commencing April 6th, 1917. The war produced an estimated 300,000 amputee survivors worldwide, of whom roughly 67,000 were German and 41,000 British (Guyatt, 2001, p. 98). Relative to the Civil War, demand associated with 4,000 amputee U.S. veterans was relatively modest. Because production capacity was low among the European powers and high in the U.S., however, the U.S.-based artificial limb industry played a major role in satisfying global demand. Linker (2011) writes, for example, that "While serving in France, American orthopedist Robert Osgood estimated that during the year 1915, French manufacturers were able to produce only 700 limbs for its 7,000 amputees" (Linker, 2011, p. 98).[12] The European powers thus utilized U.S. manufacturing capabil-

[12] A historical question of interest is why the prosthetic device industry in Europe had not developed in the wake of the Crimean War which, like the U.S. Civil War was fought using "Minie Ball" bullets, which dramatically increased the prevalence of wounds necessitating amputation (Freemon, 1993). The answer likely lies in sheer numbers. Estimates of wounded war survivors were roughly three times larger during the U.S. Civil War than during the Crimean War (Garrison, 1917). Additionally, a larger fraction

ities. Great Britain, for example, invited the largest American prosthetic companies "to set up workshops at the main amputee center" (Linker, 2011, p. 99).

2.2 Background on Civil War-Era Procurement

Although the Union's Civil War era artificial limb program was ad hoc, its design was quite sensible. As Hasegawa (2012) documents, a modest initial appropriation by Congress led General William Hammond to convene a panel of physicians to, in Hammond's words, "determine what kind of Artificial Limbs should be adopted for the use of mutilated soldiers." Hasegawa (2012) describes a series of subsequent meetings during which the panel assessed inventors' prototypes for artificial arms and legs. If satisfactory, the panel deemed an artificial limb "serviceable," allowing its subsequent purchase through the program.

Reimbursement occurred on a fixed-price basis. Artificial arm provision was temporarily delayed because no prototypes were initially deemed to be of sufficiently high quality to merit approval (Hasegawa, 2012, p. 34).[13] Artificial arms were subsequently approved at a price of $50, while the price for artificial legs was set at $75 (roughly $1,500 in 2018 dollars) for the bulk of the war (Hasegawa, 2012, p. 37-38). Importantly, with

---

of the surviving wounded was likely to be amputees during the Civil War than during the Crimean War due to improvements in surgical survival rates. Estimates suggest amputation survival rates of roughly 75 percent during the U.S. Civil War (Figg and Farrell-Beck, 1993). During the years surrounding the Crimean War, by contrast, amputation survival rates among civilians treated in the relatively favorable conditions of the London Hospital were nearly 50 percent. (Macleod, 1858, p. 168) enumerates a total of 521 amputee British survivors during the last year of the two and a half year conflict, a period extending from April 1, 1855, to March 30, 1856. He notes that the 73 percent survival rate for this latter period of the war was surely far higher than the rate under the far less favorable conditions of the war's first two years. A final point of interest is that roughly 60 percent of the Crimean War's surviving war wounded were from the Russia Empire (Garrison, 1917). While details on Russian procurement of prosthetic devices have proven difficult to come by, we speculate that its arrangements were likely less generous than those of either the Union or, for that matter, the Confederacy.

[13]This indicates that the panel took its job rather seriously, as a number of low-quality offerings were denied approval. Further, this highlights that the artificial arm patents from this period were associated with appreciable improvements in product quality, at least as assessed by the review panel.

reference to section 1's conceptual framework, these prices were quite low relative to manufacturers' stated costs.[14] Balance billing was not permitted. After the war's conclusion, the panel of physicians rated the qualities of alternative limbs and allowed soldiers to select higher-quality limbs at higher prices, with the soldier paying or taking home the difference from the allowance (Hasegawa, 2012, p. 40).

Over the decades immediately following the war, the U.S. government provided regular artificial limb replacements for veterans. Like the initial appropriation for artificial limbs, these reforms took place within the context of the Union Army's pension system. Importantly, veterans were allowed to choose between a replacement limb and cash, which was referred to as a commutation payment (Hasegawa, 2012, p. 76). Statistics from annual reports of the army's Surgeon General reveal that veterans overwhelmingly preferred cash; from 1870 to 1891, "arm amputees chose a new device over commutation only 1.4 percent of the time, and leg amputees selected a new leg 21.9 percent of the time" (Hasegawa, 2012, p. 76). As detailed below, this program's budgetary costs, coupled with societal perceptions of limbless veterans "pocketing" their allowances, greatly impacted World War I era views regarding care and rehabilitation for veteran amputees.[15]

2.3 Shifts in Approaches to Treatment, Rehabilitation, and Innovation

By World War I, the U.S. had substantively formalized the treatment of amputee veterans. This occurred within a broader effort to formalize veterans' health care, which was motivated in part by the cost overruns and seemingly endemic politicization of benefits administered through the Union Army's pension system (Cogan, 2017). In addition

[14]Hasegawa (2012) documents that a leading manufacturer told the government his costs were $150 per artificial leg. Findings from Chan and Dickstein (2019) caution, however, that providers will tend to inflate cost-assessments when their reimbursements depend on it.

[15]Linker (2011) argues that Civil War amputees who opted for cash rather than a replacement limb, or who otherwise purchased cheap peg legs, had effects on views of treatment and rehabilitation that prevailed during World War I.

to being formalized, care for amputees had also been largely centralized at large facilities including the recently built Walter Reed Hospital.[16] Progressive Era policymakers worried that amputee veterans would, like many of their Civil War predecessors, fail to return to gainful employment. As Linker (2011, p. 13) writes, "The veterans of America's First World War were expected to become citizen-workers once their military service was over; they were to make useful lives, not to languish at the expense of the US Treasury." Further, "The Limb Lab's goal was to give every man, whether legless or armless, a 'modern limb'—a limb that would make it possible for amputee soldiers to pass as normal, able-bodied citizens in the workplace and on the streets" (Linker, 2011, p. 101). The British and German governments had similar views on the importance of rehabilitation and reemployment.[17]

Between the Civil War and World War I, discretion in the choice of artificial limb shifted from soldier to government. During World War I, amputee veterans underwent extensive rehabilitation prior to their return to civilian life, including obligatory use of standard-issue prosthetic limbs. Linker (2011, p. 101) writes that "the OSG [Office of the Surgeon General] forcefully mandated artificial limb wear, creating legislation that made it virtually impossible for US amputee soldiers to be discharged from military service without months of rehabilitation and daily routine artificial limb wear." Physicians now mediated between veterans and artificial limb manufacturers.

Medical professionals of the World War I era de-emphasized the amputee's comfort

[16]Treatment of amputee veterans also took place at Letterman hospital in San Francisco. As Linker (2011, p. 80) writes, "Surgeon General Gorgas designated two general hospitals to become permanent installations for rehabilitative care: Letterman General Hospital in San Francisco and Walter Reed General Hospital in Washington. Later in the war, the list of military rehabilitation hospitals would grow to 14, but Letterman and Walter Reed remained the flagship facilities during and after the war."

[17]See, for example, Guyatt (2001, p. 311-312) regarding the British government's objectives with regards to artificial limbs. In a description of German austerity towards amputee veterans, Perry (2014, p. 124) describes the prevailing view as being that "the greatest obstacle to war-time physical rehabilitation was not the injury itself, but rather the soldier's own lack of 'will to work"' and that they were "encouraged by others to become dependent on welfare and charity." Perspectives on amputee veterans and the importance of self-sufficiency thus differed starkly from what one might expect in more recent times.

in favor of this strict rehabilitation program. In a description of prevailing views and approaches to rehabilitation, Linker (2011, p. 109-114) writes:

> Once surgical healing had been attained... the 'toughening' of the stump by 'pounding it on a firm surface' should be 'vigorously pursued'... Following stump pounding exercises, 'patients usually complained of discomfort'... Another report stated that when amputees were forced to wear artificial limbs soon after surgery, they often 'expressed gratitude when the artificial limb [was] removed.'

In addition to driving a relatively severe program of physical rehabilitation, the desire for economic and social reintegration spurred an emphasis on "disguising" veterans' disabilities. A chief of the War Risk Insurance Bureau, for example, wrote that "One of the most useful and necessary duties of this department will be to prescribe and furnish medical and surgical treatment in order that disabilities may be reduced or caused to disappear entirely" (Linker, 2011, p. 100).

A final key point differentiating Civil War and World War I prosthetic device innovation is the direct role of governments in these efforts. During the Civil War, innovation came entirely from private industry. During World War I, innovation was, in part, conducted by governments and contracted directly by governments. The U.S. "Limb Lab" is an instance of innovation conducted by the government itself. In the United Kingdom, Guyatt (2001, p. 312) notes that "the government's Ministry of Pensions decided which limb-makers were contracted and, increasingly, what kinds of devices they should make." This work was complemented by a government-run laboratory focused on "developing new materials for use in the industry."

3 Data and Text Analysis

We begin this section with a discussion of the historical patent data we use to estimate the effects of wartime demand shocks on overall patent flows. We also discuss the typical caveats for using patents as a measure of innovation and provide evidence on why the patents we analyze have relatively strong links to true technological advances. We then discuss the new data we generated through text analysis (or natural language processing) using a combination of close readings and machine learning techniques.

3.1 Historical Patent Data

The first question we attempt to answer is if wartime increases in demand for prosthetic devices increased the rate of prosthetic device patenting. This analysis requires information on 19th and early 20th century patents by technology class. Until relatively recently, the patent data sets analyzed by economists did not facilitate this type of historical analysis. The groundbreaking NBER patent database (Hall, Jaffe, and Trajtenberg, 2001), for example, begins with patents granted in 1963. Economists have recently developed databases extending to the earliest surviving records of the U.S. Patent and Trademark Office (USPTO). To identify historical patents based on their technology classes, we use the database assembled by Berkes (2018).[18]

Figure 1 provides an initial look at time series on prosthetic device patents and

[18]In a comparison of several recent efforts to compile data sets on the universe of U.S. patents, Andrews (2019) concludes that the database laid out in Berkes (2018) is "currently the gold standard." Additional analyses of 19th and early 20th century patents have been made possible by these data. Berkes and Nencka (2019), for example, analyze the effects of the original Carnegie Library donations on innovative activity, finding that the establishment of Carnegie Libraries had substantial effects on patenting rates. Berkes, Gaetani, and Mestieri (2019) use the historical patent data to analyze the rise and fall of cities. They find that diverse innovation portfolios are associated with a city's resilience to the rise and fall of particular industries, while cities with innovation in the most central fields exhibit the strongest growth over subsequent decades. A similarly historic patent data set is under analysis by Akcigit, Grigsby, and Nicholas (2017). The PATSTAT database maintained by the European Patent Office, as analyzed for example by Doran and Yoon (2018), enables patents granted by the U.S. Patent Office to be tracked as far back as 1899.

other broad categories of patents during the historical episodes we analyze. The dashed vertical lines in each panel encompass the years we subsequently associate with "war-induced" booms in prosthetic device patenting. It is quite clear from the panels of figure 1 that both the Civil War and World War I were associated with dramatic increases in the rate of prosthetic device patenting. However, quantifying the causal effect of wartime demand shocks faces the difficulty of constructing appropriate counterfactuals, which we discuss in section 4.

3.2 Patents As a Measure of Innovation

Our use of patents as a measure of innovation faces standard caveats. Not all innovations are patented (Moser, 2005, 2012), for example, and not all patents are indicative of meaningful innovation. Further, in our historical context we can use neither patent citations nor market valuations as proxies for value or scientific impact, as has been done in analyses of patents from more recent periods (Trajtenberg, 1990, 1989; Hall, Jaffe, and Trajtenberg, 2005; Kline, Petkova, Williams, and Zidar, 2019).[19] Nonetheless, there is substantial evidence to support a strong link between this period's prosthetic device patents and meaningful technological advances.

In our Civil War and World War I era contexts, two factors ease standard concerns regarding the link between patents and the underlying flows of innovation. First, the periods we analyze pre-date more recent concerns regarding "patent trolls" (Cohen, Gurun, and Kominers, 2014). Second, Khan (2015) observes that "there is ample evidence that inventors during the 19th century were especially anxious to secure their rights through patenting." Together, these factors suggest that useless patents and unpatented innovations are less common in our context than in recent settings.

[19]Standard reference sections were not included in patents until the mid-20th century. Similarly, market valuations are not available for the mid-19th and early-20th century firms we study.

Two additional points connect this period's prosthetic device patents to meaningful technological advances. First, medical histories document that these episodes were, in fact, episodes of substantial advance in artificial limb technologies. Finally and more directly, available data establish links from patents to manufacturers, from manufacturers to sales, and from both sales and manufacturers to expert assessments of quality. We present the data underlying these connections in table 1. Twelve out of the thirteen most notable manufacturers of artificial legs and eight out of the nine most notable manufacturers of artificial arms from the Civil War period can be linked to at least one patent. Through May 1866, these patent-holding manufacturers accounted for nearly all of the artificial legs and nearly 90 percent of the artificial arms furnished to Union Army veterans.

Post- and late-war rankings of artificial limbs by quality further support a link between quality and market share (Barnes, 1865; Houston and Joynes, 1866). The top three rated artificial legs accounted for just under 60 percent of sales through 1866, while the top four rated artificial arms accounted for just over 60 percent of sales through 1866. The highly-rated limbs with low market shares were those developed relatively late during the war, namely the artificial arms of John Condell and the National Arm and Leg Company. The low market shares of these limbs are thus largely mechanical, as they were not on the market when most of the limb purchases for which we have documentation occurred. Low-rated limbs with non-trivial market share tended to be either unpatented or to involve pre-war patents, suggesting an incumbency advantage.

3.3 Coding Patent Attributes

Beyond measuring patent flows, our analysis aims to understand the economic attributes that are emphasized in each patent. We pursue this to understand how inventors distributed their efforts across improving aspects of production processes and/or par-

ticular dimensions of each product's quality. Because the data required for this analysis did not previously exist, we developed a novel data set.

Our data set contains information from historical patent documents that quantifies the economic attributes that each patent emphasize. First, we created a program to scrape the historical patent documents from Google Patents. Within the text of these patent documents, we then analyzed six innovative attributes.[20] Four of these, namely cost, simplicity, adjustability, and materials, are cost-oriented production process attributes. That is, these traits emphasize dimensions of an innovation's construction. We use the term "adjustability," for example, to describe patents that emphasize uniform production of outputs that can subsequently be fitted (or "adjusted") to the needs of a specific consumer. Our other two traits, namely comfort and appearance, are quality-oriented attributes. Table 2 presents a concise verbal definition of each economic attribute.

To develop this data set, we first manually classify our six attributes for two sets of patents surrounding the war eras of interest. The first set contains patents related to prosthetic devices, or class 623, as defined using the UPSTO's patent classification system. The NBER patent database categorizes class 623 as a subset of technology subcategory 39, "Misc. Drugs and Med," which in turn is a subset of technology category 3, "Drugs and Medical." The second set contains patents from all other classes within the "Drugs and Medical" and "Mechanical" categories, which form our control groups. The patents of interest come 1840 to 1890 and 1900 to 1940. Our sample of closely-read patents is then selected using stratified random sampling. We stratified across patent classes and war episodes to ensure coverage across our treatment and control groups during both time periods of interest.

[20]Our focus on these attributes was motivated by initial close readings of patent documents from both prosthetic devices and our control groups. Useable attributes needed to be of economic interest, as well as coherently and similarly defined in both our treatment and control groups.

As summarized in table A.1, the manually coded data set contains 195 prosthetic device patents and 399 other medical or mechanical patents from the Civil War period, as well as 302 prosthetic device patents and 305 other medical or mechanical patents from the World War I period. We use these manually classified data to train a machine learning model to code the same variables for additional patents.

3.4 Text Analysis

This section provides an overview of the text analysis tools we developed and implemented. Classifying text according to its semantic content requires overcoming several important difficulties. First, synonyms and variation in word meaning across contexts can induce errors in text classification tasks. Researchers can select from a variety of available algorithms, each having strengths and weaknesses, to address these issues. Second, even a well-chosen algorithm can perform poorly if provided too little data from which to learn. A model may also perform poorly if trained on data from contexts that differ from those to which it is applied. Neglecting these issues can lead an algorithm to generate inaccurate data, which can result in biased or uninformative estimates.

Our experimentation with a set of machine learning models showed that models trained on the entirety of each patent's text generated inaccurate results and were computationally slow. We thus modified existing algorithms by constraining the text inputs they consider to a combination of keywords and the immediate textual context surrounding the keywords. We adopted this approach for two reasons. First, restricting the input data in this way improved model accuracy. Second, our approach led to efficiency gains with respect to both processing times and small sample performance.

3.4.1 Overview of Key Issues

Machine learning algorithms generally perform best when provided many observations from which to learn. The number of observations required for a particular classification task depends on the complexity of the data used. When using text as data, each observation can contain hundreds of unique words. If provided too few observations, an algorithm can struggle to ascertain each word's relative importance. Directing an algorithm to focus on the most relevant words can aid in achieving accurate results from a relatively small training set.

The complexity of language presents additional challenges for text classification. Two such challenges are "polysemy," which applies when a word has multiple meanings, and "synonymy," which applies when multiple words have the same meaning. Variation in word meaning and usage can occur within and between text documents, across domains (e.g., prosthetic device patents vs. mechanical device patents), and across time periods. An algorithm's performance depends on how effectively these issues are addressed.

Classifying text by searching for keywords, for example, will tend to yield poor results if the keywords suffer from a high degree of polysemy. Keyword searches fail to account for nuanced variations in contextual meaning, which can lead to large inclusion errors when polysemy is severe. Machine learning models, by contrast, attempt to use context to ascertain meaning. Such models can fail to detect contextual meaning, however, when the algorithm is provided a data set containing too few documents (observations) from which to learn.

3.4.2 Our Approach: Feature Selection for Machine Learning

A machine learning algorithm's performance can often be improved by limiting its attention to the most relevant words, or "features," in a document's text. This process is called "feature selection." The familiar Lasso procedure, for example, limits the number

of features in the model by applying a penalty factor within its objective function. Guyon and Elisseeff (2003) note that feature selection has been shown to help at "improving the prediction performance of the predictors, providing faster and more cost-effective predictors, and providing a better understanding of the underlying process that generated the data." We thus develop and validate an approach that selects a set of features comprised of keywords, their synonyms, and a flexible neighborhood of textual context surrounding the keywords and synonyms.

We begin by connecting each of our attributes to lists of keywords. To construct these lists, we first develop domain knowledge by closely reading 1,200 patent documents. We then supplement our initial lists of keywords using a data-driven process. Specifically, after preprocessing the patent texts we use the algorithm "Word2Vec" (Mikolov, Sutskever, Chen, Corrado, and Dean, 2013) to identify additional synonyms.

With our comprehensive list of keywords, we then implement our method for selecting the constrained set of features that are inputs for a machine learning algorithm. The steps in our method are displayed diagrammatically in figure B.1. After preprocessing, we select each occurrence of a keyword within a given document. We then pair each keyword with a neighborhood of its surrounding context. We term this neighborhood the "spread." If the spread parameter takes a value of j, for example, we keep the keyword and all words up to $j-1$ spots to the left or right of the keyword.[21]

The next step in implementing our algorithm is to arrange the words selected through

[21]Figure B.2 displays an illustrative example of a patent document to highlight how we instruct a machine learning algorithm to see only important words in an otherwise cumbersome text document. Consider the following sentence fragment from a patent in our sample: "[joints] may be moved in adjusting the fingers or thumb to any given article." Our preprocessing procedure converts this sentence fragment to "joints moved adjusting fingers thumb given article." The word "adjusting" is, unsurprisingly, one of our keywords indicating a potential case of the trait "adjustability." If our spread parameter were $j = 3$, then our feature space would include the words: joints, moved, adjusting, fingers, thumb. This technique provides contextual cues, namely "fingers" and "thumbs," that signify that the keyword "adjusting" does not denote mass-producibility. Instead, here the word "adjusting" indicates an improvement in the functionality of the prosthetic hand. See appendix B for additional discussion of the selection of the spread parameter.

the above procedure into a document-term, or "doc-term," matrix. Each unique word in the set of selected words is assigned to a column within the doc-term matrix. Each document is assigned to a row. Entry (i, j) is thus assigned a value corresponding to the relative frequency with which word j was used in document i, called the "tf-idf" value.[22]

Note that the poor performance of algorithms that do not limit the feature space is linked to the dimensionality of the doc-term matrix. When the feature space is not constrained, the number of columns in the doc-term matrix can far exceed its number of rows, with many of these columns representing largely extraneous words. This can inhibit the model from isolating the words on which it should focus. When this problem arises, restricting the feature space can enable machine learning models to obtain accurate results even at smaller sample sizes.

