Forthcoming: Journal of Financial and Quantitative Analysis

Deleting Unreported Innovation

Ping-Sheng Koh, David M. Reeb, Elvira Sojli, Wing Wah Tham, and Wendun Wang*

February 25, 2021

ABSTRACT

The absence of observable innovation data for a firm often leads us to exclude or classify these firms as non-innovators. We assess the reliability of six methods for dealing with unreported innovation using several different counterfactuals for firms without reported R&D or patents. These tests reveal that excluding firms without observable innovation or imputing them as zero innovators and including a dummy variable can lead to biased parameter estimates for observed innovation and other explanatory variables. Excluding firms without patents is especially problematic, leading to false-positive results in empirical tests. Our tests suggest using multiple imputation to handle unreported innovation.

* Koh, kohp@essec.edu, ESSEC Business School; Reeb, dmreeb@nus.edu.sg, National University of Singapore and senior fellow of Asian Bureau of Finance and Economic Research; Sojli e.sojli@unsw.edu.au, University of New South Wales, UNOVA Transdisciplinary R&D Lab; Tham, w.tham@unsw.edu.au, University of New South Wales, UNOVA Transdisciplinary R&D Lab, and AGORA Center for Market Design; Wang, wang@ese.eur.nl, Erasmus School of Economics. The simulation code and R&D imputed data are available at www.wendunwang.weebly.com/research.html. An early version of this manuscript circulated under the title "Missing Innovation Around the World." We appreciate discussions and advice from an anonymous referee, Renee Adams, Sumit Agarwal, Benito Arrunda, Richard Boylan, Lora Dimitrova, Bronwyn Hall, Jarrad Harford (the editor) Gilles Hilary, Yael Hochberg, Qin Li, Gustavo Manso, Jiaming Mao, Naci Mocan, Randall Morck, Ivan Png, Wenlan Qian, Amit Seru, Vijay Singal, Noah Stoffman, Xuan Tian and Rosemarie Ziedonis. We are grateful to seminar participants at City University of Hong Kong, Hong Kong Polytechnic University, Louisiana State University, Maastricht University, the National University of Singapore, Rotterdam School of Management, Temple University, University of Queensland, University of South Carolina, University of Toronto, Virginia Tech, Wilfrid Laurier University, University of Technology Sydney, and the 2020 PCOB Workshop on Valuation & Accounting for Intangible Assets. Koh acknowledges financial support from the ESSEC Foundation and CY Initiative of Excellence.

I. Introduction

A well-known problem in cross-disciplinary research on corporate innovation is the absence of innovation data for the majority of firms (Anton and Yao (2004), Koh, Reeb, and Zhao (2018)). For instance, over 50% of US firms do not report R&D spending and among firms with positive R&D for over a decade, roughly 60% of them do not obtain patents. Potential explanations for not reporting R&D or obtaining patents include negligible innovation inputs, unsuccessful innovation projects, or attempts to keep the innovation information secret (Png (2017)). Recent studies develop additional measures of corporate innovation, ranging from the textual analysis of financial analyst reports (Bellstam, Cookson, and Bhagat (2020)) to firm disclosures of new products (Mukherjee, Singh, and Zaldokas (2017)). These new measures provide meaningful approaches to capture different aspects of innovation but also suffer from the same concerns about missing data noted by Lerner and Seru (2017). The two most common approaches to dealing with the absence of innovation data are to exclude firms without R&D or patents (e.g., Hombert and Matray (2018), Branstetter, Drev, and Kwon (2019)) or classify these firms as zero innovators and include a dummy variable (e.g., Masulis and Zhang (2019), Koch, Panayides, and Thomas (2021)). Are these the right approaches to handling the absence of corporate innovation data?

We investigate the efficacy of different methods for handling unreported innovation data in research that focuses on innovation as either an explanatory variable of interest or includes it as a control variable. We compare six approaches that researchers could potentially use to deal with the absence of innovation data: listwise deleting (discarding) firms without R&D or patents, deterministic imputation with either zero or industry average, inverse probability weighting, Heckman selection, and multiple imputation (MI). MI is arguably the least common method in corporate finance and relies on estimating the missing variable of interest using other observable covariates and explicitly adjusting for imputation uncertainty.

