Disregarding the Shoulders of Giants: Inferences from Innovation Research

David M. Reeb
National University of Singapore

Wanli Zhao
Bocconi University

ABSTRACT

Studies proposing new determinants of corporate innovation include previously identified factors in an *ad hoc* manner. We find that only a sparse set of recently proposed innovation determinants provide material, independent information about patents and citations. We document that inferences in recent empirical studies often change when we include previously discovered innovation determinants. Commonly used econometric methods, including fixed effects and plausible shocks, do not always mitigate the need to condition on previously identified innovation determinants. Rather than randomly selecting a subset of control variables from prior studies, our analysis offers researchers a framework to consider previously proposed variables. (*JEL* G30, O30, G32, O34)

We benefited from discussions and comments from Sumit Agarwal, Tara Bhandari, Ekkehart Boehmer, Simba Chang, Ben Charoenwong, Tony Cookson, Jin-Chuan Duan, Daniel Ferreira, Huasheng Gao, Vladimir Ivanov, Chao Jiang, Kathleen Kahle, Charles Lee, Ram Mudambi, Mattias Nilsson, George Papadakis, Ivan Png, Christopher Polk, Peter Pope, Sungjune Pyun, Wenlan Qian, Srini Sankaraguruswamy, Johan Sulaeman, Parth Venkat, Matthew Wynter, Bernard Yeung, and Mengxin Zhao and seminar participants at 2018 Singapore Scholars Symposium, FMA 2018 conference, Bocconi University, Chinese University Hong Kong at Shenzhen, IESEG, Nanyang Technological University, National University of Singapore, National Cheng Chi University, Renmin University of China, the U.S. Securities and Exchange Commission, Southwestern University of Finance and Economics, University of Guelph, University of Oklahoma, and WHU-Otto Beisheim School of Management. We particularly want to thank Isil Erel (the editor) and two anonymous referees for their help in significantly improving the paper. Send correspondence to Wanli Zhao, wanli.zhao@unibocconi.it.

We address a commonly encountered problem for empirical researchers, namely, the extensive set of previously identified determinants on the same outcome variable. For instance, in the investment literature, a benchmark that subsumes the predictive power of a covariate provides no advantage to investors. In corporate finance research, we encounter similar situations in which selecting previously identified determinants to include in new studies seems *ad hoc* and random. Regrettably, little guidance exists on how to handle this problem, nor is there consensus among researchers that a problem exists. Using innovation research as the platform, we aim to aggregate existent research and facilitate future studies by exploring how to incorporate previously documented innovation determinants into new empirical research. Innovation research offers a natural laboratory because of the recent interest in this topic across finance, economics, and strategy journals. Also, the public availability of the data and the uniformity of the measures ensure that the empirical results are comparable between studies.

We identify 53 recent articles in the most prominent finance and economics journals that propose new patent determinants. These studies rely on various economic arguments and models to conjecture that their proposed variable facilities corporate innovation.[1] Yet, we observe that these studies rarely condition analysis on a similar set of control variables, even though they all use the same dependent variable(s). Thus, aggregating this literature is a necessary step; we do so by identifying which previously discovered innovation determinants provide material, independent information about patents and citations. Otherwise, uncovering new explanatory variables or

[1]A simple analysis of the 409 numerical variables in Compustat, shows that 70% of the variables are correlated with patent activity. This percentage illustrates the difficulty in interpreting or assessing this body of research. A conventional machine learning approach would use all 409 of these variables. In contrast, we focus on the previously argued economic determinants of corporate innovation, seeking to find which provides statistical evidence distinguishable from the others to identify a sparse set of control variables. Without a central theory of corporate innovation, the outcome of variable selection via the conventional blind application would yield little inferences on the findings.

features of corporate innovation becomes challenging without a clear understanding of the structure of the existing body of research.

We use machine learning and other regularization techniques to evaluate and compare these previously proposed economic determinants of corporate innovation. Our aim is not to identify which of these determinants "causes" corporate innovation but instead to offer researchers a framework to select control variables for future research. We find that only a small subset of the variables hypothesized in prior innovation research survives the "horse race" to provide material and independent explanatory power. Selecting a subset of significant variables from an extensive set of covariates, whether based on data mining Compustat or from a group of previously hypothesized economic determinants of corporate innovation, does not provide insights into which of these variables have a causal interpretation. Instead, we aggregate the existing literature on corporate innovation using a data-driven approach to evaluate the explanatory power of these previously hypothesized determinants. Without conditioning on previously identified innovation covariates, interpreting new corporate innovation studies becomes increasingly tricky. Our results resonate with the recent observations in Harvey (2017), in which he notes the escalating number of false-positive factors because of *p*-hacking.

After identifying the subset of innovation covariates proposed in prior research, which we call high explanatory power (HEP) covariates, we investigate how excluding these HEP variables affects new studies' inferences. We find these HEP factors, if included, often undermine the findings of subsequent studies. Thus, even though the decision to consider previously documented innovation determinants depends on the research question, we document that the inference or the interpretation of the empirical findings could change drastically and needs careful attention when conducting new research.

Finally, if most empirical research aims to identify marginal causal effects, perhaps this omitted variable problem is not a significant concern in studies that use an exogenous change in the independent variable. A critical issue in these assessments centers on the change's exogeneity, mainly whether the exclusion condition is satisfied. Unfortunately, the exclusion condition is untestable and cannot be verified empirically, and including a set of control variables does not solve the problem. Yet, a common theme in much of the empirical literature is that using a proposed exogenous shock alleviates the need for including previously identified factors. We explore how standard techniques, such as fixed effects and plausible shocks, potentially mitigate the necessity of including previously identified economic determinants of innovation. We find that neither approach necessarily mitigates the need to have prior economic determinants in the analysis. At a minimum, we suggest reviewing the exclusion condition by investigating whether previously discovered innovation covariates are related to the independent variable's purposed exogenous change.

Our tests take advantage of machine learning techniques as they provide a systematic method for feature selection. Rather than using machine learning to data-mine a host of firm characteristics, we use these techniques to evaluate and compare previously proposed economic determinants of corporate innovation for inclusion as conditioning variables.[2] Machine learning techniques typically rely on semiparametric algorithms, explicitly building on out-of-sample verification to compare different models from the in-sample analysis (Mullainathan and Spiess 2017). Similarities in outcomes from different regularization approaches (elastic net, group Lasso, and stepwise regressions) and different time periods suggest the results stem from the underlying data

[2] The crucial distinction between our application of machine learning versus the conventional application is in the initial variable selection. We focus on theoretically motivated variables in prior research in which the author(s) found empirical support for their predictions. In contrast, a standard machine learning approach centers on data mining a large set of observable characteristics to discover correlation patterns.

generating process rather than the weaknesses of any particular approach. Still, selecting a subset of significant variables from a large set of covariates, even if based on previously proposed economic determinants of innovation, does not answer whether these variables determine corporate innovation. Instead of randomly selecting control variables from prior research, our analysis offers researchers a framework to choose conditioning variables from previously proposed determinants. A weakness of this approach is that it cannot provide insight into which of the previously proposed determinants of innovation have causal interpretations and which do not.

We assemble 35 documented corporate innovation determinants, with our main tests focusing on firms with patent applications (see Koh et al. 2016; Lerner and Seru 2017).[3] To better ensure our results stem from the true data generating process, we implement our machine learning tests by assessing the entire sample's data as the primary approach and complement it with tests using short (2-year) rolling windows, relying on out-of-sample analysis for cross-validation. Performing the investigation across the whole period, in different subperiods, or with varying window choices emphasizes these covariates' relevance as key conditioning variables (which explain over 90% of the all-variable explanatory power specification).

Our initial analysis focuses on the quantity of innovation, investigating which previously identified covariates provide material, independent explanatory power for patent applications.[4] The number of corporate patents is a standard measure of innovation in financial market research, especially in studies about managerial incentives to invest in innovation. Among patenting firms,

[3] The 53 prior studies contain more than 35 factors for innovation. In Appendix A, we show the frequency of each variable used in those studies, and we provide explanations on how we arrive at 35 variables. In recent years, the *Journal of Financial Economics* includes 19 such articles, 10 for *Management Science*, 7 for *The Review of Financial Studies*, 6 for *The Journal of Finance*, 5 for the *Academy of Management Journal*, and 2 for the *American Economic Review*. At the same time, the *Accounting Review*, *Strategic Management Journal*, *Quarterly Journal of Economics*, and *Econometrica* each contain one such article.
[4] Our goal centers on aggregating the existing set of innovation covariates rather than increasing the predictive power of previously proposed determinants by including nonlinear terms. Studies that identify nonlinear effects (e.g., Im and Shon 2019) could be affected by such concerns.

we find 6 of the 34 previously proposed innovation determinants provide material, independent explanatory power for patents (excluding R&D stock in this test). These six variables are stock liquidity (Bernstein 2015; Fang, Tian, and Tice 2014), firm size, CEO reputation or centrality (Faleye, Kovacs, and Venkateswaran 2014), analyst following (He and Tian 2013), and industry citation and patent intensity (Hall, Jaffe, and Trajtenberg 2001). Interestingly, we find that 30 of the previously hypothesized innovation determinants do not survive the selection process.

We separately repeat the analysis for innovation productivity, which includes R&D stock, and use 35 previously documented innovation determinants. Unsurprisingly, R&D stock is an essential variable in explaining patent activity and substantively influences the analysis. We find 6 of the 35 previously proposed variables explain patents, including CEO centrality (Faleye, Kovacs, and Venkateswaran 2014), stock liquidity (Bernstein 2015; Fang, Tian, and Tice 2014), R&D stock (Balsmeier, Fleming, and Manso 2017), firm size, analyst following (He and Tian 2013), and industry citation intensity (Hall, Jaffe, and Trajtenberg 2001). These results provide a potential standard set of control variables in studies that focus on innovation productivity, such as studies that argue governance improves R&D efficiency.

We undertake a similar analysis for patent citations to gauge innovation quality, starting with the same 35 proposed innovation determinants in prior research. Again, we find that stock liquidity, industry citation intensity, CEO centrality, and firm size explain citations, which, taken together, explain over 90% of the explanatory power of the all-variable specification. Moreover, we obtain similar results across different regularization procedures, suggesting these findings stem from the data generating process rather than some particular method's peculiarity.[5]

[5] One noteworthy point is that the variable selection process does not distinguish between causal and associational inferences. A variable representing a clear causal effect may be more important than a variable with greater explanatory power. We will address this issue in Section 1.3.

After identifying the HEP variables, we explore how inferences from empirical studies may change if we include these key covariates, specifically we look at recent studies that propose new determinants of corporate innovation. Recent research proposes several novel innovation determinants, including whether the CEO has a pilot license, antitakeover devices, and state-level corporate taxes. We repeat the analysis in these studies, with and without the HEP from prior research. Using the exact model specification of Sunder, Sunder, and Zhang (2017), we confirm the positive relationship between pilot CEOs and patent activity.[6] After including the HEP covariates from prior innovation literature, we find that the pilot CEOs' coefficient estimate is insignificantly different from zero. Using a research setup similar to that of Chemmanur and Tian (2018), we confirm the positive relations they document between innovation and antitakeover devices. After including the HEP innovation determinants from prior research, we find the coefficient estimate on antitakeover devices is no longer significantly different from zero. The inclusion of previously identified industry-level innovation covariates is the reason the antitakeover variable loses significance.

Yet, it is essential to note that in other cases, adding the six surviving innovation covariates strengthens the results about newly proposed innovation covariates (e.g., Mukherjee, Singh, and Zaldokas 2017). One potential benefit of including the HEP variables from prior research is an improvement of model fit. Our analysis reveals that the inference becomes statistically more robust by including these previously documented variables in the empirical specifications.

Do we suggest that these studies not including the HEP variables are "wrong"? The answer is no. On the one hand, without the proper set of control variables, studies can discover false-positive findings (Harvey 2017). On the other hand, the answer to the question depends on the

[6] We thank Stephen B. McKeon for providing us with the pilot CEO data.

research question and the purpose of the study. For instance, suppose that CEO characteristics matter for corporate innovation, and the underlying mechanism is via risk tolerance. Since risk tolerance is hard to measure but presumably correlated with pilot CEOs, the interpretation of the pilot CEO study is that pilot CEOs have higher risk tolerance. However, pilot CEOs also may be overconfident, which is related to innovation. If some other previously proposed variable captures CEO overconfidence better than their pilot status, that variable will win the horse race. This finding does not mean that risk tolerance is not related to innovation but simply due to the poor correlation between pilot CEO and risk tolerance (conditioning on overconfidence). Our evidence suggests that a proposed covariate's interpretation or inference depends on the research question's purpose. Incorporating the HEP variables from prior research into the analysis provides information about the effect's underlying mechanism.

In our last task, we explore whether commonly used alternative methods to address omitted variable concerns can mitigate the necessity of including the HEP variables from prior research. Specifically, as a parsimonious treatment to address the omitted variable problem, studies often include industry or firm fixed effects in the analysis to account for differences in patenting choices.[7] We explore whether these HEP determinants provide additional explanatory power after including industry or firm fixed effects. Our research indicates that the seven surviving covariates of innovation typically keep substantial explanatory power after industry, industry-year pairwise, or firm fixed effects. However, firm size is the only remaining significant variable after adding both firm and industry-year pairwise fixed effects.[8] In short, these tests suggest that fixed effects

[7] Many R&D firms choose not to file patents. A common argument for using industry fixed effects is to mitigate the potential bias of R&D firms that do not seek patent protection.

[8] In many studies, controlling for fixed effects may not be practical or meaningful, especially when the independent variable is sticky and lacks time-series variations (e.g., Bertoni and Tykvova 2015). Consequently, including both firm and industry-year pairwise fixed effects is rare in practice.

rarely mitigate the need for including previously documented innovation determinants into the analysis.

Empirical research often focuses on discovering marginal causal effects, relying on exogenous shocks to identify these causal relationships. This method arguably mitigates omitted variable concerns more robustly than relying on fixed effects. However, this approach depends on satisfying the exclusion restriction, which requires the shock to be uncorrelated with other innovation covariates, even when the shock is genuinely exogenous. Thus, assessing a typical instrumental variable or natural experiment hinges on understanding whether the shock only influences innovation through the proposed dependent variable of interest or whether it also affects previously identified innovation covariates. Unfortunately, little guidance exists on which variables to use to evaluate the exclusion restriction in corporate innovation studies.

To investigate the potential use of these variables in assessing the exclusions restriction in patent-based studies, we replicate a study on institutional ownership and corporate innovation. The identification strategy evaluated relies on the inclusion of firms in the S&P 500. Our evidence shows that the S&P 500 shock also influences one of the previously identified covariates of innovation: stock liquidity. [9] In this context, the evidence that institutional ownership increases corporate innovation could arise from a direct effect theorized by Aghion, Van Reenen, and Zingales (2013) or from institutional ownership lowering the cost of capital for innovation projects.