3.4.3 Evaluating the Performance of Alternative Models

Our evaluation of alternative machine learning approaches generated several insights. First, constraining the feature space led to significant gains in the accuracy obtained by each of the machine learning algorithms with which we experimented.[23] Our preferred algorithm, from which we obtained the most accurate models, was Gradient Boosted Machines (GBM) (Friedman, 2001).

Second, reasonable text analysis algorithms can perform poorly if not provided a sufficient number of observations from which to learn. Rule-of-thumb statements regarding the sample size needed to train an accurate machine learning model are difficult to

[22]Tf-idf is defined as $\text{tf\_idf}_{i,j} = \frac{t_{i,j}}{T_j} \log\left(\frac{D}{D_i}\right)$ where $t_{i,j}$ is the total number of times a term j appears in the document i, T_i is the total number of terms in document i, D is the total number of documents, and D_j is the total number of documents containing term j.

[23]For more details on the accuracy metric we adopt and the cross-validation procedure we employ, see appendix B. The algorithms we considered were Support Vector Machines (Cortes and Vapnik, 1995), Gradient Boosted Machines (Friedman, 2001), Naive Bayes, Random Forests (Breiman, 2001), and Lasso (Tibshirani, 1996).

make. However, researchers can take reliable steps to ascertain their particular needs. A reliable way to establish an endpoint for sample collection is to iteratively measure a model's out-of-sample accuracy as additional observations are added to the training set. When the accuracy score asymptotes, the gains from further expanding the training set become small (see appendix section B.6.1 for a more thorough discussion of the exercise we conducted). We find that the gains from expanding our training set vary considerably across the variables we constructed, as the economic traits differ in their degree of semantic complexity.[24]

Third, the training set must cover the full set of contexts to which the trained model will be applied. This is because word meaning and usage can vary across contexts. To illustrate this point, we partition our data into distinct combinations of time periods and/or patent classes. We then investigate how models perform when they are trained in one context and used for prediction in others. For example, we use World War I-era patents from our control classes to train a model we then use for prediction of traits in Civil War-era prosthetic device patents (see appendix section B.6.3 for a more thorough discussion of the exercise we conducted). For quality-oriented traits, we find that different train-test contexts result in substantial losses in the model's accuracy. Cost-oriented production process traits suffer from less loss of accuracy. This suggests that the semantic indicators of production-processes transfer more readily across contexts.

Finally, we illustrate why it can be important for researchers to concede when a text analysis problem has proven too difficult to place strong weight on the ensuing analysis. As shown in section B.6.2, moderate declines in accuracy can, in our setting, result in substantially attenuated estimates. As shown in table B.2, we obtain moderately lower

[24]For the most straightforward economic traits we analyze (e.g., "simplicity," for which the problems of polysemy and synonymy are modest), we obtain high accuracy scores with training data sets containing as few as 100 observations. For more semantically complicated traits (e.g., "comfort,") accuracy continued to improve as the training set traits expanded to roughly 1,000 observations.

accuracy scores for the trait we term "materials" than for other traits. Consequently, our analysis of materials may be less reliable than our analysis of other traits.

3.5 Novel Dataset on Patent Attributes

Our final data set, produced by our machine learning approach, describes the economic attributes of 745,558 patents, with the earliest coming from 1840 and the latest from 1940. There are 814 prosthetic device patents, 19,666 other medical patents, and 725,078 mechanical patents. Our regression analyses focus on samples of our 745,558 patents for which the patent year is in relatively close proximity to each conflict. These samples extend from 1855 to 1867 and from 1910 to 1922.

Across this large set of patents, table A.3 shows that the economic traits we coded are only modestly correlated with one another. The primary exceptions are cost and simplicity. Among prosthetic device patents, these traits share a correlation of 0.378 with an associated r-squared of 0.142. Similarly, across all patents in our data set these traits share a correlation of .303 with an associated r-squared of 0.092. Correlations across all other trait pairs are between -0.1 and 0.1, highlighting that each trait captures an independent dimension of innovation.

4 Empirical Strategy

Here, we present our specifications for analyzing both changes in patenting rates and changes in the economic characteristics emphasized by inventors in their patent applications. After presenting each estimation framework, we highlight the key challenges we face when attempting to generate causal estimates of the effects of wartime procurement.

We begin by estimating the following regression equations. The first is specified as an Ordinary Least Squares model for predicting the log of patents per year:

$$ln(N_{t,c}) = \alpha_{c,w(t)} + \alpha_t + \beta_1 1\{\text{War}\}_t \times 1\{\text{Prosthetic}\}_c + \epsilon_{c,t}. \quad (4)$$

The second is specified as a Poisson model of patent counts:

$$E[N_{t,c}|X_t] = exp(\gamma_{c,w(t)} + \gamma_t + \beta_1 1\{\text{War}\}_t \times 1\{\text{Prosthetic}\}_c + \varepsilon_{c,t}). \quad (5)$$

In both equation (4) and equation (5), c denotes patent classes, t denotes time (multi-year time periods for these specifications), and $w(t)$ denotes war episodes (Civil War and World War I). $N_{t,c}$ denotes the number of patents in class c at time t. The specifications include time fixed effects (α_t or γ_t) and episode-by-patent class fixed effects ($\alpha_{c,w(t)}$ or $\gamma_{c,w(t)}$). The coefficient of interest is β_1, which is an estimate of the differential change in the patenting rate for prosthetic devices relative to the control classes during war episodes relative to pre-war periods.

The key challenge in developing causal estimates is to construct control groups that approximate the counterfactual development of patenting rates for prosthetic devices. Technology classes might generate inappropriate counterfactuals for a variety of reasons. They might, for example, be affected by very different scientific developments (e.g., nuclear technology). Alternatively, a plausibly comparable technology class will be a poor control class if it is directly affected by wars (e.g., firearms) or if it is shaped by spillovers from prosthetic device innovation.

Our selection of a complementary set of control groups follows the logic of Finkelstein (2004), whose analysis of vaccine clinical trials is analogous to our setting in key respects. The patents we use to construct control groups come from broad categories of medical and mechanical innovations. Our largest control group incorporates all such patents. We also consider sub-groups that are chosen to either increase comparability or reduce the likelihood that the control group contains patent classes that could be di-

rectly affected by the wars. Like Finkelstein (2004), we also consider data-driven control groups. For our analysis of patent flows, the data-driven approach selects the control group to match baseline flows of prosthetic device patents in levels.

For estimating the effects of wartime procurement on changes in patent characteristics, we use a simple difference-in-differences model. The variable of interest in this analysis describes the share of patents, within a given technology class, that emphasize the characteristic of interest. This removes the underlying trend of patenting rates and creates a measure of the relative intensity with which inventors focus on particular traits. We can write the estimator as follows:

$$\begin{aligned}\beta = &[\text{Prosth. Trait Share}_{wartime} - \text{Prosth. Trait Share}_{prewar}] \\ &- [\text{Other Trait Share}_{wartime} - \text{Other Trait Share}_{prewar}],\end{aligned} \tag{6}$$

where

$$\text{Category Trait Share}_{period} = \frac{\text{\# Category Patents with a Trait}_{period}}{\text{\# Category Patents}_{period}}.$$

Identifying suitable control groups for estimating β in equation (6) requires overcoming additional difficulties beyond those associated with estimating β_1 in equations (4) and (5). These additional issues stem from the fact that some traits of interest are only relevant to a small set of the technology classes within our broadest control group. The statistics in table A.2 reveal, for example, that our quality-oriented traits "appearance" and "comfort" are much more prevalent in prosthetic device patents than in other medical or mechanical patents. In contrast, the prevalence of cost-oriented production-process attributes is similar when comparing prosthetic devices to our broadest control group.

This key difference between our quality-oriented and cost-oriented traits may apply somewhat broadly across technology classes. Quality-oriented traits capture the functional and aesthetic details that, within a given technological class, create value for consumers. Cost-oriented production process traits like "simplicity" and "cost," by contrast, are abstractions that may effectively apply to production processes in many technology classes.

This issue requires that control groups for analyses of patents' attributes be selected using an approach that weeds out technology classes for which an attribute is largely irrelevant. Our primary method for constructing control groups uses the synthetic control approach to matching the levels and trajectories of a patent category's emphasis on a trait over the baseline period. We obtain similar results using a simpler "caliper method" approach. The caliper method selects control groups based on the average prevalence of a trait across the entirety of the baseline period.[25] When implementing the synthetic control approach for our Civil War sample, patent flows for many technology classes were limited, including prosthetic devices. In each of 1858 and 1861, for example, there was a single prosthetic device patent. The maximum across the pre-Civil War years was seven, which occurred in 1859. The share of patents emphasizing a given trait is thus highly volatile across the Civil War baseline when expressed at an annual frequency. Matching year-to-year trends would amount to matching noise. For our baseline method, we thus match levels and trends in four-year moving averages.[26]

Table 3 presents data on the baseline means for our patent trait variables for pros-

[25]Notably, the primary results we emphasize also differ little, with one key exception, when we use the full set of other medical and mechanical patent classes as the control group. The exception is that our estimate for the trait we term "appearance" is strongly positive during the Civil War period when we adopt either a simple matching or synthetic control approach but is strongly negative when we use the full sample. The explanation for this discrepancy likely lies in the fact that "appearance" is one of the quality-oriented traits for which the broad set of control classes constitutes a poor control group.

[26]As a natural robustness check, we have confirmed that our results are little changed by matching levels and trends on either three-year moving averages or five-year moving averages.

thetic devices, for the full sample of other medical and mechanical control classes, and the synthetic control group for each trait. The synthetic control procedure successfully brings the baseline means for the control groups much closer to the means for prosthetic devices. Notably, although the mean for appearance is matched quite closely for the World War I sample, the mean for the Civil War control group remains moderately below the mean for prosthetic devices. This reflects both the difficulty of matching quality-oriented traits and the moderate size of our samples of Civil War-era patents relative to World War I-era patents. Consequently, results for our analysis of appearance during the Civil War period ought to be interpreted with caution.

5 Results

This section presents estimates of equations (4), (5), and (6). Subsection 5.1 presents estimates of wartime procurement's effects on overall prosthetic device patenting flows during the Civil War and World War I. Subsection 5.2 presents estimates of changes in the economic attributes emphasized in prosthetic device patents during the wartime patent booms relative to the pre-war periods.

5.1 Overall Patent Flows

Figure 2 provides a graphical illustration of the changes in patenting rates, from pre-war periods to the wartime periods of elevated prosthetic device patenting, that underlie our estimates of equation (4). Panel A presents the distribution of changes for the Civil War era and Panel B presents the distribution for the World War I era. Each observation underlying these histograms represents a patent class in our broadest control group. The dashed vertical lines are placed at the value of the change for prosthetic devices. In the Civil War histogram, the change in prosthetic device patenting is the

rightmost point in the distribution, while the World War I change is quite close to the right end of the distribution. Despite having only two class-by-time period treatment events, these figures provide an initial indication of why wartime increases in prosthetic device patenting are strongly statistically distinguishable from zero when we conduct inference using "randomization tests" (Imbens and Rosenbaum, 2005).

Table 4 presents estimates of equation (4). The estimates presented across the columns differ exclusively with respect to the patent classes that are used as controls. The estimate in column 1 reveals that wartime changes in prosthetic device patenting were roughly 95 log points larger than changes in patenting in all other medical or mechanical patent classes. Columns 2 through 6 reveal that this estimate is only moderately sensitive to using subsets of the broader set of controls. The subsets include other categories matched based on baseline patenting rates (column 2), other medical categories only (column 3), the "miscellaneous" mechanical classes (column 4), metalworking mechanical classes (column 5), and materials processing mechanical classes (column 6).[27] The estimates range from 85 log points to 102 log points.

Table 5 presents estimates of equation (5). The estimates in table 5 differ from the estimates in table 4 exclusively by model choice. That is, they are estimates of the Poisson model described by equation (5) rather than the OLS model described by equation (4). All estimates are between 0.54 and 0.88, suggesting that wartime demand shocks led to large increases in flows of prosthetic device patents.

---

[27]Our restriction of the control group to other medical technology classes (column 3), is similar to the approach taken by Moser, Voena, and Waldinger (2014) in their analysis of chemicals patenting. We obtain similar, though modestly smaller, results when further narrowing our control group to the sub-category "Miscellaneous-Drugs and Medicine," which also contains Prosthesis innovation. This sub-category is quite small during these periods, however, as it comprised of only two other classes, namely "Optics: Eye Examining, Vision Testing and Correcting" and "Dentistry." A further issue facing this approach to selecting control classes is that optics and dentistry are medical categories for which it is plausible that the Civil War and World War I may have had a direct effect. This may contribute to why we obtain moderately smaller point estimates when using these control classes rather than a broader control group. For details, we refer readers to the descriptions of the technology classes that are available on the website for the NBER patent database: http://www.nber.org/patents/.

Cluster-robust standard errors are presented in parenthesis below our estimates in table 4 and table 5. The standard errors in table 5 are smaller than the standard errors from table 4, suggesting that the Poisson model may better fit the statistical properties of the count data we analyze, resulting in efficiency gains. However, both models may result in cluster-robust standard errors that are insufficiently conservative due to the small number of "treated patent class episodes" in our sample, namely two.

Our empirical setting falls into the class of settings flagged by Bertrand, Duflo, and Mullainathan (2004) and Cameron, Gelbach, and Miller (2008), where conventional cluster-robust standard errors may result in insufficiently conservative inference. In such settings, randomization inference has been found to generate p-values that confer appropriate degrees of statistical significance (Cameron, Gelbach, and Miller, 2008; Imbens and Rosenbaum, 2005). Figure 3 displays our prosthesis point estimates (dashed vertical lines) in the context of distributions generated from three distinct randomization inference procedures.[28] In each case, the "true point estimate" is larger in magnitude than nearly the entirety of the "placebo distribution." One of the 500 estimates exceeds the true estimate when using assignment algorithm A, two when using algorithm B, and zero when using algorithm C. The implication, in each case, is that our estimates are statistically distinguishable from zero at the $p < .01$ level.

As with any difference-in-differences research design, a question we face is whether our estimates might be biased by differential trends in prosthetic device patenting that pre-date the onset of the wars we analyze. The time series presented in figure 1 suggest

[28]We use three distinct procedures for assigning placebo treatment status. In each case, we assign placebo treatment status to two patent class-by-episode observations. The sample from which these are drawn includes mechanical and medical patent classes other than prosthetic devices. For the first procedure (presented in panel A of figure 3), we assign placebo treatment status at random across both treatment episodes. For the second (presented in panel B of figure 3), we assign treatment at random to one patent class from each of the treatment episodes. For the third, we restrict the sample to patent classes that appear in both the Civil War and World War I sub-samples, then assign treatment at random to a single patent class. The dispersion of the distributions of placebo point estimates are only modestly affected by these alternative assignment mechanisms.

quite strongly that this is not the case. Appendix A further investigates this potential issue by presenting estimates of equation (A.1), which is a standard "event study" specification. The resulting pre-war point estimates do not suggest a pre-existing trend.

Interestingly, wartime booms in prosthetic device patenting were not sustained over the long run. This might initially seem puzzling given that the government's commitment to providing limbs was ongoing. Historical context provides evidence, however, that sustained demand for U.S.-manufactured prosthetic limbs was short-lived during both episodes. Following World War I, demand for U.S.-manufactured devices was short-lived because, as mentioned in section 2, the European powers made conscious efforts to develop their own prosthetic device industries. By 1920, moreover, amputees in Germany, Canada, and the United States were documented to prefer adapting to life without a prosthetic (Linker, 2011, p. 114,118). As discussed in section 2, the same was true following the Civil War; an overwhelming majority of Union veterans chose cash over replacement artificial limbs when they were given that choice during the post-war years. Substantial demand for replacement limbs thus may not have materialized. In both settings, the preference for cash over replacement limbs highlights that, contemporaneous innovation notwithstanding, quality remained low in an absolute sense.

5.2 Traits of Wartime Prosthetic Device Patents

We now turn to estimating the effects of wartime procurement on the economic characteristics of prosthetic device patents. Our estimates of equation (6) are presented in table 6, while the underlying time series are presented in figures 4, 5, 6, and 7. In the time series figures, the dashed vertical lines encompass the years during which prosthetic device patenting was elevated, as first shown in figure 1. The p-values reported table 6 are generated using randomization inference within each of the historical episodes taken separately. Several facts of interest emerge from this analysis.

Our first finding concerning patent traits is that, relative to patents in other medical and mechanical categories, prosthetic device patents exhibited an increased emphasis on the trait we term "adjustability" during both war episodes (see figures 4 and 5 for the time series underlying the estimates in table 6). During the Civil War and World War I episodes, the share of prosthetic device patents exhibiting this trait increased by an average of roughly 10 percentage points more than the changes that occurred across the synthetic control groups.

Our finding on "adjustability" is consistent with an important role for economies of scale. That is, as demand increased, manufacturers appear to have shifted away from the construction of bespoke prosthetic limbs. Importantly, this linkage between wartime procurement and the rise of mass production finds support in the historical literature (Guyatt, 2001).[29]

Our next finding is that the Civil War period was associated with across-the-board increases in our cost-oriented production process traits. The average across these traits (namely "cost," "simplicity," "materials," and "adjustability") increased by an economically substantial 0.13 on a base of 0.15. The magnitude of this difference, as well as the underlying time series, is presented in figure 6. This estimate is statistically distinguishable from zero at the 0.01 level, as it is a true outlier relative to the distribution of randomization test outcomes. In contrast, the average across cost-oriented production process traits moved quite modestly during the World War I period. While both periods ushered in substantial emphases on adjustability, Civil War-era prosthetic device patents also exhibit economically substantial shifts towards emphases on "cost," "simplicity,"

[29]In discussing British manufacturing efforts during World War I, for example, Guyatt (2001, p. 311) writes "why the government turned to standardization when it came to considering how best to answer the huge new demand for artificial limbs, the impetus must also have come from the American limb-making industry, now represented in Britain by the three firms at Roehampton. For at least a generation, the American industry had embraced modern theories of manufacturing, introducing greater efficiency in the production process and a 'uniformity of all essential parts' in the limb."

and experimentation with materials. Changes in these three traits were relatively modest during the World War I episode.

Two factors likely contributed to the Civil War period's emphasis on production processes. First, the prosthetic device manufacturing industry was decentralized during this time period, which would have facilitated extensive experimentation. Second, the government's procurement arrangement, namely fixed-price reimbursement of $50 per arm and $75 per leg (roughly $1,000 and $1,500 in 2018 dollars), created a strong incentive for cost-oriented production process innovation. Crucially, these payments were set below both manufacturers' stated costs and the costs implied by a sparse set of records from the 1860 census of manufacturers.[30] As shown in the framework we developed in section 1, this is precisely the form of payment under which firms will be assured to undertake cost-reducing innovation.

Finally, the prosthetic device patents of the Civil War and World War I episodes diverged with respect to their quality-oriented characteristics. Specifically, Civil War-era prosthetic device patents exhibit a substantial increase in emphasis on comfort, while World War I-era prosthetic device patents de-emphasized comfort and increased emphasis on appearance (see figure 7 for the underlying time series). These differences are plausibly linked to changes in institutional views regarding the importance of rehabilitation, re-employment, and social re-integration, as discussed in section 2. Importantly, however, these differences likely reflect contributions from several factors that it would be difficult to empirically disentangle.

[30]Our knowledge of manufacturers' stated (and, unsurprisingly, inflated) costs comes from Hasegawa (2012), while the authors of Hornbeck and Rotemberg (2019) generously shared the relevant manufacturing census records.

5.3 Robustness of Analysis of Patent Traits

This section explores the extent to which our analysis of the direction of prosthetic device innovation is robust to using alternatives to our synthetic control procedure for generating the control groups underlying our baseline estimates. Tables A.4, A.5, A.6, and A.7 present difference-in-differences estimates using different control samples for comparison with the estimates from table 6. The tabulations and changes in table A.4 are based exclusively on our set of 1,200 manually coded patents. Table A.5 reports estimates associated with the full sample as coded using our baseline machine learning model. Table A.6 reports estimates for which the control group is restricted to medical patent classes only, while table A.7 reports estimates that use a simple matching procedure to select the control group rather than the synthetic control procedure.

The estimates in tables A.7 are quite similar to those in table 6. In the Civil War period, essentially all traits have positive point estimates. During the World War I period we see large negative estimates for comfort, while appearance and adjustability have the largest positive estimates. The consistency between the synthetic control and simple matching approaches suggests that our baseline estimates are not sensitive to the choice of matching methodology used to select control groups.