Our preliminary analysis reveals that the absence of innovation data for a firm is predictable with known determinants of innovation and other corporate outcomes of interest, rejecting the hypothesis that unobservable innovation is *missing completely at random*.[1] These results raise the concern that the commonly used methods to handle unreported innovation could lead to biased parameter estimates due to the non-representativeness of the population under study (deletion) or distortions of the variance-covariance matrix (deterministic imputation).

Our primary analysis starts with an examination of firms without reported R&D spending. First, we use data on US firms that did not report R&D spending in a particular period but reported this historical amount in subsequent financial statements. Firms that initiate reporting of R&D are required to report their R&D expenditures for prior years. To the empiricist these firms do not appear to have engaged in R&D activity in the years R&D spending was not reported, creating a natural laboratory to evaluate different methods of handling unreported corporate innovation. We denote this newly reported R&D spending in future financial statements as "*recovered R&D*." Using recovered R&D as a baseline, we compare the observed/counterfactual R&D with replacing these firms' missing R&D with zero, the industry average, and MI. Our tests show that the R&D in firms with unreported R&D significantly differs from zero R&D firms, the average industry R&D, and positive R&D firms in aggregate. Notably, we find that, on average, MI provides estimates of the missing R&D that are qualitatively similar to their actual R&D reported in subsequent financial statements.

The absence of reported R&D in a firm could arise from a lack of innovation or as a disclosure choice. One potential issue with using recovered R&D is that these firms all engage in

[1] Terminology in statistics differentiates between three types of missing data. *Missing completely at random* (MCAR) occurs when neither observables nor unobservables predict missing observations. *Missing at random* (MAR) occurs when observables can predict missing observations and *missing not at random* (MNAR) occurs when missing observations are related to observable and unobservable data (see the Internet Appendix for details).

innovation. Consequently, we use an alternative counterfactual group that allows for both innovative and non-innovative firms. Specifically, we use the text-based ranking of corporate innovation in the S&P 500 firms from Bellstam et al. (2020). Analysts' discussions that describe corporate innovation arguably arise in firms that engage in R&D, while firms without such discussions are likely those without meaningful innovative activity. Again, we find that MI is the best solution to handle missing R&D. Figure 1 shows the dramatically different innovation rankings of S&P 500 firms when using reported R&D spending and the textual analysis of Bellstam et al. (2020). The standard practice of classifying firms that fail to report R&D as zero innovators leads to a large clumping of firms at the bottom of the distribution. In contrast, the textual analysis approach, while only available for S&P500 firms, provides a more complete distribution of corporate innovation. Similarly, the MI approach to handling unreported R&D also mitigates the clumping problem inherent with deleting these firms or classifying them as zero innovators.

Next, we focus on firms without patents. Patents from the United States Patent Office (USPTO) are a commonly used metric to measure corporate innovation. To evaluate different methods of handling firms without patents or patent citations, we use new product announcements as the ground truth. We assume that firms have successful innovation when they make new product announcements, especially for major new product announcements. Yet, firms without patents tend to have fewer new product announcements than USPTO patenting firms. Consequently, classifying these firms as zero innovators, imputing them with the industry average number of patents, or deleting firms without patents is problematic. The MI estimates appear to place these non-patenting firms into the appropriate innovation categories.

Our second test on the absence of patents focuses on firms without USPTO patents. Among US firms, 69% of positive R&D firms never file for patents using USPTO data, while only 43% never file patent applications using the 30 global patent offices. This 26% wedge in unobserved patents for studies using USPTO patents provides another opportunity to examine different methods of handling unobserved innovation. In this particular case, the nature of the missingness likely differs from trade-secret based reasons, limiting external validity. Yet, a benefit of focusing on this 26% wedge is that we can directly evaluate different methods of handling unobserved innovation in a large number of empirical studies. Strikingly, we again find that MI provides much closer estimates for the patents unobserved by the empirical researcher relying on USPTO patents than in other commonly used replacement methods.