This exclusion restriction example does not imply that including the HEP variables offer a test of the exclusion condition. Nor does our analysis provide a feasible method to sort out which variables affect innovation output. Importantly, it demonstrates that the exclusion restriction's

[9] We confirm the positive relationship between institutional ownership and corporate innovation using a specification from Aghion, Van Reenen, and Zingales (2013). Importantly, Aghion, Van Reenen, and Zingales (2013) provide formal theory explaining why the direct effect might exist. Conceptually, institutional ownership could have direct and indirect impacts on corporate innovation.

untestable nature indicates that relying on exogenous shocks for identification does not obviate the need to include previously documented innovation determinants. Moreover, the HEP variables provide a better baseline for new corporate innovation studies than *ad hoc* conditioning variable selection.

This study makes several significant contributions to the growing literature on the determinants of corporate innovation. We posit that applying a data-driven approach for conditioning new analysis on previously hypothesized innovation determinants adds rigor to the process, especially when no recognized theory guides these control variable choices. Our perusal of corporate innovation studies indicates that this literature uses various potential conditional variables in an *ad hoc* manner. It is unclear from prior research which previously identified factors researchers should include as control variables in new studies of corporate innovation.

Our analysis using machine learning techniques reveals that a small set of previously identified covariates for innovation provide independent explanatory power on the quantity, productivity, and quality of corporate innovation. Remarkably, relying on industry or firm fixed effects does not invalidate the need to control patent activity's previously documented determinants in assessing any newly proposed innovation determinants. In this context, our analysis provides empirical support to the argument that machine learning methods and natural experiments provide complementary approaches in financial economics research.

Our next contribution is to show that without considering these HEP variables, the studies' inference on innovation can be drastically different or even become questionable. In addition, our results reveal the relative statistical explanatory power of the covariates in prior literature, calling for further research on some intriguing findings (e.g., stock liquidity). Our study offers a framework or starting point for researchers to rethink their newly proposed innovation determinant's inference by including previously identified determinants. Instead of randomly choosing what previously identified variables to include as conditioning variables, we suggest a small set of these hypothesized variables from previous economic studies that provide independent explanatory power.

Innovation studies often rely on exogenous shocks to provide causal evidence. However, this approach depends on the applicability of the exclusion restriction. Identifying the potential covariates to analyze in testing the exclusion restriction is challenging because guidance from the innovation literature is limited. Against this backdrop, our analysis provides some preliminary guidance on the covariates to analyze in testing the exclusion restriction. More specifically, our research suggests using the seven HEP variables from prior research to evaluate whether a shock influences previously identified covariates.

1. Data, Sample, and Variables

1.1 Data sources and sample

To capture the 35 previously identified determinants of innovation, we obtain data from multiple sources. Our main sample is a cross-section of different databases (see the appendix for details). More specifically, we use ExecuComp to capture compensation information about managers. We complement it with BoardEx data with information about other CEO characteristics, such as age, gender, and centrality. We rely on the Compustat and The Center for Research in Security Prices (CRSP) databases to capture firm characteristics. We acquire corporate governance practices of firms from the Investor Responsibility Research Center (IRRC) Risk Metrics. We obtain family firm status from Ron Anderson’s website,[10] state marginal tax from the Department

[10] http://www.ronandersonprofessionalpage.net/data-sets.html

of Labor, and board-related information from BoardEx. We collect institutional ownership information from Thomson Reuters and analyst following information from the Institutional Brokers' Estimate System (I/B/E/S). The industry characteristics are based on Compustat and CRSP information. Finally, we obtain patent and citation information from the U.S. Patent and Trademark Office (USPTO) (Hall 1990). We drop the financial (SIC 6000–6799) and utility (SIC 4900–4949) industries. Our main sample spans from 2001 to 2010 with 2,716 firm-year observations of 410 unique firms with patents. In parallel to the main sample, which is the most restrictive because of a lack of data availability, we also present results using a larger sample from 1992 to 2010 of 5,955 observations of 832 unique firms.[11]

1.2 Variable definitions

1.2.1 Dependent variables.

We use two commonly used metrics for innovation output, patents and citations, as the dependent variable. The patents are based on the patent applications, and we focus on the application year rather than the grant year as the application year is closer to the actual time of innovation (Griliches, Pakes, and Hall 1987). Specifically, we use the log of patents and log of (1 + citations), and our base tests only include firms with patents. In later tests, we will include nonpatenting R&D firms. Doing so will allow us to incorporate firms with zero patents, to provide insights into the determinants of the patenting choice.

1.2.2 Right-hand-side variables.

[11] The sample size in our main sample is largely restrained by the lack of availability of CEO compensation and family firm data. In later sections, we will loosen this restriction to evaluate the candidate covariates across time and firms. Our first analysis extends the sample to the 1990s to incorporate 25,985 observations with 2,217 unique firms covering 24 variables. Our second analysis extends the sample to include additional firms; doing so gives us 58,671 observations with 7,502 unique firms covering 19 variables.

We include 35 potential determinants of corporate innovation from prior research that we classify into four categories. The first group includes six managerial characteristics: CEO gender, CEO total compensation, CEO delta, CEO vega, CEO confidence, and CEO centrality. The second group contains firm features including firm size, (log) firm age, R&D stock, Tobin's q, stock liquidity, the firm's headquarters' distance to nearest USPTO office, tangibility, a dummy variable indicating if the firm is in manufacturing, return on assets (ROA), sales growth, organizational capital, capital structure, and state marginal tax rate. In the third category, we include corporate governance variables, including six antitakeover provisions (staggered board, poison pill, golden parachutes, limits to shareholder bylaw amendments, and supermajority requirements for mergers and charter amendments), board size, board independence, institutional ownership, a block holder dummy, analyst following, and family firm designation. Finally, we include industry characteristics in the fourth category, namely, the industry patent intensity, industry citation intensity, average industry R&D, industry competition, and industry size. We define the variables in detail in Appendix C. Table IA1 in the Internet Appendix provides summary statistics for the 35 variables.

1.3 Causality versus the explanatory power of variables

Our study aims to identify the subset of covariates identified in prior academic research that provide materials and independent information about innovation performance. We realize that explanatory power is not the only factor that should be considered in determining what variables should be included in a regression model, as causality is a conceptually more appealing property. Yet, a concern is that many of the previously proposed casual factors do not capture new economic insights beyond those previously identified. Still, one could argue that a variable representing an apparent causal effect with less explanatory power is more important than a variable with greater explanatory power but no clear causality.

Arguably among the many proposed determinants of patents and citations, some of these studies have done a better job of establishing a causal effect of the determinants on patents and citations. However, when we look back at the prior studies, we find it difficult to ascertain which ones provide robust evidence of causality, even though they all seem to argue for some causal relationship.[12] Given that it is difficult to gauge the relative strength of the causal inference of the covariates, we do not distinguish between the HEP variables in terms of their causal relationship on innovation.

2. Identifying Key Variables: A Machine Learning Approach

2.1 Machine learning method in variable selection: Adaptive Lasso

He and Tian (2018) indicate that in recent years corporate innovation research has drastically increased. Not surprisingly, multiple managerial, firm, and industrial characteristics have been proposed as influencing a firm's innovation performance. For instance, Aghion, Van Reenen, and Zingales (2013) document the effect of institutional investors, and Galasso and Simcoe (2011) posit that CEO confidence is positively associated with innovation. So far, more than 30 factors have been studied and found to be significantly related to innovation. We apply a "horse race" approach to the multiple factors that have been identified as important determinants of corporate innovation.

The traditional variable selection approach relies on stepwise regressions as it proves to be computationally tractable relative to all subset regressions. Ideally, one might select the best fit model via specific statistical criteria of model fitness (e.g., Akaike information criteria or Bayesian

[12] Arguably, the ones that rely on exogenous shocks provide more convincing evidence. However, whether these studies satisfy the exclusion conditions is not clear, given that many of these studies do not examine this issue.

information criteria), among all the possible combinations of the variables (in our case, 2^{35} = 34,359,738,368). Stepwise regression overcomes best subset model selection infeasibility by severely lowering the complexity or number of combinations to assess (630 potential models in our case). However, as convenient as it is, stepwise regression has several shortcomings. For instance, the process only focuses on a subset of the potential models among the possible combinations of the 35 factors because the outcome is contingent upon the sequence of the variables in the regression. Specifically, when a stepwise process either drops or adds a variable one at a time, the sequence of the variables becomes important. Unfortunately, there is no clear treatment on this issue. Last, but not least, the stepwise procedure lacks validation of the outcome, that is, there is no out-of-sample verification in the process. A benefit of these traditional methods, all subset or stepwise, is that they directly penalize the model for additional coefficients.

Because of these shortcomings in traditional parametric approaches, we use a machine learning method, specifically the adaptive Lasso procedure, as our primary approach. Machine learning methods divide the sample into training and validation subsamples. Adaptive Lasso (the least absolute shrinkage and selection operator) provides a strong and robust inference because it explicitly considers the "predictive" power of the selection outcome via a cross-validation process. Furthermore, adaptive Lasso gives a differential weight for penalizing different coefficients instead of applying a common penalty factor to all coefficients. We apply adaptive Lasso to our data using tenfold cross-validation and choosing the two tuning parameters (lambda and gamma) to minimize the mean square error in the out-of-sample testing (Hui, Warton, and Foster 2015). We fit the model on 9 subsets using these 10 random subsets and then test the model on the excluded or validation set. We repeat this approach for each of the excluded sets and select the model with the best out-of-sample performance across all 10 subsets. In Appendix B, we explain the adaptive Lasso method (typically labeled as an L_1 penalty) and group Lasso and elastic net methods in detail.

For our main tests, we apply adaptive Lasso with the main sample from 2001 to 2010. In complementary tests, we also present the results using a rolling window approach.[13] After selecting variables via adaptive Lasso, we then rank all the variables by their explanatory power. Finally, we show the incremental explanatory power loss when we drop the variables in reverse order from the least important to the most important variable from the full specification. In addition to adaptive Lasso, we also present results using alternative methods, namely, group Lasso, elastic Lasso, and stepwise regression. The machine learning methods, adaptive Lasso, group Lasso, and elastic net regression, split the data into 10 random subsets for testing. Still, the stepwise regressions rely on in-sample information criteria for variable selection.

2.2 Key variables identified

2.2.1 Patents as the innovation outcome measure.

Table 1 reports the variable selection results. Panel A shows results when we use patents as the output metric for innovation, while panel B does so for citations. The main sample spans from 2001 to 2010, and we also show an expanded sample of 1990 to 2010. The more extended sample does not include the variables where data are not available. We present the incremental R^2 loss for each variable, which is computed by comparing ordinary least squares (OLS) regressions after dropping that variable in the order suggested by the adaptive Lasso variable selection. In columns

[13] The rolling window approach is meant to check whether the variable selection results vary between different periods. Specifically, the approach yields nine 2-year rolling windows based on the sample of 10-year period. We show the frequency of each variable being selected via adaptive Lasso among the nine rolling windows. In untabulated results, we also try using a 1-year window, and we find similar outcomes. Industry factors become weaker because they are the average value of firms so that a single-year window dramatically decreases the cross-sectional variation. Our use of a rolling window approach means the optimal penalty factor changes from sample to sample. To facilitate the replicability of our analysis, we base our tests on the average optimal gamma of four across our rolling windows. We find that using sample-specific gammas yields outcomes similar to the average optimal gamma across our rolling windows.

1 and 2 of both panels, we focus on innovation output quantity, and, consequently, we do not include R&D stock, the innovation input factor, as one of the candidate variables. The results in panel A, column 1, indicate that the adaptive Lasso procedure yields six variables associated with patents. Specifically, we find that CEO centrality, firm size, stock liquidity, analyst following, industry citation intensity, and industry patent intensity are chosen. More importantly, the adaptive Lasso procedure applies the cross-validation process to yield results that have strong predictive power rather than only model fitness. In addition, we show that the incremental explanatory power loss of the unchosen variables is relatively small. On the other hand, the six variables identified together provide roughly 89% of the explanatory power of the full specification. Figure 1 illustrates these findings and shows the order of variable selection in terms of each variable's relative explanatory power. Each line represents a variable, and the farther right the variable is, the more influential the variable. The vertical line represents the threshold at which the variables are selected or retained. The figure shows six variables (plus two year dummy variables) are selected by the process. Overall, we conclude that most of the variables are not significantly associated with patents.[14]

In column 2, we repeat the same process with the larger sample, which spans 1992 to 2010. We drop four variables due to data availability restrictions (CEO centrality, board size, board independence, and family firm). We find the same three overlapping variables as in the column 1 results, that is, size, analyst following, and stock liquidity, are key variables for patent output. Overall, columns 1 and 2 indicate that six variables survive the horse race, namely, firm size, stock liquidity, CEO centrality, analyst following, industry patent intensity, and industry citation intensity.

---

[14] Figure IA1 shows the fraction of deviance (similar to R^2) explained by the six variables. The figure shows that roughly 55% of model deviance is explained by the six variables, while the full specification explains about 58%.

Column 3 shows the results when we include R&D stock. Compared to column 1 results, we find that R&D stock is selected, while industry patent intensity is dropped, suggesting that six variables are important factors for innovation productivity. In column 4, using the larger sample, we find that the same five variables (size, stock liquidity, R&D stock, analyst following, industry citation intensity) are selected.[15]

2.2.2 Citations as the innovation quality metric.

Table 1, panel B, shows the results when we focus on citations as another commonly used metric for innovation outcome. Again, columns 1 and 2 present results without R&D stock, and in columns 3 and 4, R&D stock is included. In column 1, the main sample result shows that only four variables are identified as key variables: CEO centrality, stock liquidity, size, and industry citation intensity. Together, the four key variables provide 88% of the explanatory power of the full specification. Column 2 shows that when we use the larger sample, the same results surface (except for CEO centrality where data are not available). In columns 3 and 4, we include R&D stock, and we repeat the process. We find that the variable selection process yields the same set of variables as do columns 1 and 2.

Interestingly, R&D stock is not chosen as an important factor for patent citations. In short, adaptive Lasso spots three variables that are significantly related to citations, namely, firm size, stock liquidity, and industry citation intensity (plus CEO centrality for the shorter sample period). Taken together, the findings so far show that only a handful of factors are related to patents and citations. At the same time, each of them provides significant incremental explanatory power.

2.2.3 A rolling-window approach.

[15] Surprisingly, we find that stock liquidity provides the strongest explanatory power, and it is indeed intriguing to observe that stock liquidity is salient in such a manner via the variable selection process. Going back to the original study, we do not find any theoretical explanation on the effect of stock liquidity. We believe it offers an interesting future research opportunity.

In addition to the full-sample variable selection, we also try a rolling-window approach. The purpose is to observe whether the variable selection outcome varies between different time periods. The stability of the variable selection outcome is an important issue to ascertain because it facilitates the robustness of the selection outcome not being due to some data randomness. Table IA2 in the Internet Appendix presents the results. Panel A shows the results for patents, and panel B does so for citations based on the main sample and the expanded sample. Specifically, in column 1, we repeat the adaptive Lasso procedure for each of the nine windows for the main sample of 2001 to 2010 (for a sample of 10 years, a 2-year rolling window yields 9 subsamples). We present the frequency of each variable being selected among the nine rolling windows. For instance, we find that adaptive Lasso selects CEO centrality in 8 of the 9 rolling windows. In sum, column 1 shows that five variables, CEO centrality, size, stock liquidity, industry citation, and patent intensity, are selected in most rolling windows. We find other variables are seldomly selected except for analyst following. These results are consistent with the full-sample outcome in Table 1. Column 2 shows the results for the expanded sample of 19 years. We find that size and stock liquidity survive almost all of the 18 rolling windows, while industry citation, patent intensity, and analyst following are selected in 8 to 12 of the rolling windows.