The estimates in tables A.4, A.5, and A.6 reveal that our estimates for appearance are sensitive to whether our analysis uses a data-driven control group. By contrast, estimates for our cost-oriented production process traits are relatively insensitive to estimation using either data-driven controls or broadly selected control classes. The same is true of our estimates for comfort. The sensitivity of our appearance estimates should not be surprising, as appearance is far more relevant to prosthetic devices than to most other medical or mechanical innovations. Selecting a control group that matches a trait's baseline prevalence can provide a more appropriate counterfactual. Nonetheless, the fact that matching is required to select an appropriate control group leads us to be cautious

in interpreting the strength of our quantitative estimates for appearance.

As a final robustness check, we have constructed synthetic controls from a sample of medical and mechanical technology classes that excludes all classes that might be directly affected by wars. In addition to classes involving firearms and ammunition, we exclude surgery, classes with plausible linkages to military uniforms (e.g., boot and shoe making, buckles, etc.) camp equipment (e.g., tents), and several others. Excluding these technology classes from the set of potential "donors" to our synthetic control groups has very little effect on our estimates.

6 Discussion and Conclusion

Our analysis of Civil War and World War I-era prosthetic device patenting yields several findings of potential interest. First, we find that wartime procurement programs were associated with large increases in the volume of prosthetic device patents. We thus add to an existing body of evidence finding that innovation can respond quite strongly to changes in demand.

Second, we find that wartime demand shocks generated increases in emphasis on mass production. During both the Civil War and World War I, manufacturers delivered prosthetics at prices below what might have initially appeared feasible. Patents from both periods suggest shifts away from the production of bespoke artificial limbs. This is consistent with an important role for economies of scale within the supply chain.

Third, cost-conscious innovation, including efforts to introduce new materials and shed extraneous parts, increased substantially during the Civil War. This highlights the potential relevance of the Civil War period's procurement model, which involved fixed-price reimbursement at modest rates. Experts observe that modern medical innovations have tended to bring costly enhancements to quality rather than cost-conscious improve-

ments in productivity (Chandra and Skinner, 2012; Skinner, 2013). Our findings provide a useful counter-example to this tendency. Demand shocks coupled with cost-conscious payment models can steer innovation in a cost-conscious direction.

Fourth, we find that the prosthetic device patents of the Civil War and World War I episodes diverged with respect to dimensions of quality. Civil War-era prosthetic device patents exhibited an increase in emphasis on comfort. By contrast, World War I-era prosthetic device patents de-emphasized comfort and emphasized appearance. These differences are plausibly linked to a World War I era shift in choice away from veterans and towards medical professionals. This shift was associated, in turn, with a heightened emphasis on veteran rehabilitation and re-employment. Importantly, however, these differences between Civil War and World War I-era prosthetic device innovations may stem from several factors that would be difficult to empirically disentangle.

Two key caveats accompany our reading of the evidence. First, we reiterate the standard caveat associated with interpreting flows of patents as flows of innovation. As noted in our discussion of table 1, we are able to directly link Civil War-era patents to manufacturers, market shares, and expert assessments of product quality. Further, medical historians recognize both the Civil War and World War I as key episodes in prosthetic device innovation's history. There is thus little doubt that meaningful advances in prosthetic device innovation occurred during these time periods. The standard caveat, however, ought nonetheless to be borne in mind.

A second caveat involves the limitations of text analysis. As discussed in section 3, seemingly modest reductions in the accuracy of our text analysis models can substantially attenuate our estimates of the effects of wartime procurement on the direction of prosthetic device innovation. While the accuracy of our models is generally quite high, it varies across the variables we construct. Moderately lower accuracy warrants caution, for example, in interpreting our analysis of the attribute we term "materials." Further,

we highlight a key difference between dimensions of product quality and aspects of the production process. Dimensions of product quality can be highly context-specific, which makes it difficult to select control groups. Consequently, we have more confidence in our analyses of attributes that relate to the production process than in our analyses of attributes that capture dimensions of quality.

We conclude by emphasizing our contribution to the use of text analysis tools for economics research. For researchers interested in our particular context, we have generated a novel data set describing the detailed economic content of prosthetic device patents, other medical patents, and all mechanical patents from 1840 to 1940. The full data set stems from our application of a modified supervised machine learning algorithm to manually coded descriptions of 1,200 closely read patents. For researchers who desire to apply similar tools in other settings, we provide a set of best-practice insights to help guide the development and evaluation of text analysis models. As text analysis becomes more popular, in particular when applied to patents, we hope that future researchers will find value in these insights.

References

Abadie, A. Diamond, A., and J. Hainmueller (2010): "Synthetic control methods for comparative case studies: Estimating the effect of California's Tobacco Control Program," *Journal of the American Statistical Association*, 105(490), 493–505.

Acemoglu, D., P. Aghion, L. Bursztyn, and D. Hemous (2012): "The Environment and Directed Technical Change," *American Economic Review*, 102(1), 131–66.

Acemoglu, D., D. Cutler, A. Finkelstein, and J. Linn (2006): "Did Medicare Induce Pharmaceutical Innovation?," *American Economic Review*, 96(2), 103–107.

Acemoglu, D., and J. Linn (2004): "Market Size in Innovation: Theory and Evidence from the Pharmaceutical Industry," *Quarterly Journal of Economics*.

Aghion, P., A. Dechezleprtre, D. Hemous, R. Martin, and J. Van Reenen (2016): "Carbon Taxes, Path Dependency, and Directed Technical Change: Evidence from the Auto Industry," *Journal of Political Economy*, 124(1), 1–51.

Aghion, P., and P. Howitt (1992): "A Model of Growth Through Creative Destruction," *Econometrica: Journal of the Econometric Society*, pp. 323–351.

Akcigit, U., J. Grigsby, and T. Nicholas (2017): "The rise of american ingenuity: Innovation and inventors of the golden age," *NBER Working Paper 23047*.

Andrews, M. (2019): "Comparing historical patent datasets," *Available at SSRN 3415318*.

Arts, S., B. Cassiman, and J. C. Gomez (2018): "Text matching to measure patent similarity," *Strategic Management Journal*, 39(1), 62–84.

Athey, S. (2018): "The Impact of Machine Learning on Economics," in *The Economics of Artificial Intelligence: An Agenda*, ed. by A. K. Agrawal, J. Gans, and A. Goldfarb. University of Chicago Press.

Barnes, J. (1865): *Artificial Limbs*, Circular Order. Office of the Surgeon General.

Barnes, J., and E. Stanton (1866): *Artificial Limbs Furnished to Soldiers*, Ex. Doc. 108. Department of War.

Baten, J., N. Bianchi, and P. Moser (2017): "Compulsory licensing and innovation–Historical evidence from German patents after WWI," *Journal of Development Economics*, 126, 231–242.

Bergeaud, A., Y. Potiron, and J. Raimbault (2017): "Classifying patents based on their semantic content," *PLoS ONE*, 12.

Bergstra, J., and Y. Bengio (2012): "Random Search for Hyper-parameter Optimization," *J. Mach. Learn. Res.*, 13, 281–305.

Berkes, E. (2018): "Comprehensive Universe of U.S. Patents (CUSP): Data and Facts," *Unpublished Working Paper*.

Berkes, E., R. Gaetani, and M. Mestieri (2019): "Cities and Technology Cycles," *Unpublished Working Paper*.

Berkes, E., and P. Nencka (2019): "Novel Ideas: The Effects of Carnegie Libraries on Innovation," *Unpublished Working Paper*.

Bertrand, M., E. Duflo, and S. Mullainathan (2004): "How Much Should We Trust Differences-in-Differences Estimates?," *Quarterly Journal of Economics*, 119(1).

Blume-Kohout, M. E., and N. Sood (2013): "Market size and innovation: Effects of Medicare Part D on pharmaceutical research and development," *Journal of Public Economics*, 97, 327–336.

Boseley, S. (2016): "NHS 'abandoning' thousands by rationing hepatitis C drugs," *The Guardian*, 28.

Breiman, L. (2001): "Random Forests," *Machine Learning*, 45(1), 5–32.

Brodersen, K. H., C. S. Ong, K. E. Stephan, and J. M. Buhmann (2010): "The balanced accuracy and its posterior distribution," in *2010 20th International Conference on Pattern Recognition*, pp. 3121–3124. IEEE.

Budish, E., B. N. Roin, and H. Williams (2015): "Do firms underinvest in long-term research? Evidence from cancer clinical trials," *American Economic Review*, 105(7), 2044–85.

Cameron, A. C., J. B. Gelbach, and D. L. Miller (2008): "Bootstrap-based improvements for inference with clustered errors," *The Review of Economics and Statistics*, 90(3), 414–427.

Chan, D. C., and M. J. Dickstein (2019): "Industry Input in Policy Making: Evidence from Medicare," *The Quarterly Journal of Economics*, 134(3), 1299–1342.

Chandra, A., and J. Skinner (2012): "Technology growth and expenditure growth in health care," *Journal of Economic Literature*, 50(3), 645–80.

Clemens, J. (2013): "The effect of us health insurance expansions on medical innovation," *NBER Working Paper 19761*.

Cockburn, I. M., R. Henderson, and S. Stern (2018): "The Impact of Artificial Intelligence on Innovation: An Exploratory Analysis," in *The Economics of Artificial Intelligence: An Agenda*, ed. by A. K. Agrawal, J. Gans, and A. Goldfarb. University of Chicago Press.

Cogan, J. F. (2017): *The high cost of good intentions: A history of US Federal entitlement programs*. Stanford University Press.

Cohen, L., U. Gurun, and S. D. Kominers (2014): "Patent Trolls: Evidence from Targeted Firms," *NBER Working Paper 20322*.

Cortes, C., and V. Vapnik (1995): "Support-Vector Networks," *Machine Learning*, 20(3), 273–297.

Cutler, D. (2004): *Your money or your life: strong medicine for America's health care system*. Oxford University Press, USA.

Dechezlepretre, A., D. Hemous, M. Olsen, and C. Zanella (2019): "Automating Labor: Evidence from Firm-level Patent Data," *Unpublished Working Paper*.

Devlin, J., M. Chang, K. Lee, and K. Toutanova (2018): "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *CoRR*, abs/1810.04805.

Doran, K., and C. Yoon (2018): "Immigration and Invention: Evidence from the Quota Acts," *Unpublished Working Paper*.

Dubois, P., O. De Mouzon, F. Scott-Morton, and P. Seabright (2015): "Market size and pharmaceutical innovation," *RAND Journal of Economics*, 46(4), 844–871.

Figg, L., and J. Farrell-Beck (1993): "Amputation in the Civil War: physical and social dimensions," *Journal of the History of Medicine and Allied Sciences*, 48(4), 454–475.

Finkelstein, A. (2004): "Static and dynamic effects of health policy: Evidence from the vaccine industry," *The Quarterly Journal of Economics*, 119(2), 527–564.

Freemon, F. R. (1993): *Microbes and minie balls: an annotated bibliography of Civil War medicine*. Fairleigh Dickinson Univ Pr.

Friedman, J. (2001): "Greedy function approximation: a gradient boosting machine," *Annals of Statistics*, 29, 1189–1232.

Garcia, D. (2013): "Sentiment during recessions," *The Journal of Finance*, 68(3), 1267–1300.

Garrison, F. H. (1917): "The statistical lessons of the Crimean War," in *The Military Surgeon*, ed. by J. V. R. Hoff, vol. XLI, pp. 457–473. The Association of Military Surgeons of the United States.

Gentzkow, M., J. Shapiro, and M. Taddy (2019): "Measuring Group Differences in High-Dimensional Choices: Method and Application to Congressional Speech," *Econometrica*, 87(4), 1307–1340.

Gentzkow, M., and J. M. Shapiro (2010): "What drives media slant? Evidence from US daily newspapers," *Econometrica*, 78(1), 35–71.

Guyatt, M. (2001): "Better legs: artificial limbs for British veterans of the First World War," *Journal of Design History*, 14(4), 307–325.

Guyon, I., and A. Elisseeff (2003): "An Introduction to Variable and Feature Selection," *J. Mach. Learn. Res.*, 3, 1157–1182.

Guyon, I., J. Weston, S. Barnhill, and V. Vapnik (2002): "Gene Selection for Cancer Classification using Support Vector Machines," *Machine Learning*, 46(1), 389–422.

Hall, B., A. Jaffe, and M. Trajtenberg (2001): "The NBER patent citation data file: Lessons, insights and methodological tools," *NBER Working Paper 8498*.

Hall, B. H., A. Jaffe, and M. Trajtenberg (2005): "Market Value and Patent Citations," *RAND Journal of Economics*, pp. 16–38.

Hanlon, W. W. (2015): "Necessity is the mother of invention: Input supplies and Directed Technical Change," *Econometrica*, 83(1), 67–100.

Hasegawa, G. R. (2012): *Mending Broken Soldiers: The Union and Confederate Programs to Supply Artificial Limbs*. SIU Press.

Hawkes, N. (2015): "NHS England Drops 16 Medicines from Cancer Drugs Fund," *British Medical Journal*.

Hochreiter, S., and J. Schmidhuber (1997): "Long Short-Term Memory," *Neural Computation*, 9(8), 1735–1780.

Hornbeck, R., and M. Rotemberg (2019): "Railroads, Reallocation, and the Rise of American Manufacturing," in *2019 Meeting Papers*, no. 396. Society for Economic Dynamics.

Houston, M.H., B. J., and L. Joynes (1866): "Report of the Richmond Medical Journal Commission," *Richmond Medical Journal*, pp. 564–571.

Howell, S. T. (2017): "Financing Innovation: Evidence from R&D Grants," *American Economic Review*, 107(4), 1136–64.

Hua, J., Z. Xiong, J. Lowey, E. Suh, and E. R. Dougherty (2004): "Optimal number of features as a function of sample size for various classification rules," *Bioinformatics*, 21(8), 1509–1515.

Iaria, A., C. Schwarz, and F. Waldinger (2018a): "Frontier knowledge and scientific production: Evidence from the collapse of international science," *The Quarterly Journal of Economics*, 133(2), 927–991.

——— (2018b): "Frontier Knowledge and Scientific Production: Evidence from the Collapse of International Science*," *The Quarterly Journal of Economics*, 133(2), 927–991.

Imbens, G. W., and P. R. Rosenbaum (2005): "Robust, accurate confidence intervals with a weak instrument: quarter of birth and education," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 168(1), 109–126.

Ito, K., and J. M. Sallee (2018): "The Economics of Attribute-Based Regulation: Theory and Evidence from Fuel Economy Standards," *The Review of Economics and Statistics*, 100(2), 319–336.

Johnstone, N., I. Hascic, and D. Popp (2008): "Renewable Energy Policies And Technological Innovation: Evidence Based On Patent Counts," *NBER Working Paper 13760*.

Khan, B. Z. (2009): ""War and the Returns to Entrepreneurial Innovation among U.S. Patentees, 1790-1870"," *Brussels economic review*, 52.

——— (2015): "The Impact of War on Resource Allocation:"Creative Destruction," Patenting, and the American Civil War," *Journal of Interdisciplinary History*, 46(3), 315–353.

Khoury, A. H., and R. Bekkerman (2016): "Automatic Discovery of Prior Art: Big Data to the Rescue of the Patent System," *The John Marshall Review of Intellectual Property Law*, 16.

Kim, Y. (2014): "Convolutional Neural Networks for Sentence Classification," *CoRR*, abs/1408.5882.

Kline, P., N. Petkova, H. Williams, and O. Zidar (2019): "Who profits from patents? rent-sharing at innovative firms," *The Quarterly Journal of Economics*, 134(3), 1343–1404.

Knittel, C. R. (2011): "Automobiles on Steroids: Product Attribute Trade-Offs and Technological Progress in the Automobile Sector," *American Economic Review*, 101(7), 3368–99.

Kolata, G., and A. Pollack (2008): "In Costly Cancer Drug, Hope and a Dilemma," *New York Times*, 157(54), 363.

Laffont, J.-J., and J. Tirole (1986): "Using cost observation to regulate firms," *Journal of political Economy*, 94(3, Part 1), 614–641.

Limb, M. (2019): "How NHS investment in proton beam therapy is coming to fruition," *British Medical Journal*, 364, l313.

Linker, B. (2011): *War's Waste: Rehabilitation in World War I America*. University of Chicago Press.

Macleod, G. H. B. (1858): *Notes on the Surgery of the War in the Crimea, with remarks on the treatment of gun-shot wounds*.

Magerman, T., B. V. Looy, B. Baesens, and K. Debackere (2011): "Assessment of Latent Semantic Analysis (LSA) text mining algorithms for large scale mapping of patent and scientific publication documents," *University of Leuven Working Paper*.

Mikolov, T., I. Sutskever, K. Chen, G. S. Corrado, and J. Dean (2013): "Distributed Representations of Words and Phrases and their Compositionality," in *Advances in Neural Information Processing Systems 26*, ed. by C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, pp. 3111–3119. Curran Associates, Inc.

Moser, P. (2005): "How do patent laws influence innovation? Evidence from nineteenth-century world's fairs," *American Economic Review*, 95(4), 1214–1236.

Moser, P. (2012): "Innovation Without Patents-Evidence from the World Fairs. Forthcoming in the," *Journal of Law and Economics*, 55(1), 43–74.

Moser, P., and A. Voena (2012): "Compulsory licensing: Evidence from the trading with the enemy act," *American Economic Review*, 102(1), 396–427.

Moser, P., A. Voena, and F. Waldinger (2014): "German Jewish émigrés and US invention," *American Economic Review*, 104(10), 3222–55.

Newell, R. G., A. B. Jaffe, and R. N. Stavins (1999): "The Induced Innovation Hypothesis and Energy-Saving Technological Change," *The Quarterly Journal of Economics*, 114, 941–975.

Perry, H. R. (2014): *Recycling the disabled: Army, medicine, and modernity in WWI Germany*. Manchester University Press.

Popp, D. (2010): "Innovation and climate policy," *Annu. Rev. Resour. Econ.*, 2(1), 275–298.

——— (2019): "Environmental Policy and Innovation: A Decade of Research," *NBER Working Paper 25631*.

Rogerson, W. P. (1989): "Profit regulation of defense contractors and prizes for innovation," *Journal of Political Economy*, 97(6), 1284–1305.

——— (1994): "Economic incentives and the defense procurement process," *Journal of Economic Perspectives*, 8(4), 65–90.

——— (2003): "Simple menus of contracts in cost-based procurement and regulation," *American Economic Review*, 93(3), 919–926.

Romer, P. M. (1986): "Increasing returns and long-run growth," *Journal of Political Economy*, 94(5), 1002–1037.

Rosenblatt, F. (1961): "Frank Rosenblatt: Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms," *Spartan Books*.

Scott Deerwester, Susan T. Dumais, R. H. (1990): "Indexing by Latent Semantic Analysis," *JASIS*, 41, 391–407.

Shapiro, A. H., M. Sudhof, and D. Wilson (2018): "Measuring news sentiment," Federal Reserve Bank of San Francisco.

Shapiro, A. H., and D. Wilson (2019): "Taking the Fed at its Word: Direct Estimation of Central Bank Objectives using Text Analytics," Federal Reserve Bank of San Francisco.

Shepard, M., K. Baicker, and J. S. Skinner (2019): "Does One Medicare Fit All? The Economics of Uniform Health Insurance Benefits," *NBER Working Paper 26472*.

Shleifer, A. (1985): "A theory of yardstick competition," *RAND Journal of Economics*, pp. 319–327.

Skinner, J. S. (2013): "The costly paradox of health-care technology," *MIT Tech Rev http://www. technologyreview. com/news/518876/the-costly-paradox-of-healthcare-technology/. Published September*, 5, 2013.

Smith, S., J. P. Newhouse, and M. S. Freeland (2009): "Income, insurance, and technology: why does health spending outpace economic growth?," *Health Affairs*, 28(5), 1276–1284.

Tibshirani, R. (1996): "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1), 267–288.

Tirole, J. (1988): *The theory of industrial organization*. MIT press.

Trajtenberg, M. (1989): "The Welfare Analysis of Product Innovations, with an Application to Computed Tomography Scanners," *Journal of Political Economy*, 97(2), pp. 444–479.

Trajtenberg, M. (1990): "A penny for your quotes: patent citations and the value of innovations," *Rand Journal of Economics*, pp. 172–187.

Turney, P., and P. Pantel (2010): "From Frequency to Meaning: Vector Space Models of Semantics," *Journal of Artificial Intelligence Research*, 37, 141–188.

Waldinger, F. (2010): "Quality matters: The expulsion of professors and the consequences for PhD student outcomes in Nazi Germany," *Journal of Political Economy*, 118(4), 787–831.