To further mitigate external validity concerns from the different counterfactuals for R&D and patents, we undertake three simulation studies. In the first two simulation analyses, we use the empirical distribution of US Compustat and patent data to evaluate the impact of differing levels of missingness of an innovation variable. The third simulation analysis uses clearly-specified data generating processes, allowing us to gauge the impact of unreported innovation in a controlled, cross-sectional setting. In these simulation exercises, we evaluate the six different approaches noted previously to handle unreported innovation. Our simulation analyses rely on two evaluation criteria: the bias (expressed as a proportion of the benchmark coefficient) and the root mean squared error (RMSE) of the regression coefficient estimates.

In these simulations, we find that deleting or excluding firms without reported R&D or patents leads to biased coefficient estimates for both innovation and any control variables correlated with innovation. Rather than providing a conservative approach, the exclusion or deletion of firms without reported R&D or patents is one of the worst methods for handling unreported innovation. For instance, if R&D is *missing at random*, then the average bias from excluding firms without reported R&D is almost 10 times greater than that found using MI. Also, we find that the RMSEs of the common methods in economics: deleting the firms without observed innovation or using a Heckman selection model, are very large in comparison to MI in our simulation exercises (i.e., 70% larger). Varying the makeup of unreported innovation data, ranging from firms that simply do not engage in innovation to firms that seek to hide their R&D spending, gives similar inferences. We also find that the bias and RMSE from deleting firms without observable innovation dramatically increase at higher rates of missingness. These simulation results suggest the high rate of absent innovation in patents, relative to R&D spending, makes excluding firms without patents especially problematic.

To illustrate the economic relevance of inference problems with common approaches to handling unreported innovation, we replicate an influential finance study that uses R&D spending. Fama and French (2002) test the empirical predictions of the pecking order and trade-off models of capital structure and classify firms without reported R&D expenditure as zero R&D firms. We find that the coefficient estimates and standard errors for R&D and capital structure are significantly different when using MI to account for unreported innovation. We cannot state which results are correct, but we do note that different approaches to handling unreported innovation give opposing results, suggesting this is an important consideration in research design.

The nature of unreported innovation is unknown to the researcher. Implicitly, researchers deciding how to handle missing innovation data are making assumptions about whether missingness can or cannot be predicted by observables. We recommend that researchers provide some basic statistics for the degree or magnitude of the missing innovation data in their sample and how it relates to their key variable(s) of interest. Instead of simply excluding firms without patents, reported R&D, financial analysts' coverage, or new product announcements, we should attempt to adjust for the non-randomness in unreported innovation. Both deterministic imputation and listwise deletion of firms with unreported innovation can provide biased coefficient estimates, making it difficult to evaluate how well one biased approach can provide a robustness test for another biased approach. Our analysis suggests using MI estimates when confronted with unreported innovation data.[2]

This study contributes to the burgeoning work on the econometric challenges faced by researchers in finance (e.g., Petersen (2009), Thompson (2011)). Koh and Reeb (2015) compare two deterministic imputation methods for handling missing R&D but are silent on the relative biases of these two methods. They do not evaluate excluding firms without reported R&D, multiple imputation, inverse probability weighting, or Heckman selection models. Our analysis shows that deterministic imputation solutions can lead to biased estimates and standard errors for both unreported R&D and patents, as well as other control variables. Importantly, our paper questions the foundations for deleting firms with unreported innovation and the impact of using deterministic imputation models that classify these firms as zero innovation firms and including a dummy variable. In addition, we show that the problem of unobservable innovation is arguably more pronounced in studies that use patent-based metrics to measure innovation instead of R&D expenditures, because of the higher rate of missingness in patents.

Finally, studies that use R&D or patents as control variables also suffer from this missing data bias. Best practices for dealing with unreported R&D and patents depend on the source or type of unreported innovation data. If country, industry, or firm characteristics predict unreported R&D or missing patents (see Lerner and Seru (2017)), then our analysis suggests that using multiple imputation provides the most reasonable solution. Surprisingly, and across a wide variety of specifications and approaches, we find that both Heckman selection and inverse probability

---

[2] Our simulations use the common MI approach of Markov Chain Monte Carlo (MCMC), which assumes the underlying data follows a multivariate normal distribution. Predictive Mean Matching (PMM) is an alternative MI approach that could be appropriate for R&D (Horton and Lipsitz 2001). In our simluations, the PMM method performs similarly to the standard MCMC approach. We provide both average MCMC and PMM imputed R&D values for the Compustat universe of US firms. The code for multiple imputation and the imputed R&D values are available at https://bit.ly/3qyd00X.

weighting rarely provide the best approaches to handling unreported innovation in our samples. We provide our code for researchers interested in performing simulations on absent innovation using their unique data and set of guidelines for unreported innovation in the conclusion.