Columns 3 and 4 show results when we include R&D stock on the right-hand side. We find a similar outcome: besides R&D stock being selected, CEO centrality, size, stock liquidity, and industry citation intensity exhibit consistent candidate variables across the rolling windows. Analyst following is selected in more than half of the rolling windows, and industry patent intensity is selected in the expanded sample.

Panel B presents the results for citations. We find that firm size, stock liquidity, and industry citation intensity are consistently selected across the two samples and two different specifications. R&D stock is again not selected for citation, same as our full-sample results in Table 1, panel B.[16] Overall, we find the same set of variables by the rolling-window approach as the full-sample results.

2.2.4 Robustness evidence with alternative variable selection methods.

We present variable selection results using multiple alternative methods in Table 2. First, we show the results using three alternative variable selection methods: group Lasso, elastic net, and stepwise analyses. In column 1, we use the group Lasso methodology, and we allow individual variables to be selected within the group. We find the same six key variables identified in the adaptive Lasso analysis, and three additional variables, namely, vega, institutional ownership, and industry size, are selected.[17] In column 2, we use an elastic net procedure to address further whether the results are driven by the fundamentals or by a specific machine learning specification. We find that the results show the same seven covariates providing significant explanatory power for innovation.

While the traditional all subset variable selection method remains unfeasible with 35 potential covariates, the commonly used alternative centers on stepwise regressions. The backward stepwise procedure starts with all the variables and takes an elimination approach; if the variable incurs the least amount of model fitness loss, the variable is dropped. The process is repeated until no more variables can be deleted without a statistically significant loss of model fitness. This variable selection approach is often labeled as “greedy” because of concerns with overfitting. To further corroborate the machine learning results, we use a backward stepwise regression and show the results in column 3. The individual R^2 determines the order of the variables as we include each

[16] In all the rolling-window selection results, we find the sign of every variable selected are always consistent for each rolling window.
[17] This result highlights that among a group of highly correlated variables (such as in a group), adaptive Lasso tends to choose one variable and ignore the others.

variable one at a time, separately. For instance, the last variable is CEO gender, as it has the lowest R^2 in a regression when it is included as the only explanatory variable. The results show that eight variables are chosen via stepwise regressions.[18] In sum, we find that across the three methods, seven variables, that is, CEO centrality, firm size, R&D stock, stock liquidity, analyst following, industry patent intensity, and citation intensity matter for patents, are identified as the key variables. This outcome is largely similar to what we obtain via adaptive Lasso in Table 1.

Turning to columns 4–6 in which we assess citations, we observe that across the three methodologies, six variables are identified as the key variables, namely, CEO centrality, firm size, R&D stock, stock liquidity, analyst following, and industry citation intensity. No other CEO characteristics are chosen, and none of the corporate governance factors are identified. In conclusion, we find that identifying key variables is robust across different methodological approaches, and they yield similar sets of key variables.

2.2.5 Correlation between variables.

A potential concern in using machine learning methods centers on the correlation among the 35 proposed variables. Our main approach for assessing this issue is the group Lasso analysis. Yet, explicitly investigating the correlation among the various variables provides another layer of robustness. Table IA4 provides the correlation matrix. Casual observation suggests that most of the correlations are rather low. More specifically, the average absolute value of correlation is 0.11 and the median is 0.07. The bottom (upper) quartile is 0.03 (0.13) and the 5th (95th) percentile is 0.01 (0.34). To further assess the consistency of our main results, we repeat the analysis after excluding the variables with the highest correlations. For instance, institutional ownership and

[18] Table IA3 in the Internet Appendix provides the results of the stepwise regressions with rolling windows. The results show that four variables are identified as a key variable. The differences between full-sample and rolling-window results imply the instability of stepwise regressions compared to machine learning methodologies.

analyst following correlate at 0.92. When we take out institutional ownership, the results remain the same. When we take out analyst following, the results also remain the same. We repeat this exercise with variables that exhibit correlations higher than 0.5 and we obtain the same findings. Coupled with the results in the group Lasso analysis, we conclude that the correlations between the 35 variables do not appear to provide spurious findings that would invalidate their use as control variables.

2.2.6 Innovation regimes.

Our selection procedure is applied to a sample that is restricted by data availability. How does the sampling process influence the outcome? Do the same variables show up across different innovation regimes? This question is important because it sheds light on the usefulness of the variables for future research. In Table 3, we use two different ways to check on this issue.

First, in panel A, we show the results using adaptive Lasso but with different samples. In panel A, column 1, we drop the CEO characteristics to loosen the sample restriction from the ExecuComp database. The sample enlarges to 25,985 observations. Again, we find that except for industry size, the same set of key variables are selected by the adaptive Lasso procedure, namely, firm size, R&D stock, analyst following, stock liquidity, and industry patent intensity. In column 2, we further drop the antitakeover provisions from the IRRC database, thus expanding the sample to 58,671 observations. Once again, we find the adaptive Lasso procedure identifies the same five key variables. In columns 3 and 4, we show that key variables for the citation test are also robust to different samples. In panel B, we show the results after focusing on an earlier time, 1990–2000, that does not overlap with the main test sample time frame. We find relatively similar results across the full time period and within various subsets of the data.

2.2.7 Firms without patents.

A common approach to dealing with firms without patents is to include them in the sample, while denoting the patent as zero. Table IA5 in the Internet Appendix provides the results; we repeat the variable selection process, but we include nonpatenting firms and denote their patents and citations as zero. We apply the adaptive Lasso technique and find that eight variables are chosen for the patent test, and three variables are selected for the citation test. These results are almost identical to our main findings, suggesting that the key variables are robust to different sampling treatments.

2.2.8 Using R&D to capture innovation.

Recent corporate innovation research often focuses on the output measures for innovation, that is, patents and citations. Still, R&D expenditure offers another metric as an important input for innovation. In Table IA6 in the Internet Appendix, we apply the variable selection process using R&D as the measure of innovation. Interestingly, we find a much smaller set of factors that are important in explaining R&D spending. In contrast to the seven variables for patents, we find only four factors are chosen in at least two-thirds of the rolling windows, namely, firm size, stock liquidity, tangibility, and industry patent intensity, regardless of the procedure we use. Taken together, these results suggest that R&D spending and patent citations share similar explanatory variables, while patents exhibit a more extensive set of covariates.

2.2.9 Industry factors revisited.

We find industry citation and patent intensity are vital variables. These industry factors, however, may explain too much for the firm-level variations on innovation. In other words, while some firm factors may provide insight into why specific industries have more patents, the industry factors could obscure this contribution. We provide further insight into this issue by dropping the industry factors and repeating our exercise. We find that the variable selection procedure yields the same set of determinants, suggesting that industry factors do not overlap or overtake the explanatory power on the firm-level determinants. In other words, the key variables clearly explain the within-industry variations in innovation performance.

3. A Reexamination of Inferences of Previous Studies

3.1 A horse-race between variables

We first examine the inferences of the factors identified in previous studies via a horse-race approach. In Table 4, we show the results of two specifications. In columns 1 and 2, we test each of the variables' significance, with the key variables always included in the specification. The purpose is to verify whether the nonkey variables survive after we include the key variables. The second specification includes all the variables simultaneously in columns 3 and 4. In columns 1 and 2, we find that several of them show statistical significance when we check each variable at a time. For instance, in column 1, we observe that CEO delta is significant and negative, inconsistent with prior studies. We also find that CEO confidence and Tobin's q exhibit significant but opposite signs as in prior studies. In columns 3 and 4, we show the inclusive approach results, with all variables in the regression simultaneously. We correct the significance of the estimation due to multiple testing. After the multiple testing correction for *p*-values via Bonferroni correction, we find that all the key variables survive the horse race.[19] More importantly, none of the nonkey variables are selected. Specifically, in column 3, we find that CEO centrality, size, stock liquidity, R&D stock, and industry citation intensity have significant explanatory power on patents. Analyst following, on the other hand, does not yield statistical significance. In column 4, where we use

[19] Specifically, for the Bonferroni correction, we multiply the original *p*-value by the number of multiple testing hypotheses/factors, in our case, 35. We mark the corrected statistical significance level accordingly. For instance, in column 4, even though the *t*-statistic of *Industry size* is 2.6 and the uncorrected *p*-value is .0093, the corrected *p*-value is $.0093 \times 35 = .33$, and we mark it as insignificant.

citation as the dependent variable, we find that after correcting the *p*-value for multiple testing, size, stock liquidity, R&D stock, and industry citation intensity still yield statistical significance, verifying the variable selection outcome. In sum, although these tests lack prediction power via cross-validation, we observe that the key variables identified are indeed a subset of these surviving factors.[20]

3.2 Evaluating recently proposed innovation determinants

This section assesses the importance and usefulness of these key covariates in recent studies of corporate innovation. Our purpose is to demonstrate the influence of omitting the key identified variables on previous studies that examine innovation determinants. We do not intend to invalidate the findings of the earlier studies. Instead, we focus on how the inferences or the interpretations of the results could differ.[21]

The first example is a recent study that shows the importance of an underlying managerial trait, CEO with pilot license, in corporate innovation. Sunder, Sunder, and Zhang (2017) suggest that firms with pilot CEOs exhibit more successful innovation and show that their companies generate more patents than comparable firms with nonpilot CEOs. In Table 5, panel A, column 1, we repeat their original specification, documenting that pilot CEOs are positively associated with patents (*t*-statistics > 1.97). At the same time, it adds limited explanatory power to the model (accounting for 0.002 of R^2).[22] In column 2, we add the key explanatory variables that we have

[20] We note that since the variables are standardized in the regression, the economic significance can be directly inferred from the coefficient estimate. The top-three variables ranked by economic significance for patent test are size, industry citation intensity, and R&D stock. For the citation test, the top-three variables are industry citation intensity, size, and stock liquidity.
[21] We choose the following three studies based on the different econometric approaches they adopt in their studies. The first study (pilot CEO) utilizes industry fixed effects, while the second (antitakeover device) adopts firm fixed effects. The third one uses exogenous shock in their study to establish identification.
[22] Our pilot data may not be identical to that of Sunder, Sunder, and Zhang (2017), who were unable to provide their data for verification purposes. Stephen McKeon kindly supplied the CEO pilot data used in their study of CEO pilot

identified from the original specification to assess these key variables' importance when including industry fixed effects.[23] This test's results reveal that pilot CEOs are not significantly related to innovation (*t*-statistics < 1.18). In contrast, the key variables add substantially to the model fit in this analysis, increasing adjusted R^2 by 0.072 (13.5%) relative to the baseline analysis in column 1. In column 3, we drop R&D stock from the key variables; we find that pilot CEO remains insignificant (*t*-statistics < 1.62). In sum, we find that this newly documented managerial trait indicator, CEOs with a pilot license, is not associated with innovation quantity or productivity after controlling for previously identified determinants of corporate innovation.

Table 5, panel B, examines the second study that uses firm fixed effects to explore the impact of antitakeover devices on corporate innovation. More specifically, Chemmanur and Tian (2018) show that a firm's corporate governance strength, via antitakeover provisions, is associated with future innovation output. In panel B, column 1, we replicate their specification using future patents as the dependent variable and incorporating firm fixed effects. Similar to their result, we find that antitakeover provisions are positively related to the firm's patents (*t*-statistic > 2.37). In column 2, we document that after including analyst following, R&D stock, stock liquidity, and industry patent and citation intensity, the antitakeover effect becomes insignificant (*t*-statistic < 1.06). In column 3, without R&D stock, we still find that the antitakeover effect is negligible. We also note that the other five key variables are significant even with firm fixed effects except for stock liquidity.[24]

In a recent study, Mukherjee, Singh, and Zaldokas (2017) rely on staggered changes in state-level corporate tax rates as an identification strategy and show that a tax rate increase is negatively

---

risk-taking (Cain and McKeon 2016). The descriptions of the data collection process in both studies appear similar but could still differ in the underlying data obtained.

[23] We do not include CEO centrality to minimize the influence on sampling.

[24] We find similar inferences when we exclude the two industry-level key variables.

associated with future patents. Table 5, panel C, column 1, using the original specification, shows that the dummy variable indicating state tax increase is negatively associated with the change in patents. In column 2, we include the key variables, and we show that the effect in column 1 remains robust. We find that the magnitude of the effect remains similar, while the statistical significance increases, suggesting that including the key variables improves model fit.

3.3 Discussions

We discuss these findings to put them in context. Maybe not surprisingly, the first two studies above show evidence consistent with the conventional wisdom about an omitted variable problem; that is, omitted variables often result in weaker significance in the focused covariate. Does that mean that these studies are wrong? The answer is "not really." It depends on the purpose of the research question. For instance, the pilot CEO study is consistent with the notion that risk tolerance is related to corporate innovation. However, it becomes insignificant after including the key variables. The explanation could be that one or more of the key variables capture risk tolerance better than pilot CEO. As such, it does not mean that pilot CEO is not related to corporate innovation. However, it does change the inference of pilot CEO on corporate innovation after the key variables are included as controls. In sum, whether or not to include the key variables as controls depends on the nature of the specific research question.

Nevertheless, it is important to consider these key controls because, at a minimum, they could change the inference or the interpretation of the newly proposed covariate for innovation. Furthermore, the third study indicates that because the key variables identified provide significant incremental explanatory power, one benefit of including them is improving model fitness, which results in a lower model's standard error. Overall, we suggest that including the key identified variables could significantly benefit new research on corporate innovation.

4. Fixed Effects and Exclusion Condition Revisited

To the extent that the purpose of most empirical studies is to understand marginal causal effects instead of developing the best predictive model, the control variables are to isolate alternative explanations and help evaluate the validity of the proposed research question. Two common approaches to achieving that goal include fixed effects in the regression analysis or relying on the exogenous shock to the covariate to establish identification. We explore the usefulness of our key identified variables in the context of these two methods.

We first assess whether controlling for industry and/or firm fixed effects makes the key variables identified above redundant. Studies in innovation often include industry fixed effects to address two empirical issues. One effect is the nonrandom distribution of firms with and without patents across different industries. In other words, many of their member firms choose to apply for patents in some industries, while in other industries, the opposite is the common practice. Including industry fixed effects is often adopted to mitigate the self-selection bias in patent choice. Yet another issue is that of omitted industry characteristics. Industry fixed effects only address this problem if the omitted industry factors are time-invariant. Likewise, firm-level fixed effects are commonly used in innovation studies to address firm-level time-invariant omitted variable concerns.[25]

Table 6 presents the results with the seven key identified variables and a mix of industry and firm fixed effects. In panel A, we show that in column 1, the base case, we only include the year fixed effects, the same as in the variable selection process. Not surprisingly, all the key variables

[25] Even though fixed effects are meant to mitigate industry or firm-level time-invariant variables and our key variables are time variant, the necessity of controlling for the key variables is still relevant because of the correlations between the time-invariant fixed effects and the key variables.

remain significant. In column 2, we add the industry fixed effects, and the results show that except for industry citation patent intensity, other key variables remain significant. This suggests that controlling for industry fixed effects does not make the key variables redundant. In column 3, we control for firm fixed effects. The results suggest that stock liquidity and industry patent intensity are still significant. Prior studies emphasize those liquidity proxies for ease of financing, with cross-sectional tests revealing that innovation is increasing in liquidity. Strikingly, when adding a firm fixed effect and no longer differentiating across firms, the sign on the liquidity coefficient estimate changes from positive to negative. This change in the sign arises because liquidity here is measured contemporaneously with innovation. That the change in stock liquidity and firm fixed effects is highly correlated explains our findings. Presumably, studies focusing on changes in liquidity and innovation would likely use lagged liquidity instead.