——— (2011): "Peer effects in science: Evidence from the dismissal of scientists in Nazi Germany," *The Review of Economic Studies*, 79(2), 838–861.

Watzinger, M., and M. Schnitzer (2019): "Standing on the Shoulders of Science," *CEPR Discussion Paper No. DP13766*.

Figures and Tables

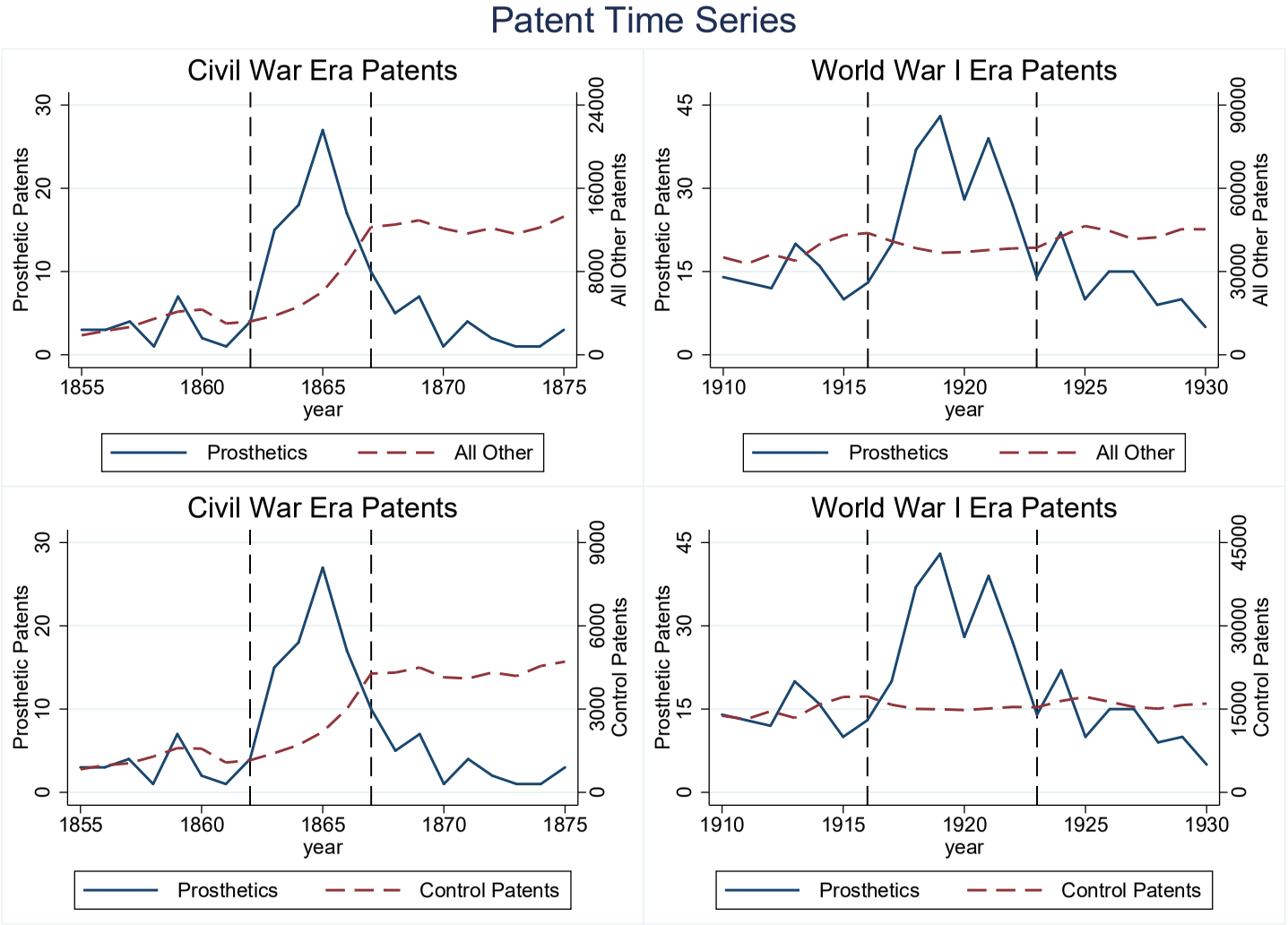
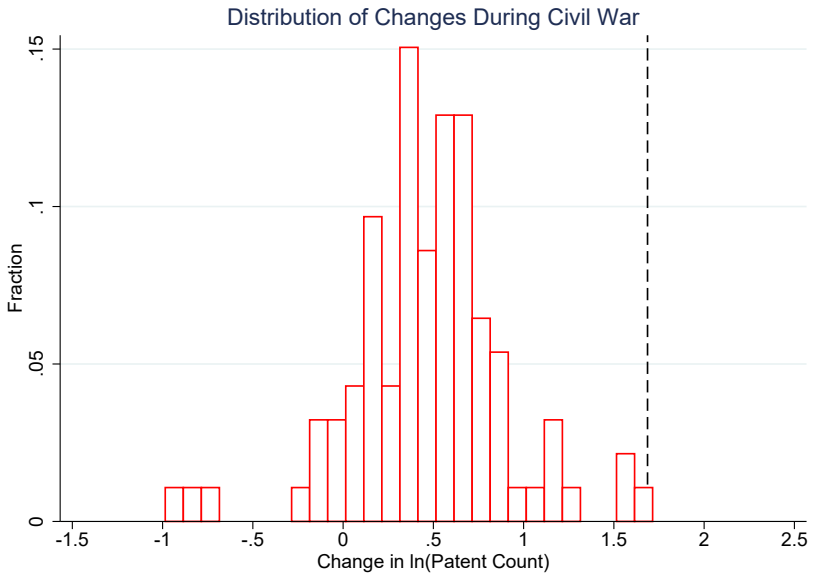


Figure 1: **Patent Time Series**

Note: This figure presents annual time series on patents, using USPTO categories as reported in Berkes (2018). The dashed vertical lines encompass the years we subsequently associate with "war-induced" boom in prosthetic device patenting. In all panels, the solid blue line corresponds with patents from USPTO class 623 "Prosthesis." In the top panels, the red dashed line corresponds with all other patent classes in the database. In the bottom panels, the red dashed line corresponds with patents from "other medical" classes and all "mechanical" classes. These categories are defined using the hierarchical structure of technological categories in the NBER Patent Database (Hall, Jaffe, and Trajtenberg, 2001). In that structure, "other medical" corresponds with all classes in technological category 3 aside from "Prosthesis," while "mechanical" corresponds with category 5.

Patents in Prosthetic Devices and Mechanical Classes

Panel A



Panel B

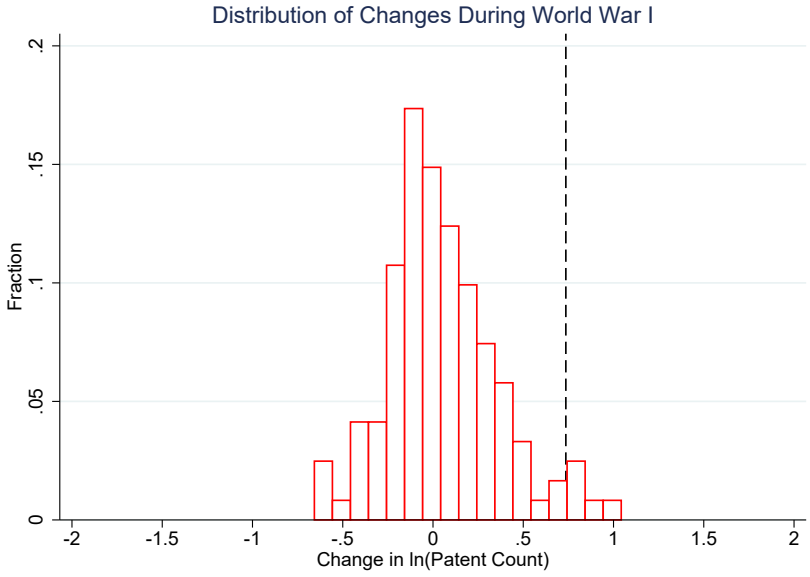


Figure 2: **Patents in Prosthetic Devices and Mechanical Classes:**
Note: This figure presents distributions of changes in the log of patents per year. Each data point in each distribution corresponds with a change for an individual USPTO class. The changes in panel A are calculated from a "base" period extending from 1855 to 1861 to a "war" period extending from 1862 to 1866. The changes in panel B are calculated from a "base" period extending from 1910 to 1915 to a "war" period extending from 1916 to 1922. The vertical dashed line in each panel corresponds with the change that occurred in USPTO class 623 "Prosthesis."

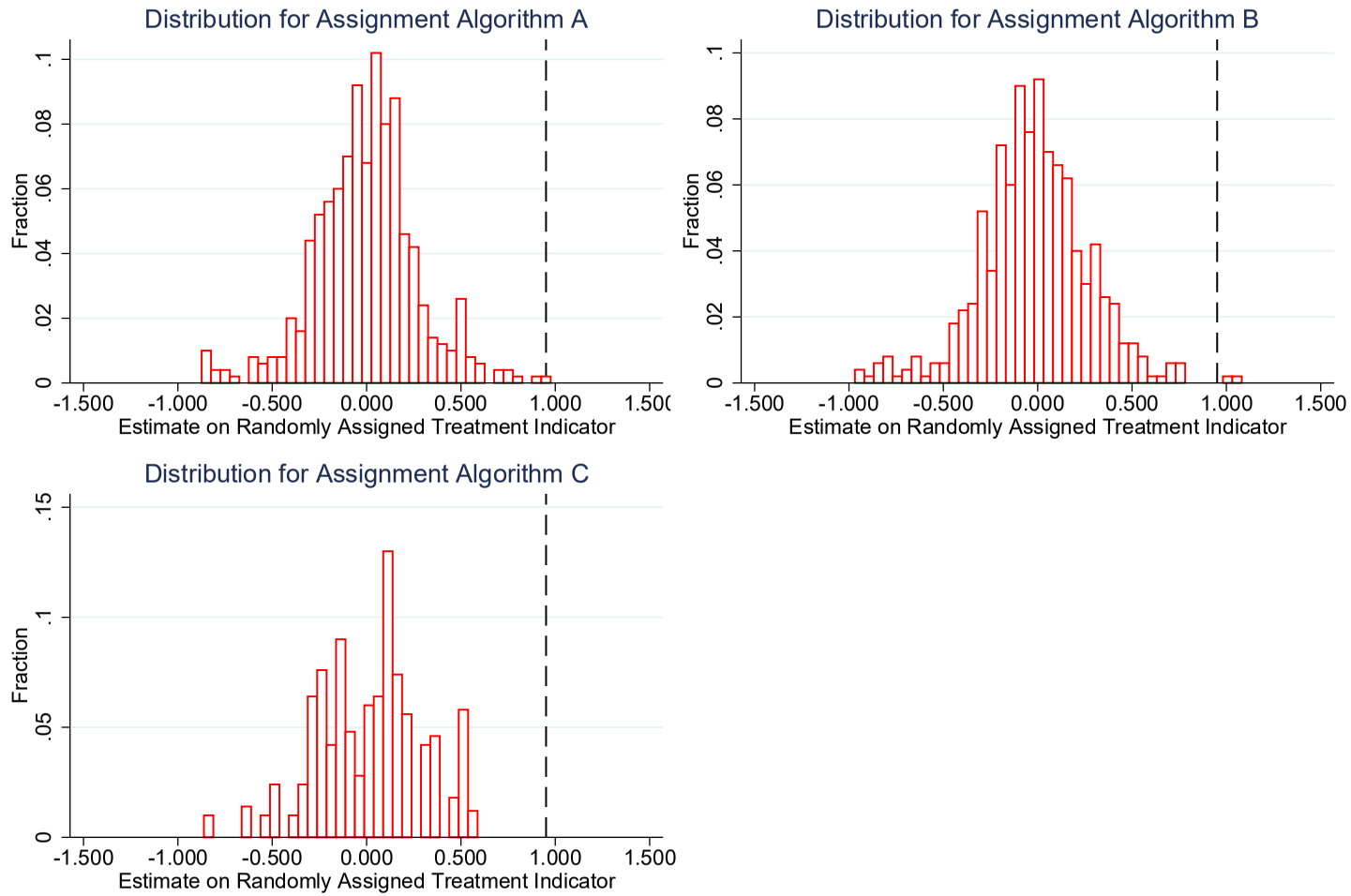


Figure 3: **Placebo Point Estimate Distributions across Three Algorithms**

Note: The figure presents distributions of placebo point estimates generated through the application of a randomization inference procedure (Imbens and Rosenbaum, 2005). The distribution in each panel corresponds with a different algorithm for assigning placebo treatment status. In each case, we assign placebo treatment status to two patent class-by-episode observations. The sample from which these are drawn includes all mechanical and medical patent classes other than prosthetic devices. For Panel A, we assign placebo treatment status at random across this full set of episodes. For Panel B, we assign treatment at random to one patent class from each of the war episodes. For Panel C, we restrict the sample to patent classes that appear in both the Civil War and World War I sub-samples, then assign treatment at random to a single patent class. In each panel, the true estimate associated with assigning treatment status to "Prosthesis" is presented by the dashed vertical lines.

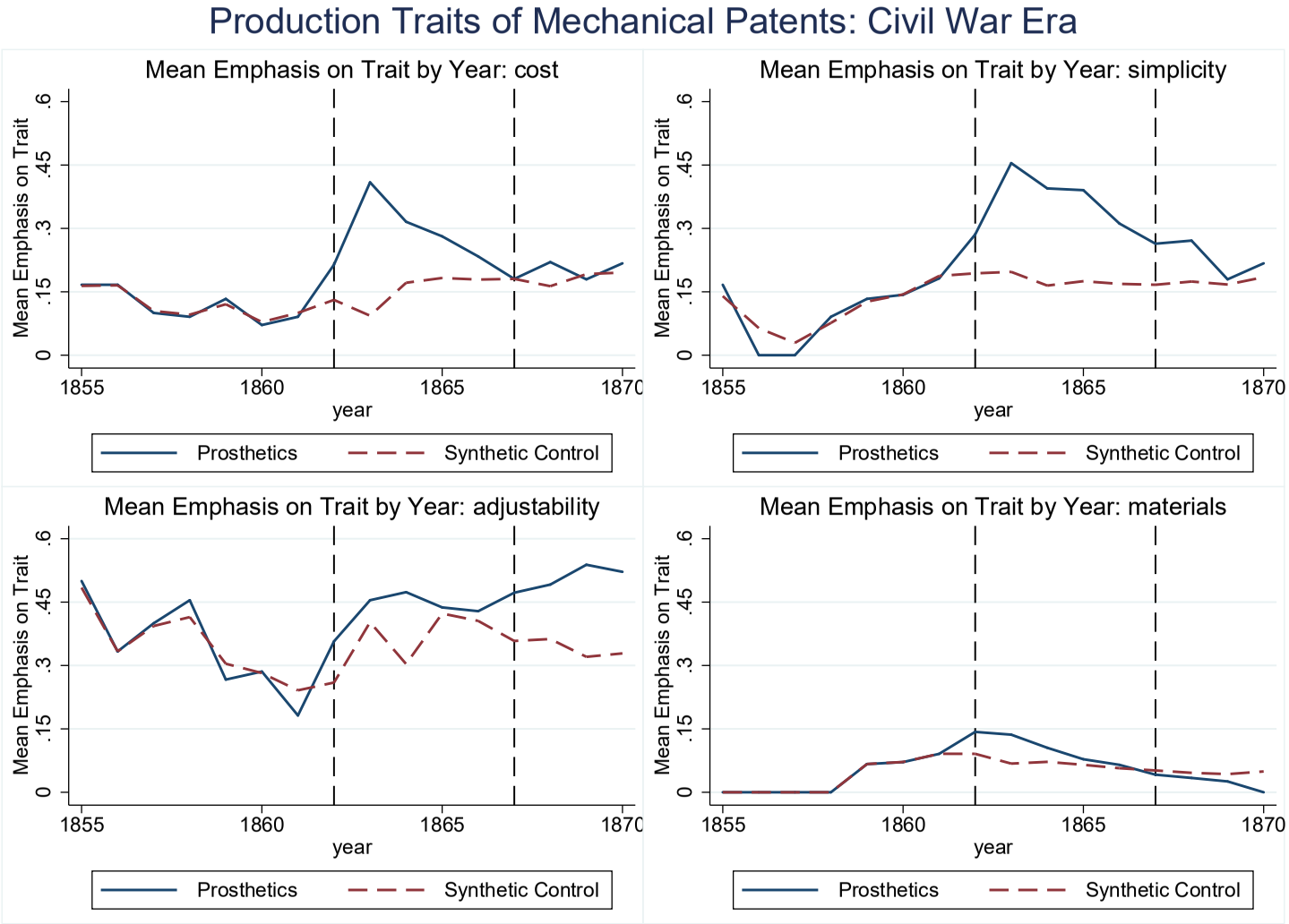


Figure 4: **Cost-Oriented Traits: Civil War Synthetic Controls**

Note: The figure presents "treatment" and "synthetic control" series that describe the evolution of patents' emphases on the traits we term "cost," "simplicity," "adjustability," and "materials." Further information on the definitions of each trait can be found in table 2 as well as in the main text. All series in the figure are calculated as 4-year moving averages. The series plot the share of patents in a given class ("Prosthesis" or the "Synthetic Control") that emphasize a given trait. We generate the synthetic control group using the "synth" package written by Abadie and Hainmueller (2010). "Donor weights" are chosen to match the treatment group on values extending from 1855 to 1861.

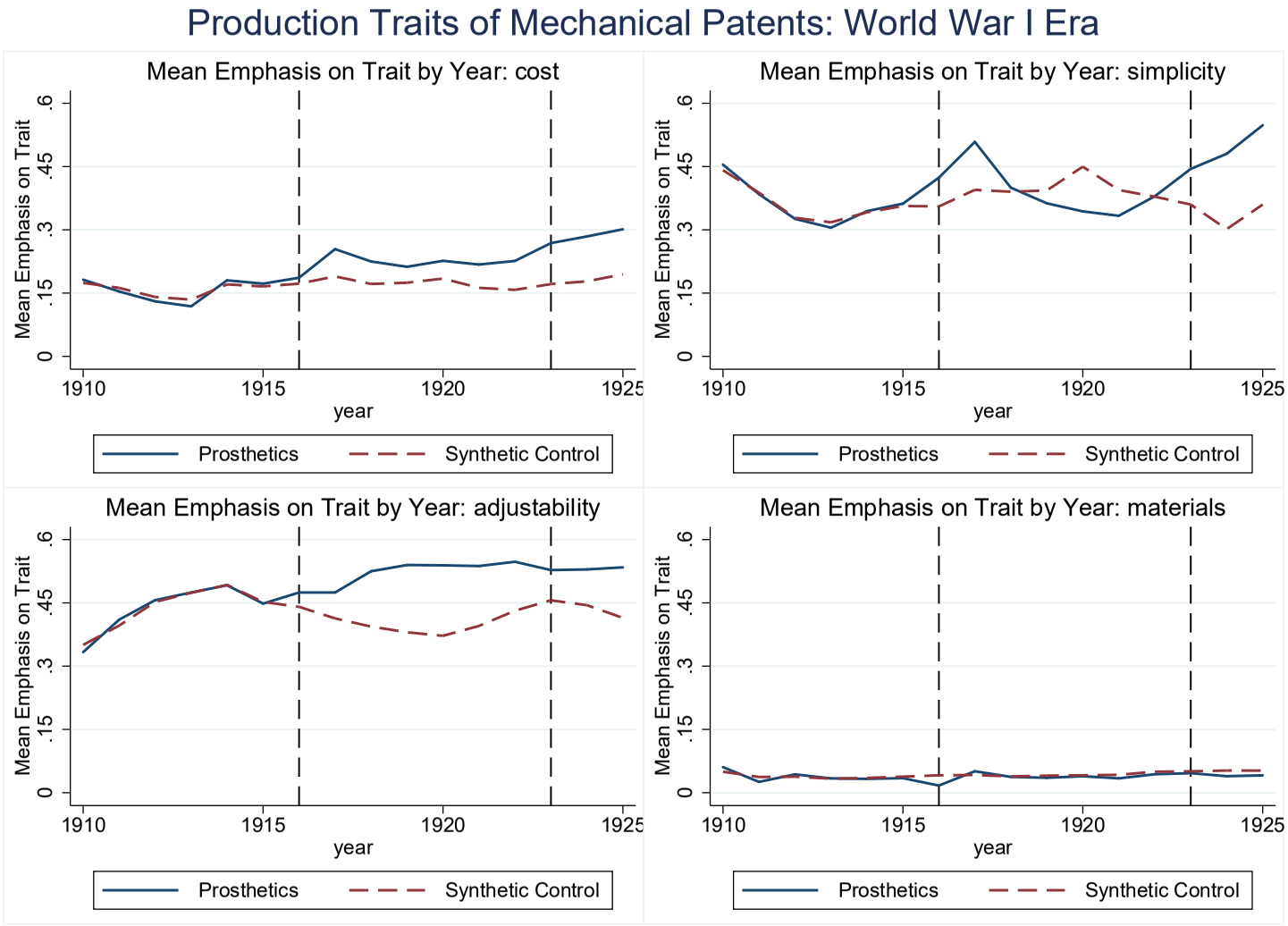


Figure 5: **Cost-Oriented Traits: World War I Synthetic Controls**

Note: The figure presents "treatment" and "synthetic control" series that describe the evolution of patents' emphases on the traits we term "cost," "simplicity," "adjustability," and "materials." Further information on the definitions of each trait can be found in table 2 as well as in the main text. All series in the figure are calculated as 4-year moving averages. The series plot the share of patents in a given class ("Prosthesis" or the "Synthetic Control") that emphasize a given trait. We generate the synthetic control group using the "synth" package written by Abadie and Hainmueller (2010). "Donor weights" are chosen to match the treatment group on values extending from 1910 to 1915.