II. Handling Unreported Innovation Data

There are numerous possible reasons for why we observe firms without innovation data. Unfortunately, the nature of missingness cannot be positively identified by examining the observable data. Missing data causes two problems: bias in the parameter estimates and loss in efficiency (Rubin (1976)). Bias stems from the non-representativeness of the population under study. Loss of efficiency arises because information loss is a direct consequence of missing data, i.e., smaller samples. Research in statistics has long recognized and studied the broad class of missing data problems (e.g., Robins and Wang (2000)), while research in machine learning also focuses on training and testing data that suffers from missing observations (e.g., Grangier and Melvin (2010)). Yet, many of these techniques and methods are relatively unused in research on corporate innovation. To illustrate the problem, we consider a simple case with one explanatory variable that contains missing observations. Let y_i be the dependent variable and z_i be the innovation variable with missingness. We have the linear relation:

(1) $$y_i = \alpha + \theta z_i + \varepsilon_i, \quad i = 1, \dots, N.$$

Let s_i be a selection indicator where $s_i = 1$, when z_i is not missing and firm i is included in the regression. Otherwise, when $s_i = 0$ firm i is deleted from the data. The validity of solutions to this problem depends on the nature of missingness.

A. Common Approaches to Unreported Innovation

One common approach to missing innovation data is to delete or exclude firms without R&D spending or patents. Listwise deletion uses only a subsample of observations, deleting firms or firm-years that contain missing values in the z-variable, in equation (1). This leads to estimating the following regression using a subsample of the data:

(2) $$y_i = s_i\alpha + \theta s_i z_i + s_i\varepsilon_i,$$

where $s_i z_i$ is now the explanatory variable and $s_i\varepsilon_i$ is the error term. If the selection is driven by observed or even unobserved variables, then $E(\varepsilon_i|z_i, s_i) \neq 0$ in general because ε_i can be correlated with s_i even if one controls for z_i, leading to biased estimates produced by deletion. Another common approach to dealing with missing innovation data is to impute the missing observations using various methods, and then treat the resulting data as given for further analysis.

B. Alternative Methods for Handling Unreported Innovation

Two other approaches to handling missing data are also viable candidates for unreported innovation. Inverse probability weighting (IPW) relies on assigning different weights to observed points depending on their probability of being observed. As this probability is unknown for unreported innovation, we can estimate it using binary choice or nonparametric models. A second approach is multiple imputation. MI is essentially an iterative version of stochastic imputation, which aims at explicitly modeling the uncertainty/variability ignored by the deterministic imputation procedures. Instead of replacing with a single value (unrelated to other covariates/observed data), MI uses the (joint) distribution of the observed data to estimate the parameters of interest multiple times to capture the uncertainty/variability in the imputation procedure (see Internet Appendix for more details on these methods).

III. The Severity of Unreported Innovation

A. Data and Sample

The sample of patents is derived from the EPO-OECD-PATSTAT database. This database, also known as the EPO Worldwide Patent Statistical Database, contains a snapshot of the European Patent Office (EPO) master documentation database with worldwide coverage. It has more than 20 tables with bibliographic data, citations, and family links for about 70 million applications from more than 90 countries, including the EPO and the USPTO.

Our sample selection begins with the October 2013 version of the PATSTAT data.[3] It contains 44,730,405 observations, including patentees who are individuals, government institutions/universities, and companies for the sample period of 1999–2012. Our analysis relies on the registered names on the original patent applications, rather than the ultimate patent owners, to better capture the entities that performed the innovation activities. We merge the patent data with all publicly listed firms in the Compustat North America and Compustat Global database for 32 countries. Our matching algorithm consists of two main steps. First, we standardize patent assignee names and firm names, focusing on unifying suffixes and dampening the non-informative parts of firm names. Second, we apply multiple fuzzy string-matching techniques to identify the firm, if any, to which each patent belongs. We randomly select firms to manually confirm the matching of patents to firms.