In column 4, we use industry-year pairwise fixed effects, which is more restrictive than the industry fixed effects shown in column 2. We find that, similar to the findings in column 2, 5 of the 7 variables (two industry variables are dropped automatically) still survive the fixed effects. Finally, in column 5, we show that adding firm fixed effects to the top of industry-year fixed effects leaves only firm size significant. Overall, these results suggest that the commonly used fixed effects approach does not rule out the necessity of using the key variables. However, the specification of all three types of fixed effects (as in column 5) may not be feasible in practice, especially when the focused variable has low time-series variations (e.g., board size, corporate bylaw, antitakeovers).

In panel B, the citation test based on four key variables yields similar inference as the patent test. Controlling for industry or firm fixed effects is insufficient to provide the explanatory power if the key variables are omitted. Notably, in columns 4 and 5, even after controlling for industry-year pairwise and/or firm fixed effects, stock liquidity is still highly significant, suggesting that it provides important incremental explanatory power, not captured by time-invariant industry firm factors.

A common and more prevalent approach for causal inferences about corporate innovation is using an instrumental variable with an exogenous event. In their study, Aghion, Van Reenen, and Zingales (2013) provide a theory explaining why institutional ownership could lead to more corporate innovation. They exploit exogenous variation in institutional ownership and show that it leads to increased innovation. Specifically, they use S&P 500 membership as an exogenous event that increases institutional ownership in the firm, finding that S&P 500 inclusion is followed by increased innovation output.

In panel A of Table 7, we replicate the instrumental variable approach in the original study in which the authors use S&P 500 membership as an instrumental variable for institutional ownership. We first show the original results in columns 1–3. Column 1 shows the baseline Poisson regression results. Columns 2 and 3 show the first- and second-stage instrument variable regression results. In column 4, we repeat column 2 with three additional key variables (the other key variables are already in the original specifications). First, we find that the S&P 500 variable is still positive and highly significant at 1%, suggesting that it is a strong instrument for institutional ownership. Second, in column 5, we show the second stage IV regression results with the key variables. We find that the institutional ownership effect flips its sign and becomes both negative and significant, indicating that the omitted key variables significantly influence the IV regression outcome.

We further explore the validity of the instrument variable. To be a valid instrument variable, the candidate variable must satisfy both relevance and exclusion conditions.[26] In this case, the results suggest that the exclusion condition is likely to be violated because, presumably, the S&P 500 index inclusion changes a firm's stock liquidity, which prior research indicates is strongly related to innovation. If that is indeed the case, the effect of the instrument variable on innovation is also manifested via other channels besides institutional ownership, which invalidates the instrument variable. In Table 7, panel B, we show that across different windows over the S&P 500 index shock, there is an evident change in stock liquidity among the treatment firms. The evidence suggests that the instrument variable does not satisfy the exclusion condition.

This result suggests another potential interpretation of the Aghion, Van Reenen, and Zingales (2013) finding that institutional ownership leads to more innovation. Whether the mechanism is: larger institutional ownership leads to more innovation directly or, alternatively, larger institutional ownership leads to more liquidity, which in turn leads to more innovation, is unresolved. It may be that higher stock liquidity correlates with a lower cost of capital, and this increases net present value (NPV) of innovation investments and leads to more patenting, but then the effect is through the cost of capital. In this view, Aghion, Van Reenen, and Zingales (2013) result can be understood as a lower cost of capital leads to more innovation; institutional ownership is able to lower the cost of capital specifically for innovation projects.

5. Concluding Remarks

[26] Roberts and Whited (2013) suggest that corporate finance research tends to emphasize the relevance of the instrument but often gives limited attention to the exclusion restriction. One impediment in testing the exclusion restriction is the identification of the relevant potential covariates to consider in the evaluation.

In this study, we accomplish three tasks. Using innovation research as a platform, we apply machine learning techniques to identify key variables that provide independent explanatory power among 35 previously hypothesized innovation determinants. Our base analysis uses the adaptive Lasso technique to assess the previously proposed determinants of corporate innovation's explanatory power. A major advantage of these techniques is their ability to limit concerns about overfitting or avoid mistaking random variations in the data as underlying trends (Caner and Fan 2015). Adaptive Lasso provides a prevalent approach to variable selection in the natural sciences. It can give the same coefficient estimates as if one knew, with high probability or asymptotically, the true underlying model, that is, the oracle property (Zou 2006; Hui, Warton, and Foster 2015).

Both parametric and nonparametric variable selection methods identify a few high explanatory variables from previously hypothesized innovation determinants. Notably, 6 of the 7 variables specified in the adaptive Lasso, group Lasso, elastic net, and the stepwise analyses are the same. Undertaking a similar investigation for citations, we document that four previously identified covariates represent a tractable set of relevant control variables. Fundamental firm and industry characteristics, such as firm size, R&D stock, stock liquidity, and industry innovation intensity, provide independent explanatory power for patents and their citations. Most surprisingly, we find that stock liquidity (e.g., Bernstein 2015; Fang, Tian, and Tice 2014) offers the most substantial explanatory power among previously proposed innovation determinants. Beyond demonstrating the importance of access to capital markets is crucial to corporate innovation, this finding merits future research on stock liquidity in a more theoretical manner.

Second, besides the seemingly random choice of control variables across disciplines in innovation research, we show how previous studies' inferences may change after including the key identified variables. We do not advocate that new research should *always* condition their empirical analysis on corporate innovation's HEP variables. Nor do we propose that our HEP variables are the only set of variables to consider. Whether to include certain control variables depends on the nature of the research question. Still, our analysis offers two benefits along this line. One is that these HEP variables help with interpreting the underlying mechanism of the newly proposed covariate. In addition, these variables provide a starting point for studies seeking to model the selection process when developing a first-stage model for the self-selection of firms' patenting choices.

Lastly, our study shows that the popular methods of using fixed effects to treat omitted variable bias do not invalidate the need to include these prior determinants of patent activity in assessing newly proposed determinants of innovation. We demonstrate that even the exogenous shock approach does not mitigate the omitted variable problem in studies providing causal evidence on corporate innovation.

References

Aghion, P., J. Van Reenen, and L. Zingales. 2013. Innovation and institutional ownership. *American Economic Review* 103:277–304.

Anderson, R. C., D. M. Reeb, and W. Zhao. 2012. Family-controlled firms and informed trading: Evidence from short sales. *Journal of Finance* 67:351–85.

Atanassov, J. 2013. Do hostile takeovers stifle innovation? Evidence from antitakeover legislation and corporate patenting. *Journal of Finance* 68:1097–131.

Balsmeier, B., L. Fleming, and G. Manso. 2017. Independent boards and innovation. *Journal of Financial Economics* 123:536–57.

Bernstein, S. 2015. Does going public affect innovation? *Journal of Finance* 70:1365–403.

Bertoni, F., and T. Tykova. 2015. Does governmental venture capital spur invention and innovation. *Research Policy* 44:925–35.

Caner, M., and Q. Fan. 2015. Hybrid generalized empirical likelihood estimators: Instrument selection with adaptive lasso. *Journal of Econometrics* 187:256–74.

Cain, M., and S. McKeon. 2016. CEO personal risk-taking and corporate policies. *Journal of Financial and Quantitative Analysis* 51:139–64.

Chemmanur, T. J., and X. Tian. 2018. Do antitakeover provisions spur corporate innovation? A regression discontinuity analysis. *Journal of Financial and Quantitative Analysis* 53:1163–94.

Einav, L., and J. Levin. 2014. Economics in the age of big data. *Science* 346:1243089.

Eisfeldt, A. L., and D. Papanikolaou. 2013. Organization capital and the cross-section of expected returns. *Journal of Finance* 68:1365–406.

Faleye, O., T. Kovacs, and A. Venkateswaran. 2014. Do better-connected CEOs innovate more? *Journal of Financial and Quantitative Analysis* 49:1201–25.

Fang, V. W., X. Tian, and S. Tice. 2014. Does stock liquidity enhance or impede firm innovation? *Journal of Finance* 69:2085–125.

Galasso, A., and T. S. Simcoe. 2011. CEO overconfidence and innovation. *Management Science* 57:1469–84.

Gompers, P. A., J. L. Ishii, and A. Metrick. 2003. Corporate governance and equity prices. *Quarterly Journal of Economics* 118:107–55.

Griliches, Z., A. Pakes, and B. H. Hall. 1987. The value of patents as indicators of inventive activity. In *Economic Policy and Technological Performance*, chap. 4, 97–124, eds. P. Dasgupta and P. Stoneman. Cambridge, UK: Cambridge University Press.

Harvey, C. R. 2017. Presidential address: The scientific outlook in financial economics. *Journal of Finance* 72:1399–440.

Hall, B. 1990. The manufacturing sector master file: 1959-1987. Working Paper, University of California at Berkeley.

Hall, B., A. Jaffe, and M. Trajtenberg. 2001. The NBER patent citations data file: Lessons, insights, and methodological tools. Working Paper, University of California, Berkeley.

He, J. (Jack), and X. Tian. 2013. The dark side of analyst coverage: The case of innovation. *Journal of Financial Economics* 109:856–78.

———. 2018. Finance and corporate innovation: A survey. *Asia-Pacific Journal of Financial Studies* 47:165–212.

Hui, F. K. C., D. I. Warton, and S. D. Foster. 2015. Tuning parameter selection for the adaptive Lasso using ERIC. *Journal of the American Statistical Association* 110:262–69.

Im, H. J., and J. Shon. 2019. The effect of technological imitation on corporate innovation: Evidence from US patent data. *Research Policy* 48:103802.

Intintoli, V., K. Kahle, and W. Zhao. 2018. Director connectedness: Monitoring efficacy and career prospects. *Journal of Financial and Quantitatively Analysis* 53:65–108.

Kaplan, S. N., and L. Zingales. 1997. Do investment-cash flow sensitivities provide useful measures of financing constraints? *Quarterly Journal of Economics* 112:169–215.

Koh, P-S., D. M. Reeb, E. Sojli, and W. Tham. 2016. Measuring innovation around the world. Working Paper, Hong Kong University of Science and Technology.

Kogan, L., D. Papanikolaou, A. Seru, and N. Stoffman. 2017. Technological innovation, resource allocation, and growth. *Quarterly Journal of Economics* 132:665–712.

Lerner, J., and A. Seru. 2017. The use and misuse of patent data: Issues for corporate finance and beyond. Working Paper, Harvard Business School.

Malmendier, U., and G. A. Tate. 2005. CEO overconfidence and corporate investment. *Journal of Finance* 60:2661–700.

Mukherjee, A., M. Singh, and A. Zaldokas. 2017. Do corporate taxes hinder innovation? *Journal of Financial Economics* 124:195–221.

Mulliainathan, S., and J. Spiess. 2017. Machine learning: An applied econometric approach. *Journal of Economic Perspectives* 31:87–106.

Roberts, M. R., and T. M. Whited. 2013. Endogeneity in empirical corporate finance. In *Handbook of the Economics of Finance* 2, Part A, 493–572. Amsterdam, the Netherlands: Elsevier.

Sunder, J., S. V. Sunder, and J. Zhang. 2017. Pilot CEOs and corporate innovation. *Journal of Financial Economics* 123:209–24.

Taddy, M. 2017. One-step estimator paths for concave regularization. *Journal of Computational and Graphical Statistics* 26:525–36.

Zou, H. 2006. The adaptive Lasso and its oracle properties. *Journal of American Statistical Association* 101:1418–29.

Appendix A. Control Variable Candidates from Prior Studies

We explain how we select the 35 variables in our study from the control variables in prior studies. It is essential to be clear about the purpose of our study that guides our selection process. Specifically, we are interested in aggregating the innovation literature on the cross-section determinants of innovation performance. As such, replicability with wide data availability is a crucial condition and a principal for choosing the variables in prior studies. There are several scenarios in the literature that limit our choice of the variable in those studies.

First, we tried our best to construct the variables based on publicly available data. That is important because, for our purpose, we need to ascertain that the variables are comparable to the original studies, if not identical. Consequently, we do not include variables that are based on proprietary data or unpublished hand-collect data, for example, pilot CEO.

The second scenario involves several studies that focus on a single industry that we deem inappropriate for our study's purpose. Similarly, studies that using international samples do not fit into our purpose as well.

Third, some variables are obtained from survey data that we cannot obtain, and as they are generally not available to other researchers, we decide not to include those factors.

Fourth, some studies (especially those in the management/strategy discipline) tend to focus on a firm's unique perspective, that is, the firm's connection with similar other firms. Again, we do not include those factors as these variables are difficult to obtain and not used in other studies.

Fifth, we do not garner the variables in the studies that examine exogenous legal changes to innovation. For one, they often rely on fixed effects as they argue that exogenous shock mitigates the need for further control variables. On top of that, these studies do not fit with our purpose because the factors do not offer cross-section inferences. These include some international studies on the legal system or financial market development.

Sixth, studies that focus on inventor-level instead of firm-level inferences are helpful if they also contain some firm-level examinations.

In sum, we focus on the cross-section determinants of firm-level innovation performance based on publicly available data and replicability. We choose 35 variables from the prior studies based on these principles.

The following table shows the statistics on the innovation determinants examined in prior studies. According to our explanation above, we select the 35 variables in our study based on data availability, the usage in the literature, and the suitability for our research purpose.