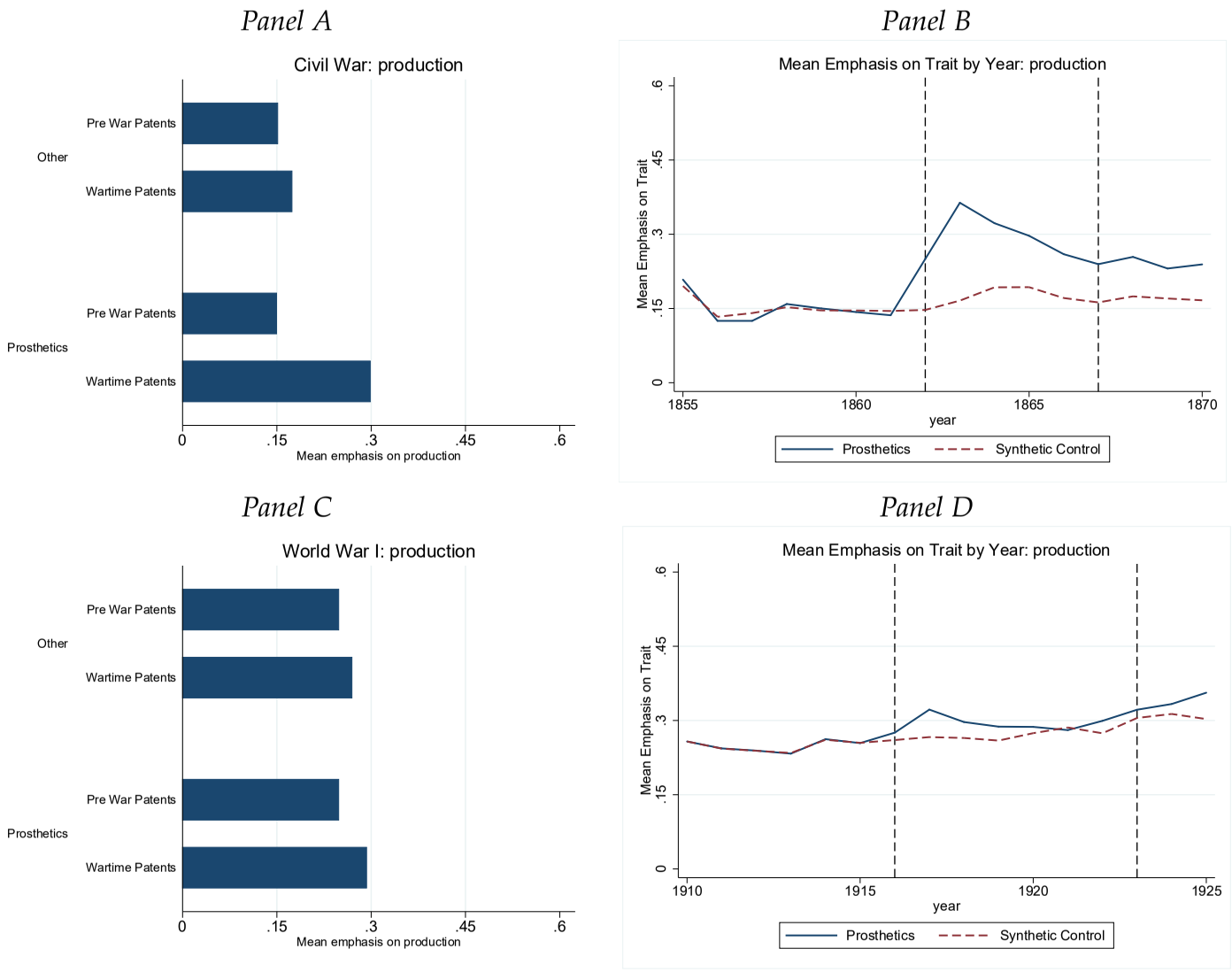


Figure 6: **Changes in the Average across Cost-Oriented Traits:** Note: The figure presents data on "treatment" and "synthetic control" series that describe the evolution of patents' emphases on averages across the cost-oriented traits we term "cost," "simplicity," "adjustability," and "materials." The time series in Panels B and D are calculated as 4-year moving averages. The bar charts in Panels A and C present averages of the "Prosthesis" and "Synthetic Control" series. The series plot the share of patents in a given class ("Prosthesis" or the "Synthetic Control") that emphasize a given trait. In Panel A, the "Pre War" baseline extends from 1855 to 1861 and the "Wartime" period extends from 1862 to 1866. In Panel C, the "Pre War" baseline extends from 1910 to 1915 and the "Wartime" period extends from 1916 to 1922. We generate the synthetic control group using the "synth" package written by Abadie and Hainmueller (2010). "Donor weights" are chosen to match the treatment group on values extending from 1910 to 1915.

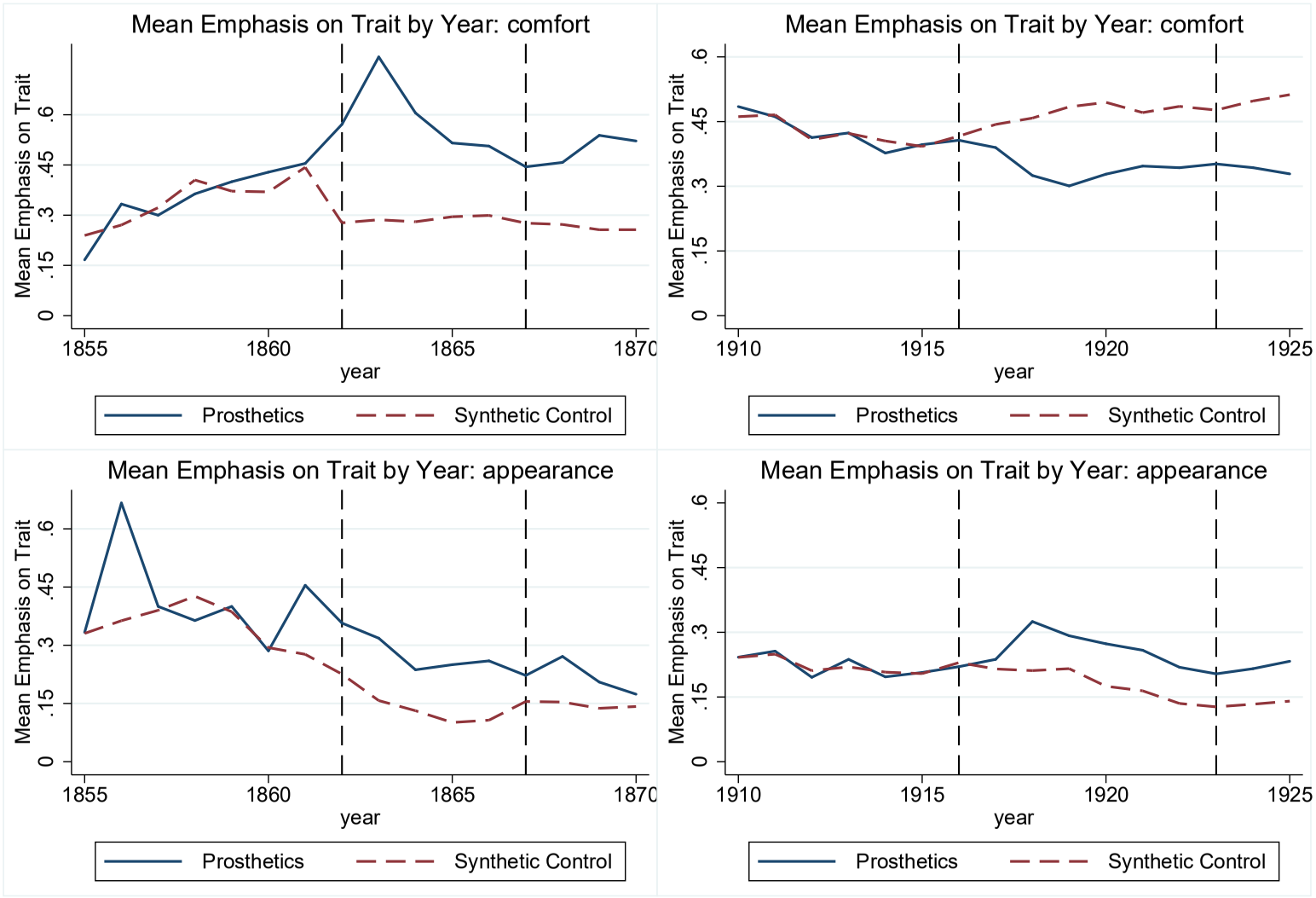


Figure 7: **Quality-Oriented Traits: Civil War and World War I Synthetic Controls**

Note: The figure presents "treatment" and "synthetic control" series that describe the evolution of patents' emphases on the traits we term "comfort" "appearance." Further information on the definitions of each trait can be found in table 2 as well as in the main text. All series in the figure are calculated as 4-year moving averages. The series plot the share of patents in a given class ("Prosthesis" or the "Synthetic Control") that emphasize a given trait. We generate the synthetic control group using the "synth" package written by Abadie and Hainmueller (2010). For the panels in column 1, "Donor weights" are chosen to match the treatment group on values extending from 1855 to 1861. For the panels in column 2, "Donor weights" are chosen to match the treatment group on values extending from 1910 to 1915.

Table 1: **Civil War Era Device Manufacturers, Patents, Early Market Shares, and Post-War Quality Rankings**

Manufacturer	Patents	First Patent	Market Share	Richmond Ranking	Union Ranking
Panel A: Artificial Legs					
B. F. Palmer	6122, 9200, 137711	1849	30.04	2	3
Douglas Bly	23656, 24002, 25238, 31438, 38549, 38550, 57666, 87624	1859	23.01	1	1
B. W. Jewett Patent Leg Company	16360, 29494	1857	19.27	9*	>4
E. D. Hudson	Copied Palmer's Design	na	10.92	>4	4
William Selpho/ Sepho and Sons	14836, 26378	1856	4.80	1	2
Salem Leg Company	35686, 35937, 44534,49528, 49529, 51593	1862	4.16	9*	>4
Charles Stafford	15831, 16420	1856	2.68	Not Considered	9*
Richard Clement	47281	1865	2.23	>4	>4
A. A. Marks	40763, 46687, 234596, 366494	1863	1.17	9*	>4
American Arm and Leg Company	40956	1863	0.72	>4	9*
National Arm and Leg Company	39599	1863	0.40	9*	9*
Marvin Lincoln	na	na	0.32	Not Considered	9*
James Hanger	Confederate Patents	1863	Large in South	9*	9*
Panel B: Artificial Arms					
Marvin Lincoln	39487	1863	45.51	2	2
Grenell & Co	44638	1864	13.02	1	4
H. A. Gildea	na	na	10.39	9*	4
D. W. Kolbe	45052, 255796	1864	8.58	9*	1
Selpho and Sons	18021	1857	8.53	9*	3
E. Spellerberg	42515, 51238	1864	6.49	9*	9*
National Arm and Leg Company	46158, 46159, 48002	1865	4.17	1	3
B. F. Palmer	22575, 22576	1859	2.45	9*	9*
John Condell	48659	1865	0.00	2	1

Note: The information in the table comes from a variety of sources. The criteria for a manufacturer's inclusion in the table is that he either a) accounted for at least 0.25 percent of the limbs furnished through May 1866, as documented in Barnes and Stanton (1866), or b) was highly rated by either the Union or Richmond post-war ranking. The Richmond Ranking comes from Houston and Joynes (1866). The Union Ranking comes from Barnes (1865). An entry of $9*$ indicates that a limb was considered and rated unfavorably or, in the case of the Union ranking, that it had been approved for reimbursement but was not included in the reported ranking. Both the Union and Richmond rankings of artificial arms had two distinct categories, which results in multiple arms rated "1," "2," etc. An entry of >4 indicates that a limb was considered and rated favorably, but outside of the top 4. Linkages between manufacturers and patents come were generated by the authors using the Google Patent Database and manufacturer names assembled from sources including Hasegawa (2012); Barnes (1865); Houston and Joynes (1866); Barnes and Stanton (1866). Patent dates come from Berkes (2018).

Table 2: **Patent Attributes with Descriptions**

Attribute	Description
Cost	Construction is cheap, economical, and less labor intensive
Simplicity	Device construction is simple and less complex/difficult
Adjustability	Manufactured product adaptable to user specifications
Materials	Made from new materials, substances, compounds, and compositions
Appearance	Natural appearance, life-like, tasteful, and neat
Comfort	Device noted as comfortable, noiseless, and promoting circulation

Note: The table describes the definitions we apply in coding each of the economic attributes on which our analysis focuses. The attributes termed cost, simplicity, adjustability, and materials are the attributes we interpret as involving the production-process, while appearance and comfort are our quality-oriented attributes.

Table 3: **Baseline Summary Statistics for Prosthetic Devices, All Control Classes, and Re-Weighted Synthetic Control Classes**

Panel A: Civil War	Prosthetics	All Controls	Synthetic Controls
cost	0.117	0.193	0.118
simplicity	0.102	0.185	0.110
adjustability	0.346	0.303	0.350
materials	0.0327	0.0550	0.0328
production	0.150	0.184	0.151
appearance	0.415	0.0952	0.352
comfort	0.350	0.0685	0.346

Panel B: World War I	Prosthetics	All Controls	Synthetic Controls
cost	0.156	0.263	0.156
simplicity	0.363	0.391	0.362
adjustability	0.436	0.411	0.436
materials	0.0385	0.0585	0.0386
production	0.248	0.281	0.248
appearance	0.223	0.0708	0.222
comfort	0.426	0.0693	0.426

Note: This table presents baseline means for three samples, namely prosthetics, the "all controls" sample, and the "synthetic controls" sample. Panel A presents baseline means for the Civil War period, for which the baseline extends from 1855 to 1861. Panel B presents baseline means for the World War I period, for which the baseline extends from 1910 to 1915. The "all controls" sample consists of patents from all mechanical classes and all medical classes other than prosthetics. The "synthetic controls" sample was selected to match baseline prosthetics on their values across each year from 1855 to 1861 in panel A, and across each year from 1910 to 1915 in panel B.

Table 4: **Relative Increases in Prosthetic Device Patenting During the Civil War and World War I**

	(1) All Controls	(2) Matched	(3) Medical	(4) Misc. Mech.	(5) Metal Works	(6) Mater. Proc.
Prosthetics x War Boom	0.951***	0.853**	0.981**	0.883***	1.015***	1.021**
	(0.267)	(0.298)	(0.294)	(0.194)	(0.269)	(0.338)
N	432	88	34	128	56	92
Clusters	216	44	17	64	28	46
Estimator	OLS	OLS	OLS	OLS	OLS	OLS
Class-by-Episode Effects	Yes	Yes	Yes	Yes	Yes	Yes
Period Effects	Yes	Yes	Yes	Yes	Yes	Yes
SEs in Parentheses	Clustered	Clustered	Clustered	Clustered	Clustered	Clustered
Randomization Inference	$P < .01$	$P < .01$	$P < .01$	$P < .01$	$P < .01$	$P < .01$

Note: The table presents estimates of equation (4), which is a difference-in-differences style OLS regression model for predicting the log of patent counts per year. Observations are at the patent class-by-time period level. The control group used for each regression is described in the column heading, with additional details available in the main text. Standard errors allow for clusters at the patent class-by-war episode level. *, **, and *** correspond with statistical significance at the 0.05, 0.01, and 0.001 levels respectively. The p-values reported in the table's bottom row are based on the true point estimates position in the distribution of placebo point estimates that are presented in figure 3. They correspond with the p-values implied by randomization inference procedures of the sort proposed and recommended by Imbens and Rosenbaum (2005).

Table 5: **Relative Increases in Prosthetic Device Patenting During the Civil War and World War I**

	(1) All Controls	(2) Matched	(3) Medical	(4) Misc. Mech.	(5) Metal Works	(6) Mater. Proc.
Prosthetics x War Boom	0.812***	0.542**	0.797***	0.776***	0.818***	0.879***
	(0.142)	(0.175)	(0.097)	(0.119)	(0.153)	(0.194)
N	432	88	34	128	56	92
Clusters	216	44	17	64	28	46
Estimator	Poisson	Poisson	Poisson	Poisson	Poisson	Poisson
Class-by-Episode Effects	Yes	Yes	Yes	Yes	Yes	Yes
Period Effects	Yes	Yes	Yes	Yes	Yes	Yes
SEs in Parentheses	Clustered	Clustered	Clustered	Clustered	Clustered	Clustered
Randomization Inference	$P < .01$	$P < .01$	$P < .01$	$P < .01$	$P < .01$	$P < .01$

Note: The table presents estimates of equation (5), which is a difference-in-differences style Poisson regression model for predicting patenting rates. Observations are at the patent class-by-time period level. The control group used for each regression is described in the column heading, with additional details available in the main text. Standard errors allow for clusters at the patent class-by-war episode level. *, **, and *** correspond with statistical significance at the 0.05, 0.01, and 0.001 levels respectively. The p-values reported in the table's bottom row are based on the true point estimates position in the distribution of placebo point estimates that are presented in figure 3. They correspond with the p-values implied by randomization inference procedures of the sort proposed and recommended by Imbens and Rosenbaum (2005).

Table 6: **Differential Changes in the Nature of Prosthetic Device Patents**

	(1)	(2)	(3)	(4)
	Civil War		World War I	
	Estimate	P-Value	Estimate	P-Value
Panel A: Quality-Oriented				
Comfort	0.303	0.016	-0.116	0.017
Appearance	0.078	0.037	0.068	0.008
Panel B: Cost-Oriented				
Adjustability	0.076	0.143	0.116	0.008
Simplicity	0.195	0.011	-0.001	0.434
Cost	0.141	0.054	0.050	0.066
Materials	0.035	0.104	-0.005	0.412
Production Mean	0.126	0.000	0.023	0.066

Note: The table presents estimates of the effect of wartime procurement arranges on the fraction of prosthetic device patents that emphasize a given economic trait. Values in columns labeled "Estimate" are estimates of β from equation (6). The control group underlying each estimate is generated using a synthetic control procedure, which is applied separately for each trait. P-values are generated using randomization inference (Imbens and Rosenbaum, 2005). Estimates and p-values in columns 1 and 2 are on the Civil War sample, while estimates and p-values in columns 3 and 4 are on the World War I sample.