We focus on countries with at least 100 publicly listed firms (excluding Hungary, Iceland, and Ireland). Thus, our primary sample contains 29 countries: Australia, Austria, Belgium, Brazil, Canada, China, Denmark, Finland, France, Germany, Greece, Hong Kong, India, Israel, Italy, Japan, Korea, Malaysia, the Netherlands, New Zealand, Norway, Singapore, South Africa, Spain,

[3] Our patent sample ends in 2012, because patents post-2012 may be affected by the truncation bias for citations.

Sweden, Switzerland, Taiwan, the UK, and the US. There are 30 patent offices in the sample because the EPO is a separate entity from each European country's patent office; European firms sometimes patent in their home patent office and other times with the EPO. Our baseline sample includes 333,920 firm-year observations and 37,272 unique firms, of which 5,374 are cross-listed firms. All accounting variables are from Compustat (North America and Global) and are defined in the Appendix.

Panel A in Table 1 reports the basic descriptive statistics of our sample firms. Only 35% of the observations in our sample report any information on R&D. Of those reporting R&D expenditures (118,264), 93% report positive R&D with an average R&D expenditure of 8% of their total assets. 7% of firms report zero R&D. The 75th percentile of R&D reporting firms has R&D that equates to roughly 6% of total assets. In addition, the sample firms invested an average of 6% of total assets in capital expenditure. Firms have an average of nine patent applications, four patents granted, and 23 citations over the sample period. On average, firms are profitable with an average ROA of 1% (median of 5%) and are highly levered, with a median leverage of 52%. In our analysis, we focus on patent applications, as these capture the R&D activity happening around the firm, but find similar results using patents granted.

We apply the Adaptive Lasso procedure to identify any additional variables to those in Table 1 that may be relevant for the prediction of unreported R&D. Following the innovation-theory based work of Reeb and Zhao (2020), we use a tenfold cross-validation method and choose the two tuning parameters (lambda and gamma) to minimize the mean square error in the out-of-sample testing (Hui, Warton, and Foster (2015)) to a set of 37 variables.[4] Similar to Reeb and Zhao

[4] One benefit of Adaptive Lasso is that it uses the preliminary coefficient estimate of a covariate as the adaptive weight and as a penalty for this covariate, which mitigates concerns about whether the variables are in the same range. For the Double Lasso method introduced in Section VI, our approach normalizes all covariates to mean 0 and a variance of 1.

(2020), the Lasso approach identifies total assets, stock liquidity, and industry patent intensity as relevant predictive variables for unreported R&D and total assets, stock liquidity, industry patent intensity, and R&D stock as relevant predictive variables for unreported patents.

B. Univariate Comparison

To better gauge the severity of the missing data problem and the potential impact of deleting firms without reported innovation, we compare samples with and without these firms. Specifically, we evaluate the effects of deleting innovation measures by comparing two approaches: deleting all observations without both R&D and patent applications and deleting all observations without either R&D or patent applications. The first group comprises only observations that have reported both R&D expenditures *and* patent information. Our benchmark group comprises of observations that have *either* reported R&D expenditures *or* patent applications with any of the 30 patent offices, R&D, and patents. We conduct a univariate comparison under different samples.

Panel B in Table 1 reports the univariate characteristics of the full sample (column 1), the sample that reports only R&D (column 2), the sample that reports only patents (column 3), and the sample with both R&D and patents (column 4). Panel B shows that deleting missing innovation data substantially reduces the number of observations and paints a very different picture in comparison to the full sample. The samples with reported R&D or patents have less than a third of the observations of the full sample. These subsamples have higher total assets than the full sample, while the rest of the variables are significantly lower (columns 5 and 6). The R&D and the patent-only sample consist of 26,273 observations. Total assets, Tobin's Q, and sales growth are larger than those in the full sample, while the rest of the variables are smaller (column 7). It is worth pointing out that ROA decreases by 400% from the full sample to the R&D and patenting sample.

These results indicate that R&D and patenting are at least not *missing completely at random* and may depend on observables.