Variables	Number of times used in prior studies	Included in our study?
Firm size	33	Yes
Tangible assets	16	Yes
Stock returns	5	Yes
Stok return volatility	2	No
ROA	17	Yes
Market-to-book ratio	7	Yes
Growth	4	Yes
R&D spending	23	Yes
Firm age	29	Yes
Stock liquidity	2	Yes
Capital expenditures	10	Yes
Capital structure	16	Yes
Competition	16	Yes

Capital labor ratio	8	No
Tobin's q	10	Yes
Organization capital	1	Yes
Public firm	3	No
Cash	7	No
Cash flow	1	No
Large shareholder	1	Yes
Institutional ownership	8	Yes
Analyst following	1	Yes
CEO confidence	4	Yes
CEO total pay	1	Yes
CEO centrality	1	Yes
CEO tenure	3	Yes
CEO delta	5	Yes
CEO vega	3	Yes
CEO gender	1	Yes
CEO age	2	No
Top university	1	No
Finance education	1	No
Technical education	2	No
PhD in technical education	2	No
Military CEO	1	No
Pilot CEO	1	No
Distance to USPTO	1	Yes
Manufacturing industry	1	Yes
Entrenchment index	1	Yes
Board size	1	Yes
Board independence	1	Yes
Family firm	1	Yes
State GDP	3	No
State tax rate	1	No
Population	2	No
Industry R&D	1	Yes
Industry citation intensity	1	Yes
Industry patent intensity	2	Yes
Industry size	1	Yes
Current ratio	1	No
Employee stock options	1	No
Ratio of value added	1	No
Colleges	1	No
Enrollment	1	No
Unemployment insurance	1	No
University density	1	No
Share of firm's preinvestment patent in class	1	No
Business combination law	1	No
Banking deregulation	1	No
VC backed	2	No
Pioneer	1	No
Early follower	1	No
Option volume	1	No
Diversification	1	No
Number of firms financed in the same quarter	1	No
Number of investors in the syndicate	1	No

Amount of money raised in the first round of funding	1	No
Firm based in Massachusetts or California	2	No
Foreign institutional ownership	1	No
Insider ownership	3	No
Foreign sales ratio	1	No
Change in gross state product	1	No
Change in total tax revenues of the state as a proportion of GSP	1	No
Change in the state's population	1	No
Change in the state's unemployment rate	1	No
State tax rate change	1	No
Change in profitability	1	No
Ability	1	No
Presence of a debt rating on the firm	1	No
Diversity	1	No
Reallocate	1	No
Banking deregulation	1	No
Financial distress	3	No
Business group	1	No
Number of research sites	2	No
Local university ties	1	No
Network size	1	No
Alliance duration	1	No
Patent stock	1	No
Technological diversity	2	No
VC failure tolerance	1	No
Wrongful discharge law	1	No
Banking competition	1	No
External financing dependence	1	No
Homestead exemptions	1	No
Employment nondiscrimination acts	1	No

Appendix B. Machine Learning and Variable Selection

Machine learning encompasses a variety of techniques for identifying patterns and relationships in the data, and it is commonly used in forecasting and to simplify model selection processes. In the realm of machine learning, our interest lies in finding an econometric model that maps the set of variables that potentially explain an output, in this case, patents or citations. Among the multiple methods available, such as subset selection, least squares, generalized additive models, trees, support vector machines, bagging, and boosting, the Lasso regression offers a balanced trade-off between interpretability and flexibility. The more flexible the method, the lower its bias has since it can better approximate the true relationship existing in the data. However, increased flexibility increases the variance of the method since it attempts to fit not only true data points but also the unavoidable noise present in the data set.

Lasso (least absolute shrinkage selection operator) is a shrinkage method that reduces (or shrinks) the values of the coefficients to zero compared with ordinary least squares. The advantage of shrinkage methods is that the estimated model exhibits lower variance than those of least squares estimates. We compare Lasso with least squares estimation as follows:

$$\text{Least squares: } \frac{min}{\beta_0,\beta_j} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_1^p \beta_j \, x_{ij})^2$$

$$\text{Lasso regression: } \frac{min}{\beta_0,\beta_j} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_1^p \beta_j \, x_{ij})^2 \text{ subject to } ||\beta||_1 \leq t,$$

where y is the vector of observations of the dependent variable, x denotes the independent variables, β are the corresponding coefficients, $|| \, ||_1$ and $|| \, ||_2$ are the L1 and L2 norms, respectively, and t is a user-specified parameter. The Lagrange formulation of the Lasso regression is

$$\text{Lasso regression: } \frac{min}{\beta_0,\beta_j} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_1^p \beta_j \, x_{ij})^2 + \lambda ||\beta||_1.$$

The least squares estimation corresponds to an unconstrained minimization problem; the Lasso regression imposes a convex but nonsmooth l_1 constraint. Least squares analysis rewards, including as many covariates as possible, since additional right-hand-side variables help to reduce the sum of the squares. However, Lasso regression imposes a penalty factor on the coefficients that reduces the value of the coefficients or the number of factors included in the model. As such, Lasso regressions appear well-suited to addressing the model selection challenge when developing forecast models that work well with out-of-sample data. In addition, the Lasso regression performs both the variable selection and the parameter estimation simultaneously. Lasso exhibits both low variability and limited computational costs, especially in high-dimensional problems.

Lasso regression solutions coincide with the least squares solution if the penalty parameter is set sufficiently small. In a Lasso regression, the penalty parameter controls both the size and the number of coefficients, with higher values leading to fewer covariates in the linear model. This increases the model's flexibility and reduces its variance but at the cost of a higher model bias. Lasso regression utilizes cross-validation, a resampling technique, to facilitate finding a parameter value that ensures a proper balance between bias and variance (or flexibility and interpretability), minimizing the estimated test error rate of the estimator. In cross-validation, a subset of the data observations, the training set, is used to estimate (or train) the model; the remaining observations are held to serve as the test set or validation set. The selected test sets serve to provide an estimate of the test error rate. Typically, the measure of the test error is the mean square error (MSE).

The K-fold cross-validation method divides the data set randomly into K different subsets, in which we set $K = 10$. Keeping one of the subsets as the validation set, we estimate the model over the remaining $K - 1$ sets for a range of penalty parameter values. We repeat this process using each of the K subsets as a validation set, yielding K estimates of the MSE for each parameter value; its K-fold estimate is simply the average value of the K estimates. The best parameter value is the one yielding the lowest K-fold estimate, which we denote as λ-min in the Lagrange formula. This parameter estimate is the one-standard-error rule parameter, $\lambda - 1SE$.

Adaptive Lasso has been developed from Lasso to address the issue that Lasso does not possess the oracle property. Note that an estimator that is consistent in variable selection is not necessarily consistent in parameter estimation. An oracle estimator must be consistent in both. Adaptive Lasso performs a different regularization for each coefficient, adjusting the penalty factor differently for each coefficient, avoiding overfitting by penalizing large coefficients. Consequently, adaptive Lasso penalizes more of the coefficients with lower initial estimates. This adjustment, as compared to Lasso, helps to achieve the oracle property of the estimators (Zou 2006).

We also use two variants of Lasso, namely, group Lasso and the elastic net regression. Group Lasso takes into consideration that the variables within predetermined categories are meant to be selected or unselected together. However, for our purpose, we loosen the “togetherness” restriction on selections at the group level, while we use a methodology that selects at both the group and individual variable levels. In addition, Lasso's penalty function is a linear combination of the coefficients, which tends to select one variable from a group and ignore the others if there is a group of highly correlated variables. To mitigate this concern, we also use the elastic net method, which adds a quadratic part to the L1 penalty function that used alone is the ridge regression penalty form. In other words, the elastic net is another regularized regression method that linearly combines the L1 and L2 penalties of the Lasso and ridge methods.

In contrast to these machine learning approaches, traditional statistical methods for variable selection focus on either the best subset or the best stepwise regressions. The best subset approach chooses the single best model from all possible combinations of the potential covariates. The best model is chosen based on some sort of prespecified information criteria, which typically penalizes the number of nonzero parameters. Two potential issues are the computational difficulty with large p and that it relies on within-sample analysis, leading to concerns of fitting noise in the data. Stepwise regressions follow the same general approach as best subset selection but limit the number of models that need to be analyzed. For instance, with 25 potential covariates, best subset analysis requires estimating 33,554,432 regressions, while stepwise regression only requires 325 regressions. Stepwise variable selection is computationally feasible, but the results are sensitive to the sequencing of the variables. As best subset selection and stepwise variable selection typically rely on fitting the data within the sample, these methods are typically labeled as "greedy" selection methods.

Appendix C. Variable Definitions

Analyst following: the number of financial analysts that follow the firm during the year

Antitakeover: the corporate governance index developed in Gompers, Ishii, and Metrick (2005)

Blockholder: a dummy variable that equals one if at least one institutional investor holds more than 5% of the common equity

Board size: log of the number of directors

Board independence: proportion of independent directors

Bylaw amendments limit: a dummy variable indicating whether the firm has a policy limiting shareholders' ability through a majority vote to amend the corporate bylaws

Capital structure: long-term debt divided by total assets

CEO centrality: a factor score based on the factor analysis of four metrics of the CEO's social connectedness. The data are drawn from BoardEx; we calculate the four metrics following Intintoli, Kahle, and Zhao (2018)

CEO confidence: a dummy variable equals one if the CEO's in-the-money exercisable options exhibit greater than 67% moneyness and it happens twice during the sample period (Malmendier and Tate 2005)

CEO delta: change in CEO wealth in dollars to a 1% change in stock price

CEO gender: a dummy variable indicating whether the CEO is female

CEO total pay: log of CEO annual total compensation (tdc1)

CEO vega: measures CEO wealth change in dollars to one annualized standard deviation of stock return

Citation: log of one plus the number of citations the firm obtains during the year

Cites: number of a firm's patents weighted by the number of future citations

Distance to USPTO: log of distance in miles between the firm headquarters and the USPTO office that oversees the firm's state

Family firm: a dummy variable indicating whether the firm is a family firm as defined in Anderson, Reeb, and Zhao (2012)

Financial distress: the firm's KZ index (Kaplan and Zingales 1997), calculated as $-1.002 \times$ Cash flow $+ 0.28 \times$ Tobin's q $+ 3.18 \times$ Leverage $- 39.368 \times$ Dividends $- 1.315 \times$ Cash holdings

Firm age: the log number of years that the firm appears in Compustat

Golden parachutes: a dummy variable indicating whether the firm has a severance agreement that provides benefits to management/board members in the event of firing, demotion, or resignation following a change in corporate control

Industry patent intensity: total industry patents divided by total industry assets; the industry is SIC two-digit

Industry citation intensity: total industry citations divided by total industry assets; the industry is a SIC two-digit code

Industry R&D: average R&D for each two-digit SIC industry

Industry competition: Herfindahl index (HHI) based on sales for each two-digit SIC industry

Industry size: log of total assets of the firms in each two-digit SIC industry

Institutional ownership: institutional ownership of common equity

Manufacturing: a dummy variable equals one if the firm is in one-digit SIC code 3 or 4

Market-to-book ratio: market value of equity divided by the book value of equity

Mergers and charter amendments: a dummy variable indicating if the firm has a provision limiting shareholders' ability through a majority vote to amend the corporate charter or requires more than a majority of shareholders to approve a merger

Organizational capital: stock of selling, general, and administrative expense (SG&A) as in Eisfeldt and Papanikolaou (2013)

Patent: log of one plus the number of patents applied

Patenting: a dummy variable indicating the firm has patent application during the year

Poison Pill: a shareholder right that is triggered in the event of an unauthorized change in control that typically renders the target company financially unattractive or dilutes the voting power of the acquirer

PPE/EMP: net property, plant, and equipment value divided by the number of employees

R&D: research and development (R&D) expenditures scaled by total assets

R&D stock: the cumulative R&D over the previous 10 years, assuming a depreciation rate of 15%

ROA: income before extraordinary items divided by total assets

S&P 500: a dummy variable equal to one if the firm is included in the S&P 500 index during the year

Sales growth: average growth rate of sales over the prior 3 years

Size: log of total assets

Staggered board: a dummy variable indicating whether the board is staggered, that is, directors are divided into separate classes with each class being elected to overlapping terms

State tax: marginal tax rate of the state

Stock liquidity: log of stock daily trading volume aggregated over the year

Tangibility: net property, plant, and equipment scaled by total assets

Tobin's q: market value of equity plus book value of debt scaled by the book value of total assets

Figure 1. Adaptive Lasso

This figure shows how the log (lambda) corresponds to the number of nonzero variables. Eight variables are chosen to the right of the vertical line of the optimal log (lambda).

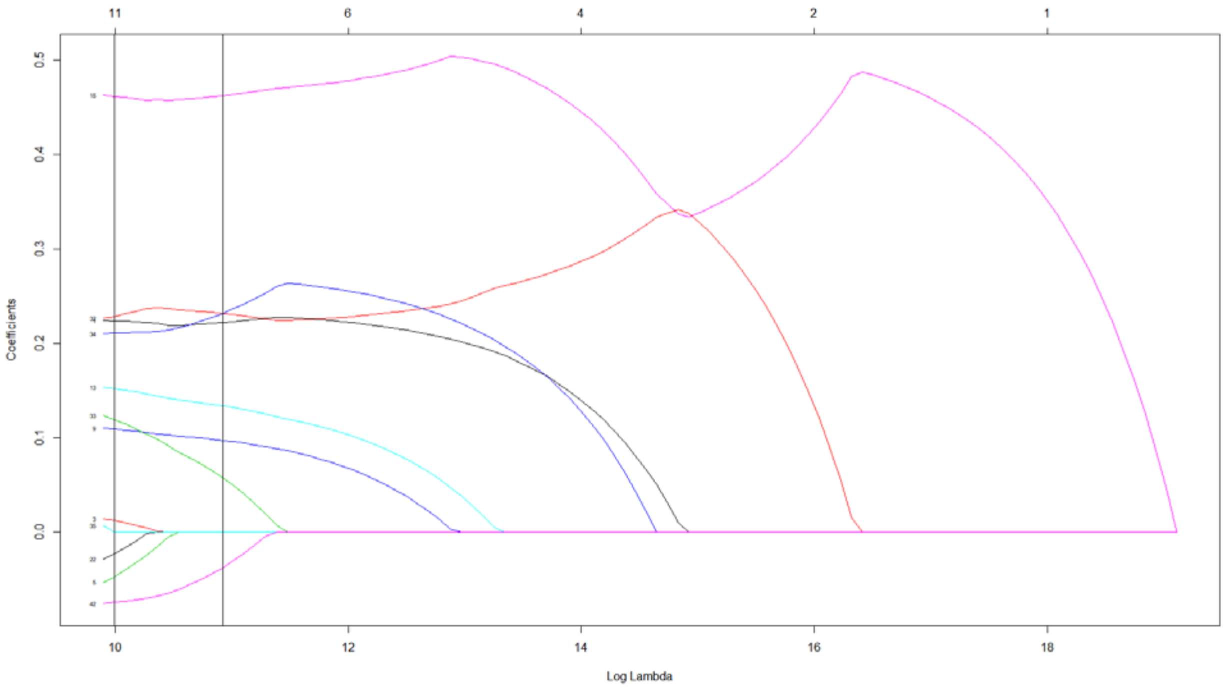

Table 1. Variable selection results: Full sample

This table presents the variable selection results on multiple managerial, governance, firm, and industry factors on patents among patenting firms by adaptive Lasso using the 2001–2010 sample (see Appendix B for a description of how the key variables are selected). Appendix C defines the variables. The incremental R^2 loss is computed by comparing OLS regressions after dropping each variable in the order suggested by the adaptive Lasso variable selection. Intuitively, this method produces results akin to a stepwise regression with the sequence from the adaptive Lasso procedure. "–" denotes variables that are not included in the selection process due to design or data availability. The bolded variables indicate the ones that are chosen by adaptive Lasso.