Appendix Material

A Figures and Supplemental Tables

This appendix expands on our description of a piece of analysis on which the main text's details are limited. Figure A.1 presents estimates of the following event-study model:

$$E[N_{t,c}|X_t] = exp(\gamma_{c,w} + \gamma_{t,w} + \sum_{t \neq 0} \beta_t 1\{\text{Prosthetic}\}_c \times 1\{\text{Year of War}\}_t + \varepsilon_{c,t}). \quad \text{(A.1)}$$

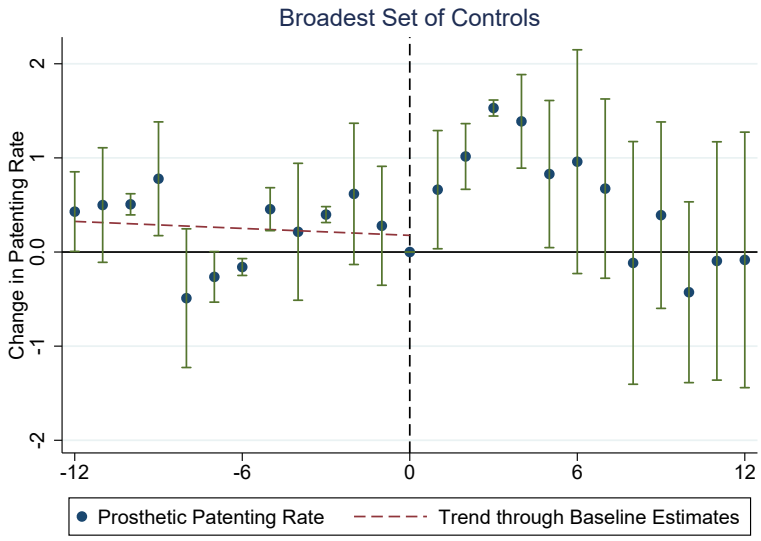
In contrast with our estimates of equations (4) and (5), for which we collapsed the data into multi-year time periods, we estimate equation (A.1) using data that are collapsed at an annual frequency. In the summation, the omitted interaction between the prosthetic device indicator variable and the time dummy variables corresponds with the first full year of either the Civil War or World War I (i.e., the year for which $t = 0$ is the first full year of either war). Each β_t can thus be described as a difference-in-differences style estimate of the change in the prosthetic device patenting rate relative to patenting rates in the control categories from year t relative to the first full year of each war. In panel A, the control patent classes consist of all classes other than prosthetic devices that are either medical or mechanical classes. In panel B, the control patent classes are restricted to other medical classes. Standard errors are clustered at the patent class-by-war episode level. For reasons discussed in the main text, these standard errors are likely to be insufficiently conservative, which motivates our use of randomization methods for inference when we assess the statistical significance of our primary estimates of interest.

The estimates trace out the differential changes one can observe through careful inspection of the time series in figure 1. Crucially, the point estimates associated with years prior to each war (i.e., $t < 0$) exhibit no discernable pattern that might be suggestive of

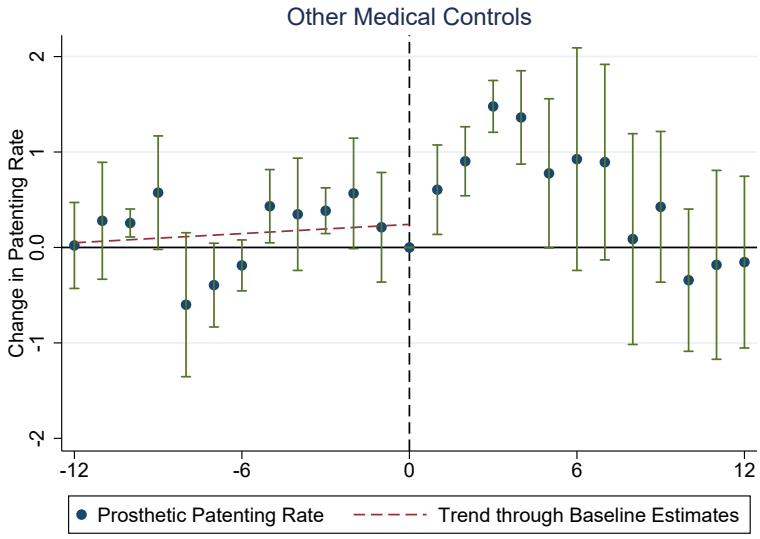
a worrisome pre-existing trend. The point estimate for year $t = -1$ is fairly close to 0, is moderately smaller than the estimates for year $t = -2$ through $t = -5$, is moderately larger than the estimates for $t = -8$ through $t = -6$ and is economically indistinguishable from the estimate for years $t = -9$ through $t = -12$. Prosthetic device patenting exhibits a strong increase relative to the control categories across years $t = 1$ through $t = 7$. There is a notable peak in years $t = 3$ and $t = 4$, which correspond with the 4th and 5th full calendar years following the onset of each war.

Event Study Estimates

Panel A



Panel B



Appendix Figure A.1: **Event Study Estimates of Changes in Prosthetic Device Patenting Rates During the Civil War and World War I:** Note: The figure presents estimates of the β_t coefficients from equation (A.1). Data are analyzed at an annual frequency. The omitted year corresponds with the first full year of either the Civil War or World War I, such that each β_t can be described as a difference-in-differences style estimate of the change in the prosthetic device patenting rate relative to patenting rates in the control categories from year t relative to the first full year of each war. In panel A, the control patent classes consist of all classes other than prosthetic devices that are either medical or mechanical classes. In panel B, the control patent classes are restricted to other medical classes. Standard errors are clustered at the patent class-by-war episode level. For reasons discussed in the main text, these standard errors are likely to be insufficiently conservative, which motivates the use of randomization methods for inference when we assess the statistical significance of our primary estimates of interest.

Appendix Table A.1: Hand-Coded Training Set Tabulations

	(1)	(2)	(3)	(4)
	Civil War		World War I	
	Prosthetics	Controls	Prosthetics	Controls
cost	0.174	0.231	0.235	0.302
simplicity	0.226	0.148	0.394	0.380
adjustability	0.328	0.301	0.450	0.387
materials	0.0462	0.0551	0.0530	0.0852
production	0.194	0.184	0.283	0.289
appearance	0.195	0.0451	0.219	0.0525
comfort	0.497	0.0551	0.371	0.0426
Observations	195	399	302	305

Note: The table presents sample means for the patents in our hand-coded training data set. For the complete hand-coded data set, the patents in the Civil War sample extend from 1840 to 1890, while the patents in the World War I sample extend from 1890 to 1940.

Appendix Table A.2: Full Sample Tabulations

	(1)	(2)	(3)	(4)
	Civil War		World War I	
	Prosthetics	Controls	Prosthetics	Controls
cost	0.186	0.200	0.245	0.299
simplicity	0.247	0.223	0.405	0.426
adjustability	0.423	0.350	0.484	0.407
materials	0.0464	0.0319	0.0435	0.0611
production	0.226	0.201	0.294	0.298
appearance	0.222	0.0605	0.276	0.0741
comfort	0.495	0.0513	0.410	0.0813
Observations	194	151,038	620	593,706

Note: The table presents sample means for the all of the "treatment" and "control" patents in the data set we generate using machine learning methods. For the complete data set, the patents in the Civil War sample extend from 1840 to 1890, while the patents in the World War I sample extend from 1890 to 1940.

Appendix Table A.3: Correlations across Patent Attributes

	cost	simplicity	adjustability	materials	appearance	comfort
cost	1					
simplicity	0.3781	1				
adjustability	0.0613	0.0532	1			
materials	0.0874	0.0207	0.026	1		
appearance	0.0515	-0.0157	0.0339	0.037	1	
comfort	0.0777	-0.0226	0.0834	0.0716	0.0895	1

Note: The table presents a simple correlation matrix across the economic traits we have defined and coded. The sample underlying the matrix is the sample of prosthetic device patents extending from 1840 to 1940.

Appendix Table A.4: Hand-Coded Training Set Tabulations and Changes

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Prosthetics		Other Mechanical			Differences	
Panel A: Civil War	Pre-Boom	Boom	Pre-Boom	Boom	Prosth. Diff	Other Diff	Diff-in-Diff
cost	0.0952	0.247	0.231	0.279	0.152	0.0473	0.104
simplicity	0.0476	0.321	0.132	0.230	0.273	0.0973	0.176
adjustability	0.143	0.370	0.289	0.344	0.228	0.0550	0.173
materials	0.0476	0.0741	0.0826	0.0492	0.0265	-0.0335	0.0599
production	0.0833	0.253	0.184	0.225	0.170	0.0415	0.128
appearance	0.381	0.222	0.0413	0.131	-0.159	0.0898	-0.249
comfort	0.381	0.506	0.0413	0.0492	0.125	0.00786	0.117
Panel B: World War I	Pre-Boom	Boom	Pre-Boom	Boom	Prosth. Diff	Other Diff	Diff-in-Diff
cost	0.188	0.237	0.248	0.360	0.0485	0.112	-0.0632
simplicity	0.365	0.406	0.369	0.393	0.0411	0.0242	0.0169
adjustability	0.424	0.473	0.369	0.400	0.0499	0.0309	0.0190
materials	0.0353	0.0628	0.0671	0.107	0.0275	0.0396	-0.0120
production	0.253	0.295	0.263	0.315	0.0417	0.0516	-0.00983
appearance	0.188	0.222	0.0336	0.0733	0.0340	0.0398	-0.00579
comfort	0.506	0.319	0.0537	0.0333	-0.187	-0.0204	-0.167

Note: The table presents sets of means and changes in means for our hand-coded training data set. The means in columns 1 through 4 are calculated separately for baseline prosthetics, wartime prosthetics, baseline controls, and wartime controls. As in our regressions, the Civil War baseline corresponds with 1855 to 1861 while the World War I baseline extends from 1910 to 1915. The Civil War "wartime" period corresponds with 1862 to 1866 while the World War I "wartime" period extends from 1916 to 1922. Column 5 presents the change from baseline to wartime for prosthetics, while column 6 presents the change from baseline to wartime for the controls. Column 7 presents the difference between these differences.

Appendix Table A.5: Full Sample Tabulations and Changes

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Prosthetics		Other Mechanical		Differences		
Panel A: Civil War	Pre-Boom	Boom	Pre-Boom	Boom	Prosth. Diff	Other Diff	Diff-in-Diff
cost	0.0952	0.247	0.197	0.187	0.152	-0.0105	0.162
simplicity	0.0952	0.333	0.213	0.218	0.238	0.00543	0.233
adjustability	0.286	0.457	0.347	0.318	0.171	-0.0291	0.200
materials	0.0476	0.0741	0.0356	0.0376	0.0265	0.00205	0.0244
production	0.131	0.278	0.198	0.190	0.147	-0.00803	0.155
appearance	0.429	0.247	0.0682	0.0547	-0.182	-0.0135	-0.168
comfort	0.381	0.531	0.0436	0.0420	0.150	-0.00151	0.151
Panel B: World War I	Pre-Boom	Boom	Pre-Boom	Boom	Prosth. Diff	Other Diff	Diff-in-Diff
cost	0.153	0.232	0.270	0.294	0.0789	0.0238	0.0551
simplicity	0.353	0.396	0.412	0.429	0.0432	0.0167	0.0265
adjustability	0.447	0.546	0.397	0.389	0.0988	-0.00807	0.107
materials	0.0353	0.0435	0.0419	0.0467	0.00818	0.00483	0.00336
production	0.247	0.304	0.280	0.290	0.0573	0.00930	0.0480
appearance	0.224	0.256	0.0566	0.0639	0.0325	0.00739	0.0251
comfort	0.447	0.329	0.0658	0.0739	-0.119	0.00815	-0.127

Note: The table presents sets of means and changes in means for the full data set we generate using machine learning methods. The means in columns 1 through 4 are calculated separately for baseline prosthetics, wartime prosthetics, baseline controls, and wartime controls. As in our regressions, the Civil War baseline corresponds with 1855 to 1861 while the World War I baseline extends from 1910 to 1915. The Civil War "wartime" period corresponds with 1862 to 1866 while the World War I "wartime" period extends from 1916 to 1922. Column 5 presents the change from baseline to wartime for prosthetics, while column 6 presents the change from baseline to wartime for the controls. Column 7 presents the difference between these differences.

Appendix Table A.6: Tabulations and Changes with Medical Control Classes Only

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Prosthetics		Other Mechanical			Differences	
Panel A: Civil War	Pre-Boom	Boom	Pre-Boom	Boom	Prosth. Diff	Other Diff	Diff-in-Diff
cost	0.0952	0.247	0.0962	0.142	0.152	0.0460	0.106
simplicity	0.0952	0.333	0.128	0.213	0.238	0.0850	0.153
adjustability	0.286	0.457	0.397	0.355	0.171	-0.0421	0.213
materials	0.0476	0.0741	0.0577	0.0508	0.0265	-0.00693	0.0334
production	0.131	0.278	0.170	0.190	0.147	0.0205	0.126
appearance	0.429	0.247	0.0833	0.0609	-0.182	-0.0224	-0.159
comfort	0.381	0.531	0.308	0.239	0.150	-0.0691	0.219
Panel B: World War I	Pre-Boom	Boom	Pre-Boom	Boom	Prosth. Diff	Other Diff	Diff-in-Diff
cost	0.153	0.232	0.251	0.271	0.0789	0.0199	0.0591
simplicity	0.353	0.396	0.388	0.410	0.0432	0.0222	0.0210
adjustability	0.447	0.546	0.424	0.412	0.0988	-0.0124	0.111
materials	0.0353	0.0435	0.0616	0.0881	0.00818	0.0265	-0.0183
production	0.247	0.304	0.281	0.295	0.0573	0.0140	0.0432
appearance	0.224	0.256	0.113	0.118	0.0325	0.00432	0.0282
comfort	0.447	0.329	0.227	0.240	-0.119	0.0128	-0.131

Note: The table presents sets of means and changes in means for our full data set, but with the control group restricted to medical patent classes only. The means in columns 1 through 4 are calculated separately for baseline prosthetics, wartime prosthetics, baseline controls, and wartime controls. As in our regressions, the Civil War baseline corresponds with 1855 to 1861 while the World War I baseline extends from 1910 to 1915. The Civil War "wartime" period corresponds with 1862 to 1866 while the World War I "wartime" period extends from 1916 to 1922. Column 5 presents the change from baseline to wartime for prosthetics, while column 6 presents the change from baseline to wartime for the controls. Column 7 presents the difference between these differences.

Appendix Table A.7: Crude Matching Sample Tabulations and Changes

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Prosthetics		Other Mechanical			Differences	
Panel A: Civil War	Pre-Boom	Boom	Pre-Boom	Boom	Prosth. Diff	Other Diff	Diff-in-Diff
cost	0.095	0.247	0.115	0.139	0.152	0.024	0.128
simplicity	0.095	0.333	0.119	0.231	0.238	0.112	0.126
adjustability	0.286	0.457	0.281	0.286	0.171	0.004	0.167
materials	0.048	0.074	0.037	0.037	0.026	0.000	0.027
production	0.131	0.278	0.149	0.183	0.147	0.034	0.113
appearance	0.429	0.247	0.423	0.094	-0.182	-0.329	0.147
comfort	0.381	0.531	0.357	0.379	0.150	0.022	0.128
Panel B: World War I	Pre-Boom	Boom	Pre-Boom	Boom	Prosth. Diff	Other Diff	Diff-in-Diff
cost	0.153	0.232	0.177	0.219	0.079	0.042	0.037
simplicity	0.353	0.396	0.370	0.399	0.043	0.030	0.014
adjustability	0.447	0.546	0.444	0.437	0.099	-0.007	0.106
materials	0.035	0.043	0.031	0.038	0.008	0.007	0.001
production	0.247	0.304	0.265	0.279	0.057	0.014	0.043
appearance	0.224	0.256	0.214	0.183	0.033	-0.032	0.064
comfort	0.447	0.329	0.420	0.417	-0.119	-0.004	-0.115

Note: The table presents sets of means and changes in means for data sets in which the control group is constrained using a simple matching procedure. Specifically, the control group is selected to include all control-group patent classes for which the baseline mean is within 6 percentage points of the mean for prosthetic devices for a given economic trait. The one exception is "comfort" during the World War I episode, for which the control-group patent classes consist of those for which the baseline mean is within 20 percentage points of the mean for prosthetic devices. This reflects the fact that there were no close matches for prosthetic devices with respect to "comfort" during the World War I period. The means in columns 1 through 4 are calculated separately for baseline prosthetics, wartime prosthetics, baseline controls, and wartime controls. As in our regressions, the Civil War baseline corresponds with 1855 to 1861 while the World War I baseline extends from 1910 to 1915. The Civil War "wartime" period corresponds with 1862 to 1866 while the World War I "wartime" period extends from 1916 to 1922. Column 5 presents the change from baseline to wartime for prosthetics, while column 6 presents the change from baseline to wartime for the controls. Column 7 presents the difference between these differences.

B Text Analysis Appendix

In this appendix we discuss our approach to designing, evaluating, and selecting our preferred machine learning algorithm for analyzing the texts of patent documents. We begin by describing our objective and comparing our setting with other uses of text analysis in economics research. We then define key terms and discuss examples of the key threats to successful text analysis, along with our approach to addressing them. Finally, we discuss several dimensions of best practice text analysis.

B.1 Generating Economic Data through Text Analysis

Our goal in conducting text analysis is to create variables that describe the economic content of patent texts. Specifically, we analyze the texts of prosthetic device patents, other medical patents, and mechanical patents to determine whether they emphasize traits we term simplicity, cost, adjustability, materials, comfort, and appearance. We code these traits as binary variables, which are our text analysis outputs.

Our text analysis task shares several key commonalities with recent "sentiment" and "partisanship" analyses, where the objective is to rate the sentiment or the degree of partisanship of a publication, writer, or speaker (Shapiro, Sudhof, and Wilson, 2018; Shapiro and Wilson, 2019; Garcia, 2013; Gentzkow, Shapiro, and Taddy, 2019; Gentzkow and Shapiro, 2010).[31] Key commonalities are as follows. First, the researcher must either obtain or create a data set containing a set of outputs (the "true values" for the variables of interest) corresponding to a set of text inputs (a subset of the texts of interest). A machine learning algorithm then learns a function, or model, that relates these input-

[31]Similarly motivated text analysis exercises have also been used quite recently to study patents. Dechezlepretre, Hemous, Olsen, and Zanella (2019), for example, use a keyword search approach to code patents based on whether they relate to "automation." Cockburn, Henderson, and Stern (2018) similarly use a keyword search approach to track the advance of artificial intelligence through references within patent texts and journal articles.

output pairs. Cross-validation is used to evaluate the model's performance by splitting the manually coded input-output pairs into two sets: one on which the model will be trained and another on which the model's performance will be tested. The train-test split is crucial for reliably evaluating performance, as testing on the same data used for training will tend to produce overly optimistic results due to over-fitting.[32] The selected predictive model is then used to assign values for the output variables of interest to the full set of text inputs. Note that these methods are typically used because resource limitations prevent researchers from closely reading and manually coding true values for the broader set of texts. In our case, for example, the broader set of texts consists of more than 700,000 patent documents.

Our preferred algorithm can be described as a modified supervised machine learning algorithm. Our algorithm is somewhat analogous to algorithms used for sentiment analysis by Shapiro, Sudhof, and Wilson (2018). Straightforward algorithms for sentiment analyses make use of "lexicons" that assign positive and negative values to the sentiment associated with extensive lists of words. A simple "Lexical Methodology," for example, is to assign a document a sentiment score based on the sum or mean of the values assigned to the words in its text by the lexicon. In our setting, this is analogous to determining that a patent emphasizes a particular economic trait if its text contains a keyword with which we associate that trait. Shapiro, Sudhof, and Wilson (2018) discuss how this basic approach can be improved upon through tools that account for context (e.g., "negation rules"). While the word "happy" conveys positive sentiment, for example, the phrase "not happy" conveys the opposite. A similar concern motivates the tool we design, which incorporates a neighborhood of contextual clues to root out false-positive errors.

[32]Testing on the left-out data gives insight regarding how generalizable a model will be to new data. Further, repeating cross-validation using randomized train-test splits decreases the likelihood that high performance is simply a result of an opportunistic split.

B.2 The Central Problems of "Polysemy" and "Synonymy"

When using algorithms to extract economic information from text, researchers must overcome errors driven by the complexity of language. In particular, errors can be generated by variations in a word's meanings across contexts and by similarities in the meanings of multiple words. These issues are commonly termed "polysemy" and "synonymy," respectively (Scott Deerwester, 1990; Magerman, Looy, Baesens, and Debackere, 2011).

Synonymy (multiple words having the same meaning) can lead to false negatives, as an algorithm may fail to account for words that are similar in meaning to an attribute's most intuitive keywords. By contrast, polysemy (when words have multiple, context-dependent meanings) elicits false positives. If an algorithm does not detect a word's distinct contextual meaning, it may falsely connect a text input with the concept of interest (Turney and Pantel, 2010). Polysemy can take multiple forms. In some cases, a word's meaning is straightforwardly negated by the words around it (e.g., the aforementioned difference between "happy" and "not happy"). In other cases, a word's meaning may differ with the subject matter contained in the full text or in a particular sentence (e.g., the meaning of "fork" in the phrases "fork in the road" versus "knife and fork"). The difficulties posed by polysemy and synonymy can be closely related, as a keyword's contextual meaning cannot be learned if the keyword itself is not initially detected.