C. Tests of the Deletion Assumptions

Next, we evaluate the validity of the assumptions underlying the common practices of deleting missing innovation. An example of an MCAR process (when deletion of observations with innovation is valid) is one in which firms decide whether to report innovation based on coin flips. We test the underlying assumption behind deletion, where the estimates of interest are consistent, in two ways. First, we use the MCAR test of Little (1988) to investigate the missing-value pattern. Second, we study if unreported innovation is more prevalent across firms with certain firm characteristics, by examining the predictability of unreported innovation through regression analysis.

Whether missing data are MCAR can be tested by investigating if there are significant differences between the means of different missing-value patterns across variables of interest. This is formalized by Little (1988), who implements the Chi-square test of MCAR for multivariate quantitative data. The test statistic takes a form similar to the likelihood-ratio statistic for multivariate normal data and is asymptotically χ^2 distributed under the null hypothesis that there are no differences between the means of different missing-value patterns. Rejection of the null provides evidence that the missing data are not MCAR.

Table 2 reports Little's MCAR test statistics for unreported R&D and the number of patents with different covariates. All p-values for various specifications are smaller than 0.01 with the χ^2 statistic ranging between 297 and 22,889 for both the global and US sample, rejecting the null hypothesis that unreported R&D and non-patenting firms are unpredictable. The test provides strong evidence that unreported innovation is not MCAR.

D. Predicting Unreported Innovation

Next, we investigate whether the observed variation in unreported R&D and patents at the firm-year level is systematically related to firm characteristics. We assess the existence of identifiable patterns in unreported innovation by conducting a regression analysis of unreported innovation on observable firm characteristics at the firm-year level for international and US firms. The chosen characteristics are based on innovation theory and the Lasso approach described above. Note that these tests do not seek to establish causality, but rather to emphasize association and predictability in the variation in unreported innovation to shed light on the nature of missingness in innovation. We estimate a panel regression model with year, industry, and country fixed effects, separately for unreported R&D and patents.[5]

In Table 3, the dependent variable is unreported R&D, which is equal to 1 when R&D is not reported and zero otherwise. For all firms (columns 1-4), firm characteristics with country, industry, and year fixed effects explain up to 38% of the variation in unreported R&D (column 3 of Table 3). In contrast, including only year or year and industry fixed effects explains 14-23% of the variation. Firm characteristics with firm and year fixed effects explain 81% of the variation in unreported R&D (column 4). Unreported R&D increases at the firm level with PPE, ROA, and SALES_GROWTH, while it decreases with ln(TOTAL_ASSETS), STOCK_LIQUIDITY, and PATENT_INTENSITY. For US firms (columns 5-7), industry and year fixed effects explain 53% of the variation in unreported R&D, while firm and year fixed effects explain 93% of this variation. Unreported R&D increases with PPE, ROA, and LEVERAGE, while it decreases with

[5] We report the results using least square estimation because it allows us to easily incorporate multi-level fixed effects. We also estimate the determinants of unreported innovation using binary choice models, logit and probit, with various specifications of fixed effects. The results remain qualitatively the same and are available upon request.

PATENT_INTENSITY for US firms. Thus, the results are inconsistent with absent R&D arising *missing completely at random.*

Table 4 presents the prediction results for firms without patents, which is set equal to 1 when a firm does not file for a USPTO patent in a given year, and zero otherwise. USPTO patents are the benchmark in this analysis, as they are widely used to measure both US and non-US firm innovation. Firm, industry, and country characteristics explain up to 26% of the variation in unreported patents (columns 1-3). The absence of patents increases at the firm level with ROA, and decrease with ln(TOTAL_ASSETS) and STOCK_LIQUIDITY. Focusing on just the subset of US firms, firm and industry characteristics explain a substantial amount of the variation in absent patents (32% to 77%; columns 6-7). Unreported patents increase with ROA while decreasing with ln(TOTAL_ASSETS) and STOCK_LIQUIDITY for US firms. Collectively, the evidence in this section indicates a significant correlation between the absence of patents and firm-specific factors. Again, the results are inconsistent with innovation *missing completely at random*. Taken together these results suggest that unreported R&D and Patents are *missing not at random*.[6]

IV. Unique Setting to Investigate Imputing Unobserved R&D

A. The Setting

The main challenge to imputation approaches relates to how close the imputed estimates are to the true yet unobservable values. In this section, we adopt an innovative approach that partially overcomes the unobservable true value problem to examine the efficacy of the various common methods used to handle missing R&D in studies of corporate innovation.