A. Patents

Sample	2001–2010	1992–2010	2001–2010	1992–2010
	Incremental R^2 loss			
	Without R&D stock		With R&D stock	
Variables	*Patent*			
Managerial characteristics				
CEO centrality	**0.0196**	–	**0.0098**	–
CEO confidence	0.0079	0.0030	0.0059	0.0003
CEO vega	0.0011	0.0034	0.0010	0.0030
CEO delta	0.0014	0.0032	0.0031	0.0012
CEO total pay	0.0001	0.0011	0.0001	0.0017
CEO gender	0.0001	0.0001	0.0001	0.0002
Firm characteristics				
Size	**0.0646**	**0.1011**	**0.0629**	**0.1011**
R&D stock	–	–	**0.0580**	**0.0361**
Tobin's q	0.0015	0.0000	0.0000	0.0007
Stock liquidity	**0.3696**	**0.3320**	**0.3696**	**0.3320**
Firm age	0.0000	0.0022	0.0000	0.0006
Distance to USPTO	0.0073	0.0052	0.0073	0.0024
Tangibility	0.0005	0.0004	0.0003	0.0024
Manufacturing	0.0009	0.0011	0.0015	0.0069
State tax	0.0000	0.0000	0.0000	0.0006
ROA	0.0001	0.0000	0.0001	0.0009
Capital structure	0.0010	0.0002	0.0010	0.0000
Sales growth	0.0001	0.0016	0.0002	0.0013
Organizational capital	0.0012	0.0006	0.0000	0.0000
Corporate governance				
Analyst following	**0.0042**	**0.0041**	**0.0039**	**0.0084**
Poison pill	0.0011	0.0023	0.0012	0.0009
Blockholder	0.0021	0.0029	0.0029	0.0016
Institutional ownership	0.0020	0.0022	0.0020	0.0015
Board size	0.0000	–	0.0000	–
Golden parachutes	0.0007	0.0004	0.0002	0.0019
Board independence	0.0008	–	0.0002	–
Mergers and charter	0.0008	0.0005	0.0008	0.0013
Staggered board	0.0000	0.0002	0.0000	0.0003
Bylaw amendments limit	0.0001	0.0008	0.0001	0.0001
Family firm	0.0022	–	0.0023	–
Industry characteristics				
Industry citation intensity	**0.0580**	**0.0212**	**0.0346**	**0.0147**
Industry patent intensity	**0.0201**	**0.0267**	0.0180	0.0267
Industry size	0.0000	0.0035	0.0008	0.0022
Industry R&D	0.0001	0.0065	0.0001	0.0010

Industry competition	0.0009	0.0011	0.0013	0.0001

B. Citations

Sample	2001–2010	1992–2010	2001–2010	1992–2010
	Incremental R^2 loss			
	Without R&D stock		With R&D stock	
Variables	*Citation*			
Managerial characteristics				
CEO centrality	**0.0301**	–	**0.0258**	–
CEO confidence	0.0040	0.0037	0.0037	0.0026
CEO vega	0.0003	0.0004	0.0003	0.0001
CEO delta	0.0008	0.0004	0.0004	0.0003
CEO total pay	0.0009	0.0017	0.0009	0.0013
CEO gender	0.0003	0.0003	0.0004	0.0013
Firm characteristics				
Size	**0.0264**	**0.0270**	**0.0264**	**0.0270**
R&D stock	–	–	0.0120	0.0090
Tobin's q	0.0031	0.0036	0.0013	0.0010
Stock liquidity	**0.2797**	**0.2394**	**0.2797**	**0.2394**
Firm age	0.0018	0.0010	0.0011	0.0003
Distance to USPTO	0.0044	0.0046	0.0006	0.0005
Tangibility	0.0001	0.0002	0.0001	0.0001
Manufacturing	0.0022	0.0025	0.0019	0.0011
State tax	0.0002	0.0004	0.0002	0.0000
ROA	0.0015	0.0015	0.0004	0.0018
Capital structure	0.0009	0.0010	0.0009	0.0006
Sales growth	0.0001	0.0001	0.0001	0.0010
Organizational capital	0.0011	0.0010	0.0010	0.0003
Corporate governance				
Analyst following	0.0020	0.0029	0.0020	0.0017
Poison pill	0.0028	0.0028	0.0033	0.0046
Blockholder	0.0019	0.0014	0.0018	0.0008
Institutional ownership	0.0055	0.0075	0.0055	0.0082
Board size	0.0000	–	0.0000	–
Golden parachutes	0.0001	0.0001	0.0001	0.0024
Board independence	0.0001	–	0.0001	–
Mergers and charter	0.0009	0.0010	0.0004	0.0007
Staggered board	0.0007	0.0007	0.0005	0.0005
Bylaw amendments limit	0.0012	0.0000	0.0015	0.0014
Family firm	0.0000	–	0.0000	–
Industry characteristics				
Industry citation intensity	**0.0350**	**0.0340**	**0.0350**	**0.0340**
Industry patent intensity	0.0007	0.0005	0.0008	0.0094
Industry size	0.0051	0.0058	0.0070	0.0083
Industry R&D	0.0016	0.0006	0.0019	0.0019
Industry competition	0.0019	0.0017	0.0000	0.0025

Table 2. Variable selection results: Alternative methods

This table shows the group Lasso, elastic net regression, and stepwise results using the 2001–2010 sample. Appendix C defines the variables. The bolded variables indicate one that are chosen by each model.

	Patent			Citation		
	Group Lasso	Elastic net	Stepwise	Group Lasso	Elastic net	Stepwise
Managerial characteristics						
CEO centrality	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	No
CEO confidence	No	No	No	No	No	No
CEO vega	**Yes**	No	No	No	No	No
CEO delta	No	No	No	No	No	No
CEO total pay	No	No	No	No	No	No
CEO gender	No	No	No	No	No	No
Firm characteristics						
Size	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**
R&D stock	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**
Tobin's q	No	No	No	No	No	No
Stock liquidity	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**
Firm age	No	No	No	No	No	No
Distance to USPTO	No	No	**Yes**	No	No	No
Tangibility	No	No	No	No	No	No
Manufacturing	No	No	No	No	No	No
State tax	No	No	No	No	No	No
ROA	No	No	No	No	No	No
Capital structure	No	No	No	No	No	No
Sales growth	No	No	No	No	No	No
Organizational capital	No	No	No	No	No	No
Corporate governance						
Analyst following	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	No
Poison pill	No	No	No	No	No	No
Blockholder	No	No	No	No	No	No
Institutional ownership	**Yes**	No	No	**Yes**	No	No
Board size	No	No	No	No	No	No
Golden parachutes	No	No	No	No	No	No
Board independence	No	No	No	No	No	No
Mergers and charter	No	No	No	No	No	No
Staggered board	No	No	No	No	No	No
Bylaw amendments limit	No	No	No	No	No	No
Family firm	No	No	No	No	No	No
Industry characteristics						
Industry citation intensity	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**
Industry patent intensity	No	**Yes**	**Yes**	**Yes**	No	No
Industry size	**Yes**	No	No	**Yes**	No	No
Industry R&D	No	No	No	**Yes**	No	No
Industry competition	No	No	No	No	No	No

Table 3. Variable selection and regime stability

This table presents the results of different tests checking the regime stability of the selected variables. Panel A shows the variable selection outcome via adaptive Lasso using different samples as we expand the sample size by data availability. Columns 1 and 3 drop the CEO characteristics. Columns 2 and 4 drop the antitakeover provisions. Panel B shows the variable selection outcome with adaptive Lasso when we check the 1990–2000 subperiod.

A. Variable selection across regimes

	(1)	(2)	(3)	(4)
Sample	1990–2010			
	Patent		*Citation*	
Variable chosen	*Firm size* *R&D stock* *Analyst following* *Stock liquidity* *Industry patent Intensity* *Industry citation Intensity*	*Firm size* *R&D stock* *Analyst following* *Stock liquidity* *Industry patent Intensity* *Industry citation Intensity*	*Firm size* *R&D stock* *Analyst following* *Stock liquidity* *Industry citation Intensity*	*Firm size* *Analyst following* *Stock liquidity* *Industry citation Intensity*
Observations	25,985	58,671	25,985	58,671

B. Variable selection for 1990–2000

	(1)	(2)	(3)	(4)
Sample	1990–2000			
	Patent		Citation	
Variable chosen	*Firm size* *R&D stock* *Analyst following* *Stock liquidity* *Industry patent Intensity* *Industry citation Intensity*	*Firm size* *R&D stock* *Analyst following* *Stock liquidity* *Industry patent Intensity* *Industry citation Intensity*	*Firm size* *R&D stock* *Analyst following* *Stock liquidity* *Industry citation Intensity*	*Firm size* *Analyst following* *Stock liquidity* *Industry citation Intensity*
Observations	13,024	30,208	13,024	30,208

Table 4. Factors on innovation: An inclusive test

This table presents the results of multiple managerial and firm factors on innovation among patenting firms. Appendix C defines the variables. In columns 1 and 2, we check the significance of each variable when we include them one at a time, with the key variables included always. In columns 3 and 4, we present the results when we include all the variables simultaneously. The Huber-White Sandwich estimator is clustered at the firm level. We correct the significance level for multiple testing in columns 3 and 4 via Bonferroni correction. *$p < .1$; **$p < .05$; ***$p < .01$.

	(1)	(2)	(3)	(4)
	With key variables		Include all	
Dependent variable	*Patent*	*Citation*	*Patent*	*Citation*
CEO total pay	-0.047	-0.059	-0.043	-0.047
	(-1.49)	(-1.25)	(-1.30)	(-1.05)
CEO delta	-0.046**	-0.028	-0.105	-0.041
	(-2.01)	(-1.14)	(-2.71)	(-0.98)
CEO vega	0.029	0.007	0.126	0.068
	(1.08)	(0.21)	(3.01)	(1.48)
CEO centrality	–	–	0.156**	0.139
			(3.53)	(1.79)
CEO gender	0.000	0.028	0.009	0.029
	(0.00)	(0.74)	(0.30)	(0.75)
CEO confidence	-0.059**	-0.058**	-0.059	-0.062
	(-2.34)	(-2.24)	(-2.15)	(-2.06)
Size	–	–	0.489***	0.316***
			(8.37)	(4.63)
Tobin's q	-0.052*	-0.045	-0.001	-0.005
	(-1.79)	(-1.33)	(-0.04)	(-0.16)
Stock liquidity	–	–	0.195***	0.211**
			(4.14)	(3.43)
R&D stock	–	–	0.235***	0.137**
			(5.29)	(3.28)
Manufacturing industry (one-digit SIC = 3, 4)	0.086**	0.050	0.064	0.053
	(2.27)	(1.25)	(1.53)	(1.18)
Firm age	0.032	0.047*	0.047	0.063
	(1.29)	(1.88)	(1.87)	(2.41)
Distance to USPTO	-0.094***	-0.035	-0.098	-0.041
	(-2.59)	(-0.88)	(-2.81)	(-1.11)
Tangibility	0.021	-0.025	0.017	-0.026
	(0.65)	(-0.68)	(0.49)	(-0.72)
Analyst following	–	–	0.228	0.240
			(2.82)	(1.92)
Institutional ownership	-0.146*	-0.213**	-0.126	-0.188
	(-1.93)	(-2.08)	(-1.73)	(-1.87)
Blockholder	-0.042*	-0.058*	-0.045	-0.039
	(-1.78)	(-1.69)	(-2.02)	(-1.27)
Family firm	-0.004	0.027	-0.017	-0.008
	(-0.13)	(0.81)	(-0.51)	(-0.25)
ROA	-0.036	-0.043	-0.013	-0.022
	(-1.18)	(-1.29)	(-0.46)	(-0.66)
Capital structure	-0.005	-0.019	-0.017	-0.039
	(-0.20)	(-0.78)	(-0.66)	(-1.55)
Sales growth	-0.021	-0.024	-0.004	-0.006
	(-0.80)	(-1.17)	(-0.20)	(-0.29)
Board size	0.004	0.002	0.004	0.009
	(0.16)	(0.07)	(0.16)	(0.33)
Board independence	-0.018	-0.034	0.007	-0.005
	(-0.66)	(-1.13)	(0.27)	(-0.17)
Staggered board	0.006	-0.055*	0.031	-0.027
	(0.21)	(-1.70)	(1.03)	(-0.83)

Bylaw amendments limit	-0.006	-0.065*	-0.000	-0.053
	(-0.17)	(-1.95)	(-0.00)	(-1.60)
Poisson pill	-0.059*	-0.075**	-0.056	-0.056
	(-1.95)	(-2.18)	(-1.87)	(-1.58)
Golden parachutes	-0.018	-0.028	-0.012	0.003
	(-0.61)	(-0.87)	(-0.43)	(0.08)
Mergers and charter amendments	-0.004	-0.026	-0.019	-0.024
	(-0.16)	(-1.08)	(-0.79)	(-1.04)
Organizational capital	0.031	0.024	0.023	0.030
	(1.02)	(0.66)	(0.84)	(0.80)
State tax	-0.031	-0.041	0.003	-0.012
	(-1.06)	(-1.44)	(0.09)	(-0.40)
Industry patent intensity	–	-0.069	0.082	-0.141
		(-1.57)	(1.50)	(-2.11)
Industry citation intensity	–	–	0.214***	0.317***
			(4.97)	(6.58)
Industry competition	0.046*	0.018	0.037	0.002
	(1.67)	(0.57)	(1.50)	(0.08)
Industry R&D	-0.070*	-0.047*	0.086	0.112
	(-1.72)	(-1.65)	(1.93)	(2.02)
Industry size	-0.114***	-0.088**	-0.122*	-0.107
	(-3.39)	(-2.57)	(-3.07)	(-2.60)
Year dummy	Yes	Yes	Yes	Yes
Observation	2,716	2,716	2,716	2,716
Adjusted R^2	–	–	.600	.424

Table 5. Evaluating previous studies

This table compares the results of including and excluding the key identified variables. We replicate the results in Sunder, Sunder, and Zhang (2017) in panel A, column 1 (columns 2 and 3) without (with) the key identified variables. In panel B, we replicate Chemmanur and Tian's (2018) work without (with) key identified variables in column 1 (columns 2 and 3). In panel C, we replicate the study of Mukherjee, Singh, and Zaldokas (2017). Tax increase/decrease is a dummy variable equal to one in the year of corporate tax increase/deduction for the firms headquartered in a state, and zero otherwise. *log(Sales)* is the natural logarithm of total sales in 2000 dollars. *log(K/L)* is the natural logarithm of the capital-to-labor ratio, where capital is net property, plants, and equipment, and labor is the number of employees. *HHI* refers to the Herfindahl-Hirschman index, computed as the sum of squared market shares of all firms based on sales for each three-digit SIC industry in each year. *Profitability* is earnings before interest and taxes to sales. *Tangibility* is net plant, property, and equipment divided by book assets. *Debt rating* is a dummy variable for firm-years rated by Standard & Poor's. *R&D/sales* is R&D expense divided by sales. *log(RGSP)* is the natural logarithm of real gross state product. *%taxes* is tax revenue as a percentage of gross state product. *log(Population)* is the natural logarithm of state population. *Unemployment rate* is the unemployment rate in a state. Appendix C defines all other variables. **p* < .1; ***p* < .05; ****p* < .01.