B.3 Illustrative Examples from Patent Texts

The attributes we analyze exhibit varying degrees of "polysemy" and "synonymy." The attribute we term "simplicity," for example, was relatively straightforward. This is because the language linked to "simplicity" is relatively common across texts; it is unlikely to have ambiguous meaning or numerous synonyms. One prosthetic device

patent, for example, quite explicitly stated that "The object of my invention is to imitate this eccentric motion of the knee-joint in the simplest manner." Another states, "The advantages of my invention are as follows: ... great simplicity, and therefore cheapness." The meaning of simplicity extended quite well to patents in our control classes. One such patent highlights, for example, "that the machinery which we use, as hereinafter described, is simple in construction." The relative ease of classifying simplicity is shown in the high performance, which we define more precisely below, we obtain when training the models we consider. Notably, our preferred model performed quite well in predicting "simplicity" even when the training set contained as few as 100 observations.

By contrast, the attribute we term "comfort" was relatively difficult to work with. Difficulties arose because the language used to indicate a product's "comfort" regularly suffered from ambiguity. Sometimes, the meaning of comfort was quite clear. A straightforward example from prosthetics states "My present invention has for its object the production of an artificial leg constructed on such principles that it will give more strength and durability to the limb, and also ease and comfort to the wearer." A straightforward true positive from a different mechanical class states that "Until the external pressure becomes too great... air [is] allowed to enter the box A, until the person sitting in it feels comfortable." Difficulties arose, however, from polysemous words used to describe discomfort. For example, the word "disturbing" often connotes bodily discomfort in prosthetic device patents. In mechanical classes, by contrast, the word "disturbing" tends to have meanings connected to the device's functionality (e.g., "disconnecting or disturbing the pump"). The difficulties created by such cases translated into poor predictive accuracy when we attempted to train our preferred model on relatively small training sets.[33]

[33]As discussed below, comfort is a trait for which accuracy experienced substantial gains as the size of our training data set increased.

B.4 Assessing a Model's Accuracy

A model's accuracy in a binary classification problem can be well described by the evaluation metrics of "sensitivity" and "specificity." Sensitivity refers to the rate of true positives as a share of all positives, while specificity refers to the rate of true negatives as a share of all negatives. These metrics were particularly well suited for our study as they directly ascertain an algorithm's ability to confront the issues of polysemy and synonymy.

Sensitivity and specificity are related. When specificity is reasonably high, sensitivity measures how well an algorithm addresses synonymy by directly revealing the algorithm's ability to correctly detect the desired characteristics: If included keywords inadequately detect patent characteristics due to excluded synonymous keywords, sensitivity would be low. Whereas, when sensitivity is reasonably high, specificity measures the algorithm's ability to ascertain a keyword's context-specific meaning: If the algorithm correctly detects the absence of a given characteristic in the presence of a keyword, it is identifying contextual cues that nullify a keyword's relevance, causing specificity to increase. If either sensitivity or specificity is very low, however, then the algorithm may arbitrarily assign positive or negative outcomes depending on which outcome occurs most frequently in the training data.

The simple average of sensitivity and specificity is commonly termed the "balanced accuracy score." The balanced accuracy score, averaged across "repeated 10-fold cross-validations," is the criterion we use for model evaluation. We used balanced accuracy, as opposed to other evaluation metrics, as it accounts for class imbalance in the dependent variable—a potential issue common in binary classification tasks.[34] As a rough rule of

[34]In the context of a binary classification problem, class "imbalance" means that there are more/less negative outcomes compared to positive outcomes. See Brodersen, Ong, Stephan, and Buhmann (2010) for a widely cited discussion of the balanced accuracy score's attractive properties in settings where this holds.

thumb, we targeted balanced accuracy scores of at least 90 percent.[35] As shown below, however, incremental improvements in an algorithm's accuracy can have meaningful implications for a research project's estimates of primary interest.

We contrast the performance of our preferred model with models generated by a variety of alternative algorithmic techniques. In cases where text classification is well defined by a set of important words, a natural benchmark for assessing alternative tools is a keyword search. A keyword search algorithm codes patents as emphasizing a particular trait if the document contains any words that are strong markers for the trait. As highlighted below, a keyword search is highly effective at identifying positive outcomes for tasks like ours. It may produce false positives, however, by ignoring contextual cues that nullify a keyword's relevance. Whether this shortcoming outweighs a keyword search's ability to detect positive outcomes depends on the degree of polysemy in a researcher's particular task.

B.5 Our Preferred Algorithm: A Novel Modified ML Approach

We considered several classes of algorithms as potential tools for constructing our data set. These included "unsupervised" machine learning algorithms, "supervised" machine learning algorithms, modified supervised learning algorithms, and simple keyword searches. Our preferred algorithm can be described as a modified supervised learning algorithm. The key modification, which involves constraining the feature space from which the algorithm learns, generated advantages with respect to both accuracy and computing requirements.

Unsupervised learning tools are meant to form meaningful groupings of input data

[35] Another common measure of model performance in binary classification tasks is AUC, the area under the receiver operating characteristic curve. For our "comfort" trait we achieve an AUC score of 0.92 and for our "simplicity" variable we attain an AUC score of 0.95. These scores are quite high, suggesting that positive and negative outcomes are quite distinctly separated as the majority of outcomes are simply determined by the presence of a keyword.

based on some predefined metric (Athey, 2018). In our context, we found that such tools struggled to form groupings that coalesced around the economic attributes we sought to analyze. This problem cannot be resolved through the analysis of larger samples.

Standard supervised machine learning tools take as inputs a feature space generated from the entirety of each document's text. We find that these tools struggled to overcome the problems of synonymy and polysemy.[36] For supervised machine learning tools, we find that the performance of existing algorithms improved, to varying degrees, as we expanded the size of our training set. It is thus possible that these algorithms would reach tolerable accuracy thresholds on training samples of sufficient size. Our analysis is suggestive, however, that generating training samples of sufficient size may be beyond many research projects' scope. Closely reading thousands of patent texts or other context-relevant documents is a resource-intensive process.

We find that simple keyword searches performed quite well in our setting. Notably, the development of our lists of keywords benefited from our experimentation with machine learning. At our project's early stages, we attempted keyword searches based on a combination of intuition and close readings of a small set of patents. This "procedure" performed poorly. The accuracy of our keyword searches increased substantially as we learned more about our domain through close readings of 1,200 patent documents in total. Success with either keyword searches or our modified machine learning approach will tend to require substantial knowledge of the domain one is attempting to analyze.[37]

[36]This may stem from the fact that even after processing the text data (removing stop words, word fragments, etc.), the full sample of patent texts contained over 18,000 features. In a simulation analysis using synthetic data, Hua, Xiong, Lowey, Suh, and Dougherty (2004) simulate error rates across alternative feature space sizes, sample sizes, and algorithms. In their context, they find that the optimal feature size is $N-1$ for uncorrelated features (where N is the sample size) and that the optimal feature size becomes proportional to $\sqrt{N}$ for highly correlated features. Although these findings are not necessarily generalizable, in our case the number of features (when using the full processed patent texts) was $15N$, suggesting that the relatively high number of features is plausibly linked to suboptimal performance.

[37]The success of our modified machine learning tool depended on a combination of manually gathered keywords through close readings and data-driven synonym determination. Although this form of feature selection required extensive domain knowledge, feature selection can be effectively executed using entirely

Both sets of approaches provide ample evidence of the idiom "garbage in, garbage out."

Although keyword searches ultimately performed quite well for our task, their general limitations are worth emphasizing. A keyword search does not, by construction, allow context to inform a word's meaning. This can lead to false-positive errors. In general, it should thus be possible to improve upon keyword searches by allowing contextual clues to inform a word's true meaning within each text.

Our preferred, modified approach connects the knowledge we obtained reading patent documents to the Gradient Boosted Machines algorithm (Friedman, 2001).[38] When constructing this model we directly targeted the issues of synonymy and polysemy. First, while reading 1,200 patent documents, we compiled a non-comprehensive list of keywords that indicate each characteristic. To gather each keyword's synonyms, we mapped all our considered patent text corpora to a vector space.[39] This allows us to model the degree of contextual similarity between words using spatial word proximity, resulting in spatial groupings of keywords and their most relevant synonyms. After adding keywords and their synonyms into the feature space, we then include a flexible neighborhood of text surrounding these words to provide contextualization.[40] We then train the machine learning algorithm with this reduced feature space to obtain more

---

data-driven algorithms (see Guyon, Weston, Barnhill, and Vapnik (2002) and Guyon and Elisseeff (2003)). In our case, however, these purely data-driven approaches selected features that induced worse performance than simply using the full patent text. Accuracy gains only occurred when we used a combination of hand-picked and data-driven feature selection.

[38]This is a "boosted" version of Random Forests (Breiman, 2001) where error terms from previous decision tree predictions inform the construction of subsequent trees.

[39]We use Word2Vec (Mikolov, Sutskever, Chen, Corrado, and Dean, 2013) to construct these word embeddings. Word2vec uses shallow neural networks to map words within text documents to a vector space that captures word relationships through a distance metric. Words within this space are mapped as being close together if they occur in similar contexts in the text corpora.

[40]These steps are well described as a type of "feature selection." Feature selection has been shown to help at "improving the prediction performance of the predictors, providing faster and more cost-effective predictors, and providing a better understanding of the underlying process that generated the data" (Guyon and Elisseeff, 2003),

accurate and efficient results.[41]

Relative to alternative machine learning methods, our modified approach generated accuracy gains when predicting each of our economic characteristics. Improvements relative to machine learning approaches that attempt to learn from the entirety of each patent's text were quite large. The relative success of our modified approach, when compared to other pure machine learning methods, is driven by the amount of extraneous information in patents' full texts, figure descriptions, and detailed claims. The presence of extraneous features reduced these algorithms' ability to pinpoint specific, economically relevant patent characteristics. Constraining the feature space to include only keywords, their synonyms, and neighboring contexts allows the machine learning algorithm to learn more efficiently.

Relative to a keyword search, our algorithm's greatest improvements in accuracy were gains of three percentage points for the quality-oriented traits we term "comfort" and "appearance." The improvement in accuracy comes entirely from gains in specificity: The modified approach learns to discriminate keywords whose context nullifies their meaning. Although a three percentage point gain in accuracy is modest, researchers will tend to realize larger gains for text analysis problems with greater degrees of polysemy.

B.6 Lessons for Implementing Best Practice Text Analysis

In this section, we illustrate several key inputs to best practice text analysis. While text analysis tasks necessarily confront many setting-specific challenges, the dimensions of best practice we discuss should apply quite generally. They include an approach for assessing the optimal size of a training set, the importance of generating a training

[41]Computation time was dramatically reduced using our approach when compared to other machine learning algorithms. This stems from the reduced feature space, allowing quicker model training.

set that covers all contexts that a researcher targets, and an approach for assessing the implications of inaccurate predictions for the estimates in which a study is ultimately interested.

B.6.1 Determining Optimal Sample Size

We conducted a systematic analysis of how the performance of various algorithms evolved as we expanded the size of our training data set. Text analysis tasks may differ substantially with respect to the complexity of each piece of text and with respect to the severity of setting-specific sources of polysemy and synonymy. Consequently, it is not possible to prescribe a "rule-of-thumb" size for a training set. One can nonetheless use the relationship between accuracy and sample size to make inferences regarding the returns to further expansions of the training set.

Using our preferred modified approach, the size of the training set required to reach tolerable balanced accuracy scores varied across traits. For the trait we term simplicity, for example, our balanced accuracy score exceeded 90 percent with training sets containing fewer than 200 observations. For the trait we term comfort, by contrast, the accuracy score approached 90 percent as training sets contained roughly 700 observations. For the trait we term materials, the accuracy score remained below 90 percent even on our full training set of 1,200 observations.

On what basis should the size of the training set be determined? Expanding a training set requires project resources. On the margin, the key question is whether increases in the size of the training set yield non-trivial returns. As a way to gauge the relevant returns, we recommend constructing "learning curves," like those displayed in figure B.3. We constructed these figures by evaluating our model's accuracy when trained and tested on samples of varying sizes. More specifically, we executed a bootstrap estimation of our model's balanced accuracy score when trained on different sample

sizes from our manually coded data, with the remaining un-sampled data forming the test set. The solid green line in each panel traces the mean of the balanced accuracy score across 400 iterations of this procedure at ascending sample sizes. The shaded green area extends from the 10th to the 90th percentiles of the distribution of results. The bootstrap approach assures that our estimate for any given sample size is not skewed by particularly "favorable" or "unfavorable" draws, meaning draws on which the algorithm happens to have a particularly easy or difficult time with its prediction task.

Panel A of figure B.3 shows that the balanced accuracy score for "comfort" is relatively low with small samples. Further, the score for comfort exhibits non-trivial improvement as the training set expands to include as many as 1,000 patents. The band extending from the 10th to the 90th percentiles of the distribution is quite large in comparison with the band presented in panel B, for the trait we term simplicity.

Panel B of figure B.3 shows that the balanced accuracy score for "simplicity" is high with small samples. Further, the score asymptotes quickly. It exhibits no further improvement once the training set includes 400 observations. Notably, the band extending from the 10th to the 90th percentiles of the distribution is relatively tight. This further supports the point that the performance of the algorithm is not particularly dependent on the patent documents used to train it.

Our analysis of alternative machine learning algorithms provides additional evidence that performance can depend crucially on sample size. On samples of the sizes we consider, we found that non-neural network machine learning algorithms perform better than deep learning algorithms and that our modified machine learning approach performs better than both deep learning and non-neural network machine learning models trained on the entire text of each patent.[42]

[42]These results are fairly consistent across the economic traits we analyze. All machine learning hyperparameters are tuned using randomized grid-search methods (Bergstra and Bengio, 2012). Deep learning models we considered were Bidirectional Encoder Representations from Transformers (Devlin, Chang,

B.6.2 Assessing the Stability of Economic Estimates

What constitutes an acceptable accuracy threshold? Alternatively, how can one gauge the implications of incremental changes in model accuracy for the primary estimates of an analysis? We shed light on this question through a simulation of how our estimates evolve as we systematically *reduce* the accuracy of our preferred algorithm's estimates.

The procedure we conduct is straightforward. Starting with the data generated by our preferred modified approach, we inject noise by altering the coding of a given fraction of the observations for an outcome variable of interest. We do this for fractions ranging from 1 percent to 50 percent. We select the observations we miscode at random, then estimate β_1 from equation (6). As in our analysis of "learning curves," we implement a bootstrap-style procedure. That is, for each degree of noise, we repeat the basic procedure 40 times to generate a range of new estimates. Figure B.4 reports the resulting means and distributions.[43]

Panel A of figure B.4 presents estimates for the trait we term "comfort" during the World War I period. Our baseline estimate for comfort is -0.14, indicating that wartime prosthetic device patents were 14 percentage points less likely than pre-war prosthetic device patents (net of the equivalent change for the synthetic control group) to emphasize comfort. As we reduce the accuracy of our comfort variable's coding, this estimate quite rapidly converges towards zero. The magnitude of the estimate for comfort was halved before we had reduced accuracy by 10%.[44]

Lee, and Toutanova, 2018), Convolutional Neural Networks (Kim, 2014), Recurrent Neural Networks with long short-term memory(Hochreiter and Schmidhuber, 1997), and Multi-Layer Perceptrons (Rosenblatt, 1961).

[43]Note that the estimate we produce using the data generated from our preferred model serves as the benchmark. Since our modified approach does not predict with perfect accuracy, the current observations already have a small amount of measurement error corresponding to the error associated with the model's performance in predicting "comfort."

[44]As the accuracy of the data approaches 50%, the estimate converges to zero. As the algorithm's accuracy dips below 50% the estimate will begin to converge to the opposite sign of the true estimate. To

Panel B of figure B.4 presents the sensitivity of estimates of β_1 from equation (6) for "simplicity." Our baseline estimate for simplicity is 0.13, indicating that wartime prosthetic device patents were 13 percentage points more likely than pre-war prosthetic device patents (net of the equivalent change for the synthetic control group) to emphasize simplicity. Interestingly, the rate of convergence to zero differs non-trivially when comparing the estimates for comfort and simplicity. Estimates for simplicity converge more slowly, as the magnitude of the estimate is halved when we had reduced accuracy by roughly 20%.

Coding accuracy is clearly important for generating unbiased estimates in analyses of both comfort and simplicity. In both cases, 20% reductions in accuracy would render the estimates from our analyses much smaller economically. In addition to being economically smaller, the attenuated estimates are less likely to be statistically distinguishable from zero. Differences in the rate of convergence towards zero suggest that the tolerability of error may be higher in the case of simplicity than in the case of comfort. It is not obvious why this is the case. A natural hypothesis, into which more research is needed, is that estimates' sensitivity to reductions in accuracy may depend in part on a trait's baseline prevalence within both the treatment and control groups.

B.6.3 Context Specificity

The performance of a trained model may be limited outside the context of its training data. We term this concept "context specificity." Limitations on a model's validity outside of its training set can result from variations in word meanings and usage across domains and across time. In our case, a model trained to recognize the traits in artificial limb patents may perform poorly when applied to patents from classes we use as con-

---

see why note that altering the coding of 100% of the observations would yield a variable that is the inverse of the original variable.

trols. A model's performance might be impaired if the training set lacks sufficient data from all considered domains.

To illustrate this point, we conduct the following exercise. Our data can be described as consisting of four contexts, namely Civil War-era prosthetic devices, Civil War era control categories, World War I-era prosthetic devices, and World War I era control categories. We train our model on a single context, then asses its accuracy in all four contexts. Doing this for each of the contexts separately generates a total of sixteen balanced accuracy scores, four of which involve applying the model to the context on which it was trained. To ensure that differences in accuracy scores across contexts are not driven by differences in sample size, we constrain the size of the training set to be equal in all cases.

The results of conducting this exercise for our "comfort" and "simplicity" traits can be found in table B.1. In each panel, the main diagonal of the matrix of balanced accuracy scores corresponds to our model being applied to the context on which it is trained. This is done using cross-validation within the given domain and time period. The antidiagonal entries correspond to our model being trained on a different patent class (prosthetic devices vs. the control classes) and historical episode (Civil War vs. World War I) than the corresponding left-out test data set. Differences in the average value of the balanced accuracy scores along the main diagonal relative to the antidiagonal provide information on the relevance of context-specificity.

Consistent with our priors, we find that context-specificity is more important for traits for which the problems of polysemy and synonymy are relatively severe. In the examples presented in table B.1, we find that the difference in accuracy scores when comparing the main diagonal to the antidiagonal is greater for "comfort" than it is for "simplicity." The differences in accuracy scores for comfort are non-trivial. On average, the score along the main diagonal is 92.5 percent, while the average score along the

antidiagonal is 86.5. The difference of 7 percentage points is non-trivial when put in the context of our analysis from the previous section. For comfort, injecting a 7 percentage point reduction in accuracy led our estimate of β_1 from equation (6) to decline by nearly half.

More generally, we find that it is important to account for context specificity when predicting attributes whose meaning is domain- and time-dependent. In our setting, attributes that exhibited this time- and domain-dependence include "appearance", "materials", and "comfort." By contrast, accuracy scores were relatively insensitive to the training set's context for the traits we term "cost," "simplicity," and "adjustability."

B.6.4 Acknowledging Limitations

In some cases, even a well-chosen algorithm trained using a large data set may yield low accuracy scores. Even with our preferred algorithm, for example, we obtained an accuracy score of 87 percent when predicting the trait we term materials. What drives this result and how should it shape our presentation of the evidence?

"Materials" was a difficult trait to predict because keywords that describe the introduction of novel materials tend to have no previous mentions. When few observations contain a keyword, an algorithm's opportunities to learn how best to classify out-of-sample observations with that keyword are limited. Keywords that were consistently used to describe new materials—like material, alloy, chemical, composition, or mixture—also tended to be used in the description of a device's construction whether or not the associated materials were new. Further, new material innovations were relatively rare. They occurred in only six percent of the observations in our sample, resulting in a small number of reliable positive observations.