[6] Results in Table IA1 in the Internet Appendix show that focusing only on the Lasso inferred variables does not qualitatively affect the inference.

Except for the first year of operation, firms are required to disclose their prior-year financial numbers on their financial statements to enable across-time comparisons by users of general-purpose financial statements (FASB ASC 205-10-45). This enables us to identify a unique (albeit narrow) setting where we can "recover" the previously unreported R&D expenditure information that serves as the true (yet previously unobservable) value. Specifically, when firms switch from not reporting to reporting R&D expenditures, they are required to report both the current year and prior year's R&D expenditures. In this instance, we can identify the previously unreported R&D expenditures. Our unique setting is thus especially appropriate to investigate how close the imputed estimates from various imputation methods are to the "recovered" true values.

Using the sample of US firms for the period 1992 to 2016, we identify firms that switch between reporting and not reporting R&D expenditure. We find 738 unique firms that switch between reporting and not reporting R&D. We then manually collect data from the annual reports (10Ks) of these firms on their prior years' R&D expenditure, collecting information on the reported R&D in the year of the switch and up to two years before the switch in reporting. We restrict our analysis to firms without any major corporate events (e.g., mergers and acquisitions) over the past two years that would have altered the underlying business operations of the firm (e.g., Bena and Li, 2014). We denote these as "recovered R&D" firms. This provides us with 763 observations for the switch year (some firms switch between reporting and not reporting R&D more than once during our sample period) and 1,032 recovered observations (some firms report amounts for one year, while others for two years before the switch).

B. Comparing Recovered R&D Firms to Zero

We begin our analysis by comparing the characteristics of recovered R&D firms with zero R&D firms and positive R&D firms. Panel A of Table 5 presents the results. Panel A shows that the average R&D investment for the switching firms with recovered R&D is $6.69 million a year, and compares the "recovered R&D" firms to firms that report zero R&D and positive R&D for the comparative years (t-1 and t-2). The R&D expenditure and R&D value of recovered firms are statistically different from the zero R&D firms. In addition, the recovered firms differ from zero R&D firms across several different dimensions like total assets, PPE, and leverage.

[TABLE 5 about here]

The R&D expenditure by recovered firms is significantly lower than positive R&D firms, but the R&D expenditure of the two groups are not distinguishable from each other. Recovered R&D firms also differ from positive R&D firms in total assets and PPE. Untabulated multivariate tests provide similar inferences, showing that recovered R&D is predictable by many common firm characteristics. Overall, results in Panel A of Table 5 show that unreported R&D expenditure firms differ from both zero R&D firms and positive R&D firms, suggesting that deleting them or classifying them as zero innovators are problematic solutions. More specifically, if an innovation covariate is correlated with any of the variables predicting recovered R&D firms, then excluding or classifying these firms as zero innovators can lead to biased inferences.

C. Comparing Different Imputation Methods

We test the different imputation techniques using the "recovered R&D" sample as a counterfactual for the true R&D in Panel B of Table 5. In the Compustat data, this recovered R&D appears as missing, and we impute this R&D with zero, with average industry R&D (two-digit SIC), and with multiple imputed R&D. We compare the recovered R&D with the imputed R&D and calculate the difference and related *t-statistics*.

We use two samples for the R&D multiple imputation. First, we base our multiple imputation estimation on the whole US sample for the period 1992 to 2016 (not just the recovered

R&D sample): *MI Full Sample*. Second, we base the multiple imputation only on the sample of recovered R&D firms and firms matched within industry and size quartile: *MI Sub Sample*. Our imputation is based on the three Adaptive Lasso determined variables: total assets, stock liquidity, and industry patent intensity and estimated by industry (two-digit SIC). We use 200 iterations for the imputation in the analysis.

Panel B of Table 5 shows that recovered R&D, for the sample with SIC information, is statistically different from zero, i.e., replacing with zero underestimates the recovered R&D values. In terms of the dollar amount of R&D, the average recovered R&D is $6.91 million, while imputing with the industry average gives an estimate of $77.86 million. The industry average imputed value is over 10 times their actual R&D spending and significantly different from the recovered value. On the other hand, the two MI methods generate an average of $6.36 million and $8.66 million, which are not statistically different from the recovered R&D values. The relatively large variance in the MI values points to the difficulty in using these as exact point estimates for innovation in firms with missing R&D, even though it could provide less biased coefficient estimates in OLS models.