A. Pilot CEOs and innovation

	(1)	(2)	(3)
	Without additional controls	With additional controls	
Dependent variable		*Patent*	
Constant	2.990***	3.132***	3.307***
	(6.41)	(7.26)	(8.40)
Pilot CEO	0.350**	0.198	0.275
	(1.97)	(1.17)	(1.61)
Size	1.003***	0.935***	0.696***
	(13.46)	(12.49)	(9.11)
log(PPE/EMP)	0.315***	0.231***	0.196***
	(4.99)	(4.12)	(3.37)
Stock return	0.046**	0.045**	0.083***
	(2.21)	(2.37)	(4.16)
Tobin's q	-0.091**	-0.104**	-0.181***
	(-2.05)	(-2.56)	(-4.32)
Institutional ownership	0.054	-0.126	-0.200*
	(1.13)	(-1.24)	(-1.93)
CEO tenure	0.008	0.014	0.018
	(0.25)	(0.49)	(0.60)
Delta	-0.018	-0.024	-0.042
	(-0.40)	(-0.54)	(-0.88)
Vega	0.204***	0.106**	0.131**
	(3.59)	(2.00)	(2.28)
CEO age	-0.010	0.000	0.002
	(-0.24)	(0.01)	(0.05)
CEO confidence	-0.036	-0.060	-0.042
	(-0.51)	(-0.90)	(-0.60)
Stock liquidity	–	0.339***	0.523***
		(5.34)	(7.74)
Analyst following	–	0.233**	0.319***
		(2.12)	(2.84)
R&D stock	–	0.486***	–
		(9.16)	
Industry patent intensity	–	0.543***	0.620***
		(4.75)	(5.44)
Industry citation intensity	–	0.058	0.025
		(1.07)	(0.46)
Industry and year fixed effects	Yes	Yes	Yes
Observations	4,426	4,426	4,426

Adjusted R^2	.532	.604	.573

B. Firm fixed effects

	(1)	(2)	(3)
	Without additional controls	With additional controls	
Dependent variable		*Patent$_{t+1}$*	
Constant	-0.908***	-1.496***	-1.136***
	(-3.55)	(-4.73)	(-5.73)
Antitakeover	0.096**	0.072	0.069
	(2.37)	(1.05)	(1.00)
Size	0.269***	0.250***	0.271***
	(8.18)	(7.63)	(8.12)
ROA	-0.303**	-0.179	-0.199
	(-2.05)	(-1.26)	(-1.46)
Leverage	-0.214***	-0.188**	-0.231***
	(-2.82)	(-2.50)	(-2.61)
HHI	-3.647***	-2.373***	-2.165***
	(-4.40)	(-2.99)	(-2.87)
HHI^2	7.546***	4.920***	5.105***
	(4.81)	(3.26)	(3.14)
R&D	0.605	-0.249	-0.286
	(1.05)	(-0.45)	(-0.58)
PPE/assets	0.279**	0.306***	0.316***
	(2.44)	(2.77)	(2.66)
CAPX/assets	-0.138	-0.092	-0.100
	(-0.90)	(-0.59)	(-0.31)
Tobin's q	-0.006	-0.007	-0.009
	(-1.00)	(-1.28)	(-1.04)
Financial distress	-0.004	-0.008	-0.007
	(-0.30)	(-0.58)	(-0.59)
Institutional ownership	-0.177***	-0.237***	-0.220***
	(-2.96)	(-3.67)	(-3.70)
Analyst following	–	0.022*	0.019*
		(1.75)	(1.75)
R&D stock	–	0.663***	–
		(3.02)	
Stock liquidity	–	0.016	0.019
		(1.33)	(1.43)
Industry patent intensity	–	3.890***	3.945***
		(4.52)	(4.49)
Industry citation intensity	–	6.281**	5.837**
		(2.26)	(2.49)
Firm fixed effects	Yes	Yes	Yes
Year fixed effects	Yes	Yes	Yes
Observations	19,274	19,274	19,274
Adjusted R^2	.915	.917	.909

C. State tax and innovation

	(1)	(2)
	Without additional controls	With additional controls
Dependent variable	Δ*log(1+#patent)*	
Tax increase	-0.027***	-0.029***
	(-2.66)	(-3.08)
Tax decrease	0.007	0.007
	(0.76)	(0.81)
Δ*log(Sales)*	0.058***	0.059***
	(5.86)	(5.62)
Δ*log(K/L)*	0.010	0.012
	(1.06)	(1.30)
Δ*profitability*	-0.100***	-0.088***
	(-5.35)	(-4.45)
Δ*tangibility*	0.090	0.071
	(1.63)	(1.31)
Δ*debt rating*	0.037	0.038
	(1.17)	(1.25)
Δ*R&D/sales*	0.012**	0.010*
	(2.03)	(1.89)
Δ*HHI*	0.825	0.685
	(1.45)	(1.27)
ΔHHI^2	-4.327**	-3.614**
	(-2.50)	(-2.10)
Δ*log(RGSP)*	0.000*	0.000***
	(1.71)	(2.78)
Δ*%tax*	0.003	0.001
	(0.79)	(0.56)
Δ*log(Population)*	0.995***	0.788***
	(3.56)	(2.95)
Δ*unemployment rate*	0.013	0.012
	(0.94)	(0.97)
Δ*stock liquidity*	–	0.008*
		(1.73)
Δ*analyst following*	–	0.010*
		(1.70)
Δ*industry patent intensity*	–	3.784***
		(13.15)
Δ*industry citation intensity*	–	0.322
		(1.48)
Year fixed effects	Yes	Yes
Observations	34,752	34,752

Table 6. Industry and firm fixed effects

This table checks the robustness of the key factors that we identify when we include industry- and firm-level fixed effects. The OLS regression based on the main sample is applied. "-" denotes variables missing in the specification because they are absorbed by the fixed effect. *$p < .1$; **$p < .05$; ***$p < .01$

A. Patents

	(1)	(2)	(3)	(4)	(5)
Dependent variable			*Patent*		
Stock liquidity	0.252***	0.218***	-0.092**	0.224***	-0.064
	(6.22)	(5.28)	(-2.36)	(5.07)	(-1.41)
Industry citation intensity	0.185***	0.011	0.025	-	-
	(4.95)	(0.30)	(0.90)		
Size	0.449***	0.516***	0.461***	0.513***	0.446***
	(9.38)	(10.18)	(6.85)	(9.50)	(6.36)
R&D stock	0.213***	0.237***	0.069	0.238***	0.107
	(4.79)	(5.46)	(0.84)	(5.21)	(1.30)
Industry patent intensity	0.137***	0.172***	0.235***	-	-
	(3.10)	(3.20)	(5.09)		
CEO centrality	0.154***	0.158***	0.034	0.157***	0.032
	(3.18)	(3.27)	(1.29)	(3.03)	(1.22)
Analyst following	0.112***	0.118***	-0.014	0.121***	-0.018
	(2.97)	(3.02)	(-0.16)	(2.95)	(-0.19)
Firm fixed effect	No	No	Yes	No	Yes
Industry fixed effect	No	Yes	No	No	No
Year fixed effect	Yes	Yes	Yes	No	No
Industry-year fixed effects	No	No	No	Yes	Yes
Observations	2,716	2,716	2,716	2,716	2,716
Adjusted R^2	.562	.609	.907	.592	.910

B. Citations

	(1)	(2)	(3)	(4)	(5)
Dependent variable			*Citation*		
Stock liquidity	0.311***	0.272***	-0.088	0.271***	-0.111*
	(6.17)	(5.35)	(-1.60)	(4.90)	(-1.66)
Industry citation intensity	0.244***	0.198***	0.242***	–	–
	(6.88)	(4.21)	(5.07)		
Size	0.304***	0.384***	0.298***	0.406***	0.332***
	(5.71)	(5.89)	(3.20)	(5.77)	(2.97)
R&D stock	0.140***	0.157***	0.087	0.160***	0.071
	(3.21)	(3.54)	(0.92)	(3.30)	(0.73)
Firm fixed effect	No	No	Yes	No	Yes
Industry fixed effect	No	Yes	No	No	No
Year fixed effect	Yes	Yes	Yes	No	No
Industry-year fixed effects	No	No	No	Yes	Yes
Observations	2,716	2,716	2,716	2,716	2,716
Adjusted R^2	.372	.408	.759	.377	.753

Table 7. An example of exclusion condition violation

This table presents the results of examining the exclusion condition for the instrument variable in Aghion, Van Reenen, and Zingales (2013). We demonstrate these tests using the effect of institutional ownership on innovation after their inclusion in the S&P 500. Appendix C defines the variables. *p < .1; **p < .05; ***p < .01.

A. Instrument variable regression

	(1)	(2)	(3)	(4)	(5)
	Original results			With additional controls	
	Poisson	OLS (first-stage)	OLS (second-stage)	OLS (first-stage)	OLS (second-stage)
Dependent variable	Cites	Institutional ownership	Cites	Institutional ownership	Cites
Institutional ownership	0.007***	-	0.029**	-	-0.036*
	(2.97)		(2.16)		(-1.73)
S&P 500	–	8.872***	-	5.493***	-
		(3.77)		(2.68)	
Stock liquidity	–	–	–	3.559***	0.391***
				(7.20)	(3.84)
Industry patent intensity	–	–	–	13.200	-0.288
				(0.45)	(-0.22)
Industry citation intensity	–	–	–	95.594	9.986
				(0.93)	(1.61)
Industry fixed effects	Yes	Yes	Yes	Yes	Yes
Year fixed effects	Yes	Yes	Yes	Yes	Yes
Observations	6,208	6,208	6,208	6,208	6,208

B. Exclusion condition check

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	1-year window			2-year window			3-year window		
	Pre	Post	*t*-test	Pre	Post	*t*-test	Pre	Post	*t*-test
Stock liquidity	17.92	17.70	0.85	17.81	17.56	1.25	18.03	17.61	2.35**

Internet Appendix
Not for Publication

Figure IA1: Fraction of Deviance Explained
This figure shows the fraction of model deviance explained by the number of variables. It shows that the six variables chosen roughly explain 55% of the deviance while the whole model explains 58%.

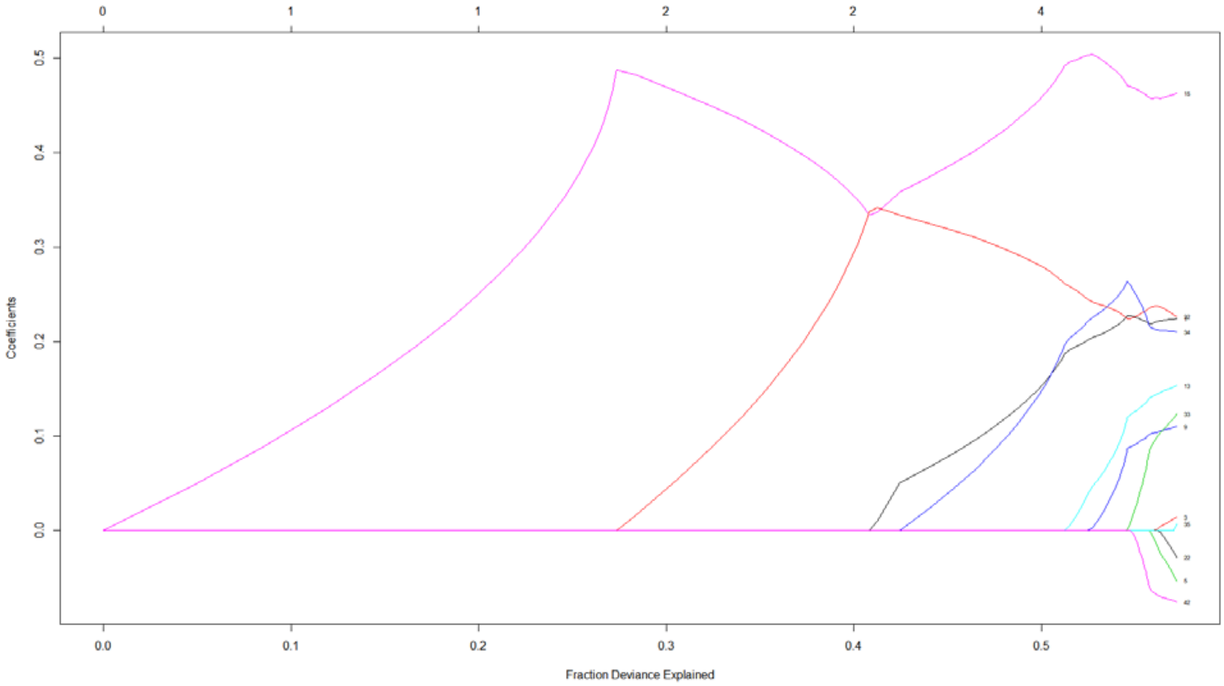

Table IA1 Summary Statistics

This table shows the summary statistics of the sample with firm-year observations from 2001 to 2010. All variables are defined in Appendix C.

	Mean	Median	Std. Dev.	Lower Quartile	Upper Quartile
Innovation Characteristics:					
Patent	3.041	2.833	1.598	1.792	4.060
Citation	0.513	0.000	0.900	0.000	0.693
CEO Characteristics:					
CEO Total Pay	8.381	8.441	1.031	7.725	9.098
CEO Delta	4.390	4.370	1.513	3.435	5.397
CEO Vega	3.518	3.596	1.554	2.551	4.590
CEO Centrality	0.326	-0.042	1.305	-0.334	0.445
CEO Gender	0.019	0.000	0.137	0.000	0.000
CEO Confidence	0.450	0.000	0.498	0.000	1.000
Firm Characteristics:					
Size	7.644	7.561	1.491	6.594	8.788
R&D Stock	0.364	0.262	0.328	0.110	0.555
Tobin’s q	1.273	0.879	1.467	0.494	1.508
Stock Liquidity	19.587	19.499	1.540	18.486	20.679
Firm Age	3.226	3.258	0.688	2.639	3.932
Distance to USPTO	7.081	7.283	1.071	6.462	7.970
Tangibility	0.193	0.161	0.134	0.095	0.259
State Tax	0.294	0.338	0.093	0.298	0.348
ROA	0.128	0.129	0.092	0.087	0.177
Manufacturing	0.637	1.000	0.481	0.000	1.000
Sales Growth	0.254	0.189	0.700	-0.013	0.417
Organization Capital	-0.023	-0.047	0.130	-0.066	-0.018
Capital Structure	0.173	0.155	0.158	0.030	0.264
Governance Characteristics:					
Analyst Following	1.813	1.386	1.872	0.000	3.738
Institutional Ownership	0.373	0.340	0.382	0.000	0.757
Poison Pill	0.654	1.000	0.476	0.000	1.000
Golden Parachutes	0.769	1.000	0.422	1.000	1.000
Family Firm	0.175	0.000	0.380	0.000	0.000
Blockholder	0.905	1.000	0.293	1.000	1.000
Board Size	2.496	2.398	0.566	2.079	2.708
Mergers and Charter Amendments	0.217	0.000	0.412	0.000	0.000
Staggered Board	0.552	1.000	0.497	0.000	1.000
Board Independence	0.793	0.818	0.118	0.750	0.889
Bylaw Amendments Limit	0.476	0.000	0.500	0.000	1.000
Industry Characteristics:					
Industry Patent Intensity	0.190	0.222	0.073	0.129	0.243
Industry Citation Intensity	0.020	0.017	0.013	0.010	0.030
Industry Competition	0.099	0.062	0.097	0.050	0.120
Industry R&D	0.091	0.097	0.058	0.042	0.114
Industry Size	14.555	14.694	0.928	14.268	15.185

Table IA2 A Rolling Window Approach with Adaptive Lasso

This table presents the variable selection results on multiple managerial, governance, firm, and industry factors on patents among patenting firms. We apply the adaptive Lasso process on a two-year rolling window basis. All variables are defined in Appendix C. "-" denotes variables that are not included in the selection process due to design or data availability. The bolded variables indicate ones that are chosen by adaptive Lasso.