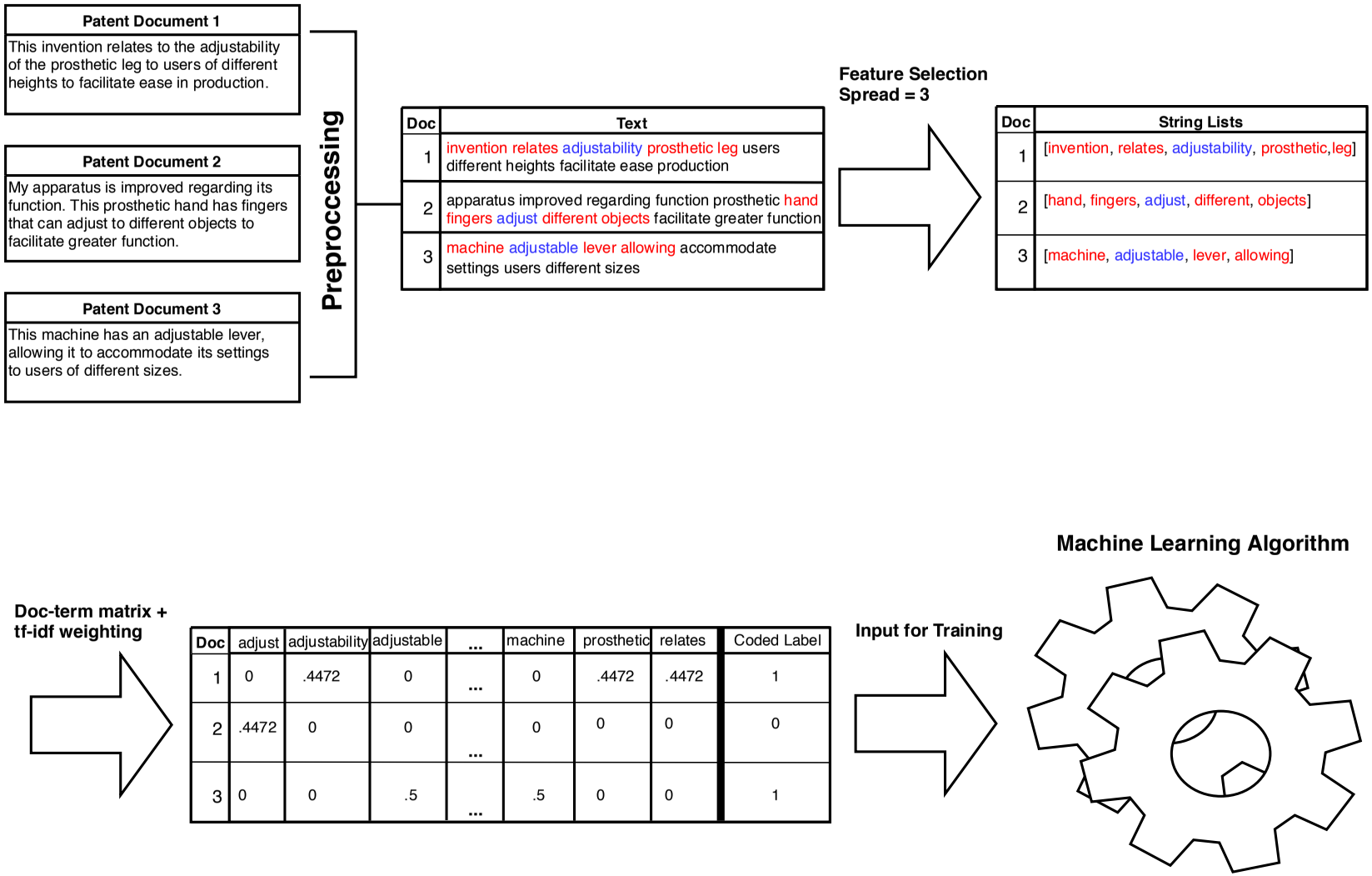
As shown earlier, reductions in model accuracy tend to attenuate our estimates. Properly interpreting our estimates thus requires knowing the accuracy of the model used to

generate the dependent variable. We recommend presenting two key pieces of information. First, analyses of this sort should present readers with an accuracy metric that is appropriate to the setting.[45] In table B.2, for example, we present the full set of balanced accuracy scores along with the underlying sensitivity and specificity scores. Second, "stability curves" of the sort we present in section B.6.2 provide valuable information for inferring the biases associated with inaccurate predictions. We thus recommend coupling these key pieces of information within a discussion of the implications of prediction errors.

In some cases, predictive accuracy may be sufficiently low that the resulting biases will lead point estimates to be highly misleading. In such cases, we recommend that readers be directly warned to interpret the estimates "with caution." In some cases, it may be possible to pair this caution with the best estimate of the potential magnitude of the associated bias. If the only bias is a straightforward form of attenuation bias, then interpretable estimates can be recovered by applying a correction factor. If a correction factor cannot be estimated, the best approach may be to describe estimates as being useful for "illustrative purposes" only.

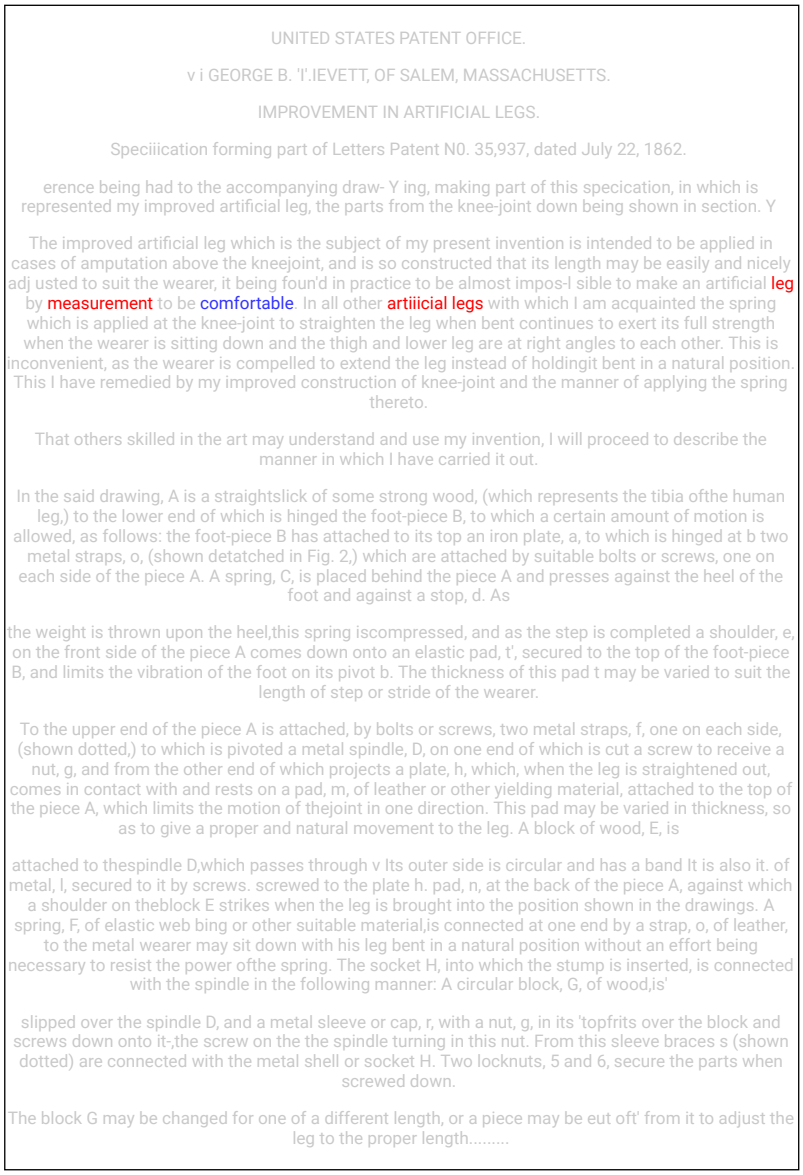
[45]While the balanced accuracy score is a sensible metric for our setting, alternative metrics might be more suitable elsewhere.

Appendix Figure B.1: Flowchart of Modified Approach for Adjustability Characteristic



Note: The figure presents a flowchart of our modified approach. First, the text documents are preprocessed by correcting spelling errors, setting characters to lowercase, removing stop words, punctuation, word fragments, numbers, and extremely frequent or rare words. Then we select keywords and their surrounding context as features. After, we create a doc-term matrix with each entry representing the tf-idf weighting of relative importance. Lastly, this doc-term matrix is fed into the machine learning algorithm for training.

Appendix Figure B.2: Patent Document Example for "Comfort" with Spread $= 3$



UNITED STATES PATENT OFFICE.

v i GEORGE B. 'I'.IEVETT, OF SALEM, MASSACHUSETTS.

IMPROVEMENT IN ARTIFICIAL LEGS.

Speciiication forming part of Letters Patent N0. 35,937, dated July 22, 1862.

erence being had to the accompanying draw- Y ing, making part of this specication, in which is represented my improved artificial leg, the parts from the knee-joint down being shown in section. Y

The improved artificial leg which is the subject of my present invention is intended to be applied in cases of amputation above the kneejoint, and is so constructed that its length may be easily and nicely adj usted to suit the wearer, it being foun'd in practice to be almost impos-I sible to make an artificial leg by measurement to be comfortable. In all other artiiicial legs with which I am acquainted the spring which is applied at the knee-joint to straighten the leg when bent continues to exert its full strength when the wearer is sitting down and the thigh and lower leg are at right angles to each other. This is inconvenient, as the wearer is compelled to extend the leg instead of holdingit bent in a natural position. This I have remedied by my improved construction of knee-joint and the manner of applying the spring thereto.

That others skilled in the art may understand and use my invention, I will proceed to describe the manner in which I have carried it out.

In the said drawing, A is a straightslick of some strong wood, (which represents the tibia ofthe human leg,) to the lower end of which is hinged the foot-piece B, to which a certain amount of motion is allowed, as follows: the foot-piece B has attached to its top an iron plate, a, to which is hinged at b two metal straps, o, (shown detatched in Fig. 2,) which are attached by suitable bolts or screws, one on each side of the piece A. A spring, C, is placed behind the piece A and presses against the heel of the foot and against a stop, d. As

the weight is thrown upon the heel,this spring iscompressed, and as the step is completed a shoulder, e, on the front side of the piece A comes down onto an elastic pad, t', secured to the top of the foot-piece B, and limits the vibration of the foot on its pivot b. The thickness of this pad t may be varied to suit the length of step or stride of the wearer.

To the upper end of the piece A is attached, by bolts or screws, two metal straps, f, one on each side, (shown dotted,) to which is pivoted a metal spindle, D, on one end of which is cut a screw to receive a nut, g, and from the other end of which projects a plate, h, which, when the leg is straightened out, comes in contact with and rests on a pad, m, of leather or other yielding material, attached to the top of the piece A, which limits the motion of thejoint in one direction. This pad may be varied in thickness, so as to give a proper and natural movement to the leg. A block of wood, E, is

attached to thespindle D,which passes through v Its outer side is circular and has a band It is also it. of metal, I, secured to it by screws. screwed to the plate h. pad, n, at the back of the piece A, against which a shoulder on theblock E strikes when the leg is brought into the position shown in the drawings. A spring, F, of elastic web bing or other suitable material,is connected at one end by a strap, o, of leather, to the metal wearer may sit down with his leg bent in a natural position without an effort being necessary to resist the power ofthe spring. The socket H, into which the stump is inserted, is connected with the spindle in the following manner: A circular block, G, of wood,is'

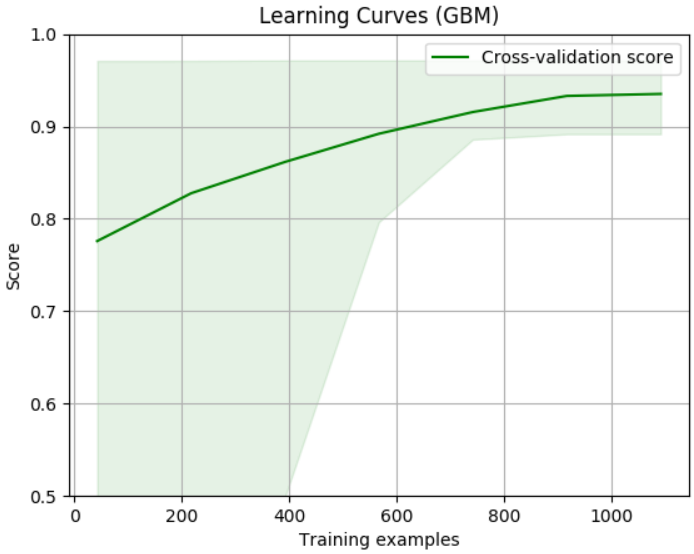
slipped over the spindle D, and a metal sleeve or cap, r, with a nut, g, in its 'topfrits over the block and screws down onto it-,the screw on the the spindle turning in this nut. From this sleeve braces s (shown dotted) are connected with the metal shell or socket H. Two locknuts, 5 and 6, secure the parts when screwed down.

The block G may be changed for one of a different length, or a piece may be eut oft' from it to adjust the leg to the proper length.....

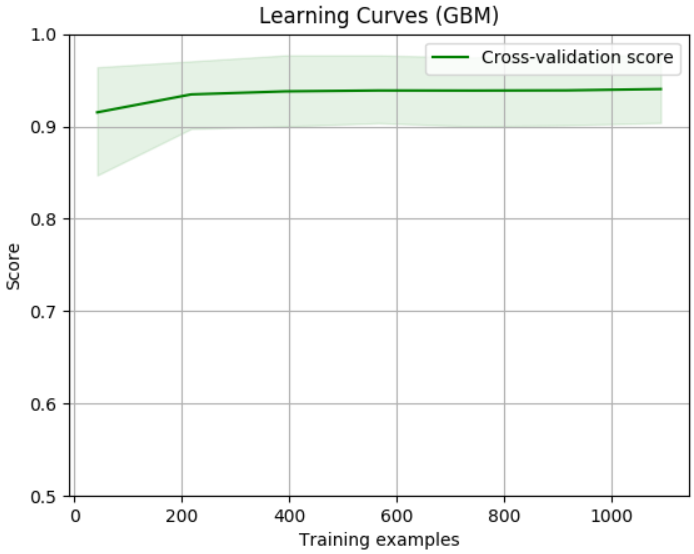
Note: The figure presents a patent document example. We focus the machine learning algorithm's attention to the keywords (blue) and the surrounding context (red). In this case spread $= 3$ and the trait of interest is "comfort". We correct spelling errors using a preprocessing procedure.

Appendix Figure B.3: Learning Curve Balanced Accuracy Score

Panel A: Comfort



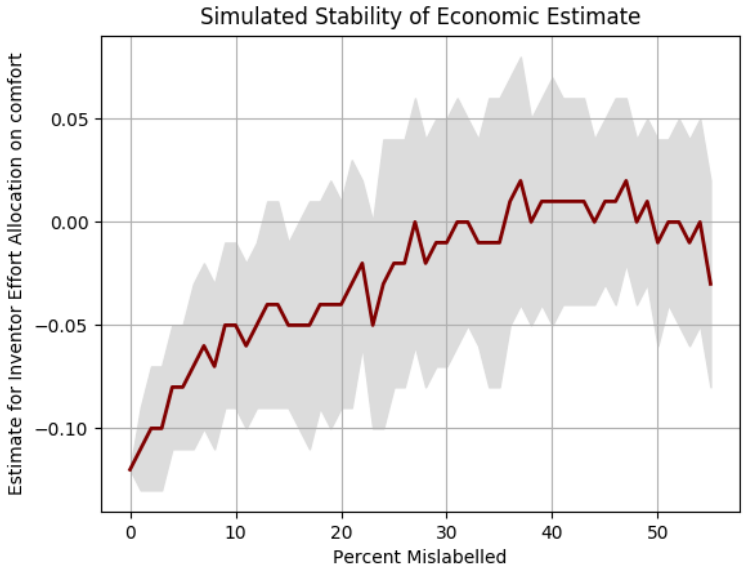
Panel B: Simplicity



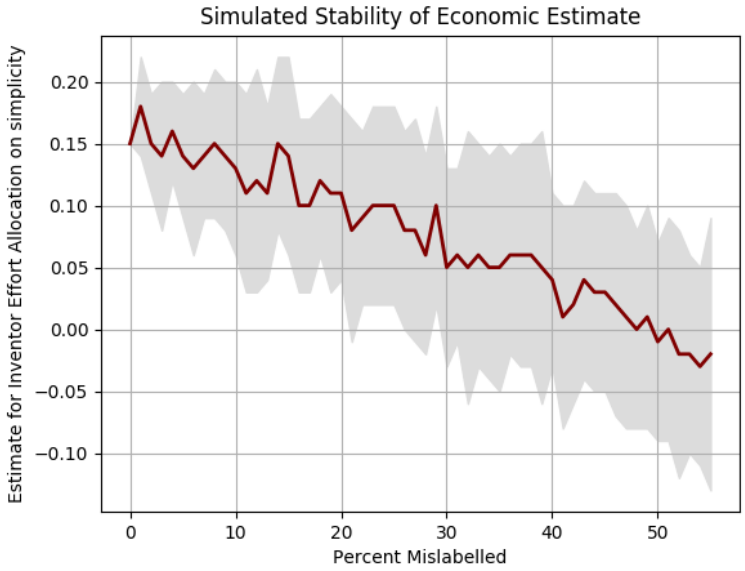
Note: The figure presents the "learning curves" for our preferred modified approach using a GBM algorithm when predicting the presence of our traits in patent documents. Panel A shows the learning curve for "comfort" and panel B shows the learning curve for "simplicity." The solid green line in each panel traces the mean of the balanced accuracy score across 400 iterations of a bootstrap cross-validation procedure at ascending sample sizes. Each bootstrap iteration randomly selects a training set of the "training examples" size to train the model and the model's accuracy is then tested on the remaining un-sampled data. The shaded green area extends from the 10th to the 90th percentiles of the distribution of results. Balanced accuracy is reported in decimals (0.9 = 90% correctly predicted).

Appendix Figure B.4: Estimate Stability To Reductions in the Accuracy Score

Panel A: Comfort



Panel B: Simplicity



Note: The figure shows the simulated stability of our economic estimates as we reduce the accuracy of our preferred algorithm. Panel A shows the simulated stability for our "comfort" variable and panel B shows the simulated stability of our "simplicity" variable. Using all the data generated by our preferred modified approach, we inject noise at random by altering the coding of a given percentage of the observations for our estimates of interest. We then re-estimate β_1 from equation (6) using a synthetic control procedure. We do this 40 times, sampling with replacement, for each percent mislabeled. The red line in each panel traces the mean of the estimates of β_1 from equation (6) at each percent mislabeled. The shaded grey area shows one standard deviation above and below the mean.

Appendix Table B.1: Balanced Accuracy Scores Across Training and Test Set Contexts

Panel A: Comfort

		Test Data			
		CWP	CWC	WWP	WWC
Training Data	CWP	93.9	84.4	91.8	78.4
	CWC	93.1	91.6	91.8	75.8
	WWP	93.6	84.4	92.7	78.4
	WWC	91.3	84.0	90.0	91.6

Panel B: Simplicity

		Test Data			
		CWP	CWC	WWP	WWC
Training Data	CWP	97.0	86.0	94.8	89.1
	CWC	96.7	94.8	93.8	93.0
	WWP	95.8	86.0	94.8	89.1
	WWC	98.4	92.7	95.4	93.5

Note: The table shows the ability of our preferred modified approach applied to a GBM model to predict our traits within and outside the context of the model's training data. We present balanced accuracy scores across wars and broad patent technological classes. Panel A shows the balanced accuracy scores when predicting "comfort" and panel B shows the balanced accuracy scores when predicting "simplicity". Balanced accuracy is reported in percentage terms (78.4 = 78.4% correctly predicted). The main diagonal presents the balanced accuracy means that are obtained through repeated 10-fold cross-validation, using the same context for training and testing. Off-diagonal entries present the model's once-calculated balanced accuracy on the given left-out test set of a different context. The (i, j) entry corresponds to using the data from row header context i in GBM training to predict the left-out data from column header context j. CWP uses Civil War prosthesis patents, CWC uses Civil War control patents, WWP uses WWI prosthesis patents, and WWC uses the WWI control patents. To ensure that differences between balanced accuracy scores across contexts are not driven by differences in sample size, we constrain the size of the training set to be equal in all cases.

Appendix Table B.2: Performance of Algorithm Across Attributes Using All Patents

Characteristic	Sensitivity	Specificity	Balanced Accuracy
adjustability	94.8	91.0	92.9
	(3.2)	(3.3)	
comfort	91.8	96.3	94.0
	(5.6)	(2.3)	
simplicity	92.7	94.3	93.5
	(5.3)	(2.6)	
materials	81.6	92.4	87.0
	(15.7)	(2.6)	
appearance	91.8	96.1	93.9
	(7.1)	(1.7)	
cost	94.7	98.9	96.8
	(4.3)	(1.1)	

Note: The table shows the performance of our modified approach applied to a GBM algorithm across our traits of interest. We present the sensitivity (true-positive rate), specificity (true-negative rate), and the balanced accuracy (simple average of mean sensitivity and specificity). Sensitivity and specificity means are taken across repeated 10-fold cross-validation and the corresponding standard errors are reported below each point estimate in parenthesis. All evaluation metrics and standard errors are reported in percentage terms (94.8 = 94.8% correctly predicted). All manually coded observations are used in the cross-validation procedure.