Panel A: Patents

Sample:	*2001-2010*	*1992-2010*	*2001-2010*	*1992-2010*
	Frequency Chosen			
	Without R&D Stock		*With R&D Stock*	
Variables	**Patent**			
Managerial Characteristics:				
CEO Centrality	**8/9**	-	**9/9**	-
CEO Confidence	0/9	1/18	1/9	0/18
CEO Vega	1/9	4/18	3/9	0/18
CEO Delta	1/9	1/18	1/9	0/18
CEO Total Pay	1/9	0/18	1/9	1/18
CEO Gender	0/9	0/18	0/9	0/18
Firm Characteristics:				
Size	**9/9**	**18/18**	**9/9**	**18/18**
R&D Stock	-	-	**9/9**	**15/18**
Tobin's q	0/9	0/18	0/9	0/18
Stock Liquidity	**9/9**	**17/18**	**9/9**	**14/18**
Firm Age	0/9	0/18	0/9	0/18
Distance to USPTO	1/9	1/18	1/9	1/18
Tangibility	0/9	0/18	0/9	1/18
Manufacturing	1/9	1/18	4/9	4/18
State Tax	0/9	0/18	0/9	1/18
ROA	0/9	2/18	0/9	1/18
Capital Structure	0/9	0/18	0/9	0/18
Sales Growth	1/9	1/18	2/9	0/18
Organizational Capital	0/9	0/18	1/9	0/18
Corporate Governance:				
Analyst Following	5/9	8/18	**6/9**	9/18
Poison Pill	0/9	1/18	1/9	2/18
Blockholder	0/9	1/18	0/9	1/18
Institutional Ownership	0/9	1/18	0/9	0/18
Board Size	0/9	-	1/9	-
Golden Parachutes	0/9	0/18	0/9	0/18
Board Independence	0/9	-	0/9	-
Mergers and Charter	0/9	0/18	0/9	0/18
Staggered Board	0/9	0/18	0/9	0/18
Bylaw Amendments Limit	0/9	2/18	0/9	1/18
Family Firm	0/9	-	0/9	-
Industry Characteristics:				
Industry Citation	**8/9**	9/18	**9/9**	9/18
Industry Patent Intensity	**6/9**	11/18	3/9	**12/18**
Industry Size	1/9	5/18	3/9	6/18
Industry R&D	1/9	12/18	2/9	6/18
Industry Competition	0/9	0/18	0/9	0/18

Panel B: Citations

Sample:	*2001-2010*	*1992-2010*	*2001-2010*	*1992-2010*
	Frequency Chosen			
	Without R&D Stock		*With R&D Stock*	
Variables	**Citation**			
Managerial				
CEO Centrality	3/9	-	3/9	-
CEO Confidence	0/9	1/18	0/9	0/18
CEO Vega	1/9	0/18	0/9	0/18
CEO Delta	0/9	0/18	0/9	2/18
CEO Total Pay	0/9	0/18	0/9	0/18
CEO Gender	0/9	0/18	0/9	1/18
Firm Characteristics:				
Size	5/9	**14/18**	5/9	**16/18**
R&D Stock	-	-	2/9	4/18
Tobin's q	0/9	0/18	0/9	1/18
Stock Liquidity	**9/9**	**15/18**	**9/9**	**15/18**
Firm Age	0/9	0/18	0/9	0/18
Distance to USPTO	0/9	0/18	0/9	0/18
Tangibility	0/9	0/18	0/9	0/18
Manufacturing	0/9	0/18	0/9	2/18
State Tax	0/9	0/18	0/9	0/18
ROA	0/9	0/18	0/9	0/18
Capital Structure	0/9	0/18	0/9	0/18
Sales Growth	0/9	0/18	0/9	0/18
Organizational Capital	0/9	0/18	0/9	0/18
Corporate Governance:				
Analyst Following	0/9	3/18	0/9	3/18
Poison Pill	0/9	0/18	0/9	2/18
Blockholder	2/9	1/18	2/9	1/18
Institutional	0/9	0/18	0/9	1/18
Board Size	0/9	-	0/9	-
Golden Parachutes	0/9	0/18	0/9	0/18
Board Independence	0/9	-	0/9	-
Mergers and Charter	0/9	0/18	0/9	0/18
Staggered Board	0/9	0/18	0/9	1/18
Bylaw Amendments	0/9	0/18	0/9	2/18
Family Firm	0/9	-	0/9	-
Industry Characteristics:				
Industry Citation	**9/9**	**12/18**	**9/9**	**14/18**
Industry Patent	1/9	1/18	1/9	1/18
Industry Size	0/9	1/18	0/9	1/18
Industry R&D	1/9	3/18	1/9	4/18
Industry Competition	0/9	0/18	0/9	0/18

Table IA3 Variable Selection: A Rolling Window Approach via Stepwise Procedure

This table presents the results of multiple managerial, governance, firm, and industry factors on innovation among patenting firms. We apply the stepwise process on a two-year rolling window basis. All the variables are defined in Appendix C. The Huber-White Sandwich estimator is clustered at the firm level. The bolded variables indicate ones that are chosen by stepwise process.

Sample:	*2001-2010*	*1992-2010*	*2001-2010*	*1992-2010*
	Frequency Chosen			
Variables:	**Patent**		**Citation**	
Managerial Characteristics:				
CEO Centrality	4/9	-	0/9	-
CEO Vega	2/9	2/18	1/9	2/18
CEO Confidence	0/9	3/18	1/9	2/18
CEO Total Pay	0/9	0/18	0/9	0/18
CEO Delta	2/9	1/18	0/9	1/18
CEO Gender	0/9	4/18	0/9	1/18
Firm Characteristics:				
Size	**9/9**	**18/18**	**6/9**	**18/18**
R&D Stock	**8/9**	**17/18**	3/9	9/18
Tobin's q	0/9	0/18	0/9	0/19
Stock Liquidity	**9/9**	**12/18**	**9/9**	**14/18**
Firm Age	0/9	0/18	0/9	0/18
Distance to USPTO	3/9	1/18	0/9	1/18
Tangibility	0/9	4/18	0/9	0/18
State Tax	0/9	2/18	0/9	1/18
ROA	0/9	3/18	0/9	2/18
Manufacturing	1/9	7/18	0/9	2/18
Sales Growth	1/9	1/18	0/9	0/18
Organizational Capital	0/9	0/18	0/9	0/18
Capital Structure	0/9	0/18	0/9	2/18
Corporate Governance:				
Analyst Following	4/9	6/18	0/9	2/18
Institutional Ownership	0/9	3/18	0/9	2/18
Poison Pill	2/9	3/18	2/9	5/18
Golden Parachutes	0/9	0/18	0/9	0/18
Family Firm	0/9	-	0/9	-
Blockholder	0/9	1/18	0/9	2/18
Board Size	0/9	-	0/9	-
Mergers and Charter Amendments	0/9	0/18	0/9	2/18
Staggered Board	0/9	0/18	0/9	3/18
Board Independence	0/9	-	0/9	-
Bylaw Amendments Limit	0/9	0/18	1/9	5/18
Industry Characteristics:				
Industry Citation Intensity	**8/9**	**10/18**	**9/9**	**13/18**
Industry Size	5/9	8/18	0/9	4/18
Industry R&D	4/9	10/18	0/9	4/18
Industry Patent Intensity	4/9	6/18	0/9	2/18
Industry Competition	0/9	0/18	0/9	0/18

Table IA4 Correlation Table

This table shows the correlation matrix between the 35 variables based on the sample of 2001-2010.

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1	CEO Total Pay	1.00																	
2	CEO Delta	0.42	1.00																
3	CEO Vega	0.46	0.78	1.00															
4	CEO Centrality	0.19	0.18	0.19	1.00														
5	Firm Age	-0.01	0.02	-0.01	-0.09	1.00													
6	CEO Gender	0.04	0.03	0.01	0.06	-0.08	1.00												
7	CEO Confidence	-0.03	0.17	0.08	-0.06	0.19	-0.08	1.00											
8	Size	0.62	0.43	0.42	0.22	0.06	0.03	-0.17	1.00										
9	R&D Stock	-0.11	-0.02	0.01	0.03	-0.13	0.00	0.08	-0.40	1.00									
10	Tobin’s q	0.12	0.38	0.21	0.02	-0.06	0.02	0.22	-0.01	0.17	1.00								
11	Stock Liquidity	0.49	0.38	0.38	0.28	-0.12	0.05	-0.09	0.54	0.21	0.26	1.00							
12	Distance to USPTO	-0.06	-0.08	-0.09	-0.03	0.06	0.01	-0.04	0.07	-0.22	-0.11	-0.27	1.00						
13	Tangibility	-0.09	-0.13	-0.09	-0.08	0.12	-0.06	-0.13	0.11	-0.34	-0.27	-0.20	0.15	1.00					
14	State Tax	0.24	0.30	0.26	0.02	0.06	0.02	0.13	0.39	-0.34	0.17	-0.03	0.10	0.00	1.00				
15	ROA	0.23	0.30	0.23	0.02	0.02	-0.04	0.10	0.40	-0.29	0.30	0.06	0.10	0.17	0.55	1.00			
16	Manufacturing	-0.16	-0.13	-0.10	-0.12	0.04	-0.06	0.11	-0.16	0.04	-0.05	-0.08	-0.01	-0.17	-0.09	-0.12	1.00		
17	Sales Growth	-0.03	0.10	0.07	-0.02	-0.02	0.02	0.08	-0.19	0.19	0.22	-0.01	-0.05	-0.09	-0.08	-0.22	-0.05	1.00	
18	Organization Capital	-0.12	-0.14	-0.14	0.02	-0.01	0.06	-0.03	-0.12	0.05	-0.11	-0.12	0.00	0.05	-0.20	-0.05	-0.01	-0.07	1.00
19	Capital Structure	-0.01	-0.07	-0.06	-0.04	0.11	-0.03	-0.07	0.08	-0.18	-0.13	-0.07	0.11	0.14	-0.15	-0.08	-0.09	0.05	0.09

		19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34
19	Capital Structure	1.00															
20	Analyst Following	0.10	1.00														
21	Institutional Ownership	0.10	0.92	1.00													
22	Poison Pill	-0.02	-0.02	0.05	1.00												
23	Golden Parachutes	0.13	0.07	0.14	0.23	1.00											
24	Family Firm	-0.08	-0.14	-0.13	-0.18	-0.27	1.00										
25	Blockholder	0.02	-0.11	0.01	0.07	0.16	0.00	1.00									
26	Board Size	-0.02	0.14	0.05	-0.04	-0.01	-0.01	-0.10	1.00								
27	Mergers and Charter Amendments	0.02	0.11	0.11	0.04	0.08	0.02	0.04	-0.01	1.00							
28	Staggered Board	0.02	0.04	0.09	0.25	0.18	-0.08	0.12	0.02	0.19	1.00						
29	Board Independence	0.03	0.17	0.20	0.06	0.25	-0.25	0.05	0.03	0.06	0.03	1.00					
30	Bylaw Amendments Limit	0.00	0.04	0.05	-0.05	0.14	-0.07	0.10	0.04	0.22	0.10	0.20	1.00				
31	Industry Patent Intensity	-0.12	-0.11	-0.09	0.10	0.03	-0.03	0.00	-0.09	-0.01	0.01	-0.06	-0.10	1.00			
32	Industry Citation Intensity	-0.14	-0.12	-0.13	0.05	-0.08	-0.01	-0.04	-0.08	-0.08	-0.10	-0.10	-0.13	0.65	1.00		
33	Industry Competition	0.11	0.14	0.14	-0.07	0.01	-0.06	-0.03	0.00	-0.01	-0.06	0.03	0.03	-0.22	-0.13	1.00	
34	Industry R&D	-0.09	-0.08	-0.06	0.09	0.03	0.00	-0.03	-0.01	-0.01	0.03	0.00	0.04	0.60	0.13	-0.34	1.00
35	Industry Size	-0.12	0.02	-0.02	0.07	-0.03	-0.07	-0.09	0.08	-0.08	-0.03	0.08	0.06	0.31	0.29	-0.18	0.53

Table IA5 Including Non-Patenting Firms

This table presents adaptive Lasso variable selection results when we include firms with zero patents.

	Frequency Chosen	
Variables	**Patent**	**Citation**
Managerial Characteristics:		
CEO Centrality	9/9	4/9
CEO Delta	0/9	0/9
CEO Confidence	0/9	0/9
CEO Vega	1/9	0/9
CEO Total Pay	1/9	0/9
CEO Gender	0/9	0/9
Firm Characteristics:		
Size	9/9	6/9
R&D Stock	9/9	2/9
Tobin's q	0/9	0/9
Stock Liquidity	9/9	9/9
Firm Age	0/9	0/9
Distance to USPTO	7/9	0/9
Tangibility	0/9	0/9
State Tax	0/9	0/9
ROA	0/9	0/9
Manufacturing	0/9	0/9
Sales Growth	0/9	0/9
Organizational Capital	0/9	0/9
Capital Structure	0/9	0/9
Corporate Governance:		
Analyst Following	7/9	0/9
Blockholder	1/9	1/9
Bylaw Amendments Limit	2/9	0/9
Institutional Ownership	0/9	0/9
Poison Pill	1/9	0/9
Board Size	0/9	0/9
Golden Parachutes	0/9	0/9
Family Firm	0/9	0/9
Mergers and Charter Amendments	0/9	0/9
Staggered Board	0/9	0/9
Board Independence	0/9	0/9
Industry Characteristics:		
Industry Citation Intensity	9/9	9/9
Industry Patent Intensity	7/9	0/9
Industry Size	4/9	0/9
Industry R&D	0/9	0/9
Industry Competition	0/9	0/9

Table IA6 Variable Selection: R&D Expenditures

This table presents the results of variable selection in two-year rolling windows using R&D as the dependent variable. All the variables are defined in Appendix C.

	Frequency Chosen	
	Stepwise	**Adaptive Lasso**
Variables	**R&D**	
Managerial Characteristics:		
CEO Centrality	1/9	0/9
CEO Confidence	0/9	0/9
CEO Gender	0/9	0/9
CEO Vega	4/9	0/9
CEO Total Pay	0/9	0/9
CEO Delta	1/9	0/9
Firm Characteristics:		
Size	9/9	9/9
Tobin's q	1/9	0/9
Stock Liquidity	9/9	9/9
Firm Age	0/9	0/9
Distance to USPTO	0/9	0/9
Tangibility	9/9	1/9
Manufacturing	1/9	0/9
State Tax	4/9	3/9
ROA	0/9	0/9
Capital Structure	1/9	0/9
Sales Growth	0/9	1/9
Organizational Capital	0/9	0/9
Corporate Governance:		
Analyst Following	0/9	0/9
Poison Pill	0/9	0/9
Blockholder	0/9	0/9
Institutional Ownership	0/9	0/9
Board Size	0/9	0/9
Golden Parachutes	0/9	0/9
Board Independence	0/9	0/9
Mergers and Charter Amendments	0/9	0/9
Staggered Board	0/9	0/9
Bylaw Amendments Limit	0/9	0/9
Family Firm	0/9	0/9
Industry Characteristics:		
Industry Citation Intensity	0/9	0/9
Industry Patent Intensity	8/9	8/9
Industry Size	0/9	0/9
Industry R&D	1/9	1/9
Industry Competition	0/9	0/9