Panel C of Table 5 compares MI with an alternative innovation benchmark, text-based innovation (Bellstam et al., 2020). Text-based innovation measures firm innovation using Latent Dirchlet Allocation (LDA) analysis for analyst reports of the S&P 500 firms. This measure is standardized with a mean of zero and variance of 1, and therefore we can effectively only compare the rank correlation between MI and the text-based innovation measure. The results in Panel C of Table 5 show that there is a strong and statistically significant rank correlation, 30%, between MI and text-based innovation. Multiple imputed R&D is more highly correlated with the text-based innovation measure than USPTO patents, global patents, or reported R&D. An alternative approach is to analyze text-based negative innovation, which focuses on the subset of analyst innovation coverage, where the analyst reports have negative aggregate sentiment. This measure captures the negative assessment of analysts concerning potential innovation outcomes for a firm. These results again show that R&D measures are more highly correlated with text-based innovation than patent measures. However, in this case, the correlation is stronger for reported R&D than for imputed R&D.

Overall, the results in Table 5 show that firms with recovered R&D differ from firms that explicitly report zero R&D, they are dissimilar to the average firm in the industry, and MI provides the closest imputation to the true value of their R&D investment. The textual analysis results arguably provide evidence that captures missing innovation in firms with and without meaningful R&D. We interpret the differences between the two textual analysis results to suggest that MI is more closely associated with average innovation.

V. Patents

Many studies of corporate innovation rely on patent data from the USPTO, with studies of international firms also tending to rely on this patent database. We investigate the performance of different imputation methods using two sets of counterfactuals: first, we use the product innovation measure of Mukherjee et al. (2017), and second, we use non-USPTO patents of US firms as counterfactuals. Different from R&D, patents represent the outcomes of innovation. Consequently, new product announcements arguably provide a good counterfactual for patents but not for R&D.

A. Firms without Patents: New Products as Counterfactuals

We use MI to predict the number of patents of firms with new major products, as in Mukherjee et al. (2017), without USPTO patent applications. New product announcements are measured using a manually-constructed database of major new product announcements by a textual search of the LexisNexis News database for company press releases, followed by an analysis of equity returns from the Center for Research in Security Prices (CRSP) around the announcement day. New products (NEW_PRODUCTS) capture the cumulative abnormal returns around product announcements made by each firm each year, and major new products (MAJ_NEW_PRODUCTS) count the number of announcements with cumulative abnormal returns above the 75^{th} percentile in the respective calendar year for each firm. Table 6 presents a comparison between two different MI methods. MI M1 includes the Lasso variables only: total assets, stock liquidity, R&D stock, and industry patent intensity using PATSTAT patents; MI M2 is the multiple imputation of non-USPTO patents using the same model as M1 and total assets, ROA, PPE, capital expenditure, sales growth, and leverage as conditioning information.

Panel A presents the innovation characteristics of firms with various coverage of patents and new products. The majority of the sample has no patents and no new products (column 2). These firms have the lowest reported R&D expenditures and the lowest percentage of R&D reporting. Just over 3% of the sample has both USPTO patents and new products (column 3), but these firms have the highest reporting rates for R&D (91%) and the largest R&D expenditures. About 6% of the sample has no USPTO patents but announce new products (column 1). The difference between firms with no patents that have new products (column 1), and firms with patents but no new products (column 4) is large and statistically significant (*t-statistics* ≥ 40.18).

Panel B of Table 6 presents the single and multiple imputations for the four categories above and compares firms without patents that have new products (column 1) to firms with patents but no new products (column 4). Imputing the number of patents with zero or industry average is statistically different from the counterfactuals of firms with patents and no new products (|*t*-

statistics| ≥ 4.00). In contrast, imputing with multiple imputation results in patent numbers that are not statistically different from firms with patents and no new products (|*t-statistics*| ≤ 1.26).

B. Firms without Patents: Global Patents as Counterfactuals

Innovation-related studies across accounting, economics, and finance mainly use data from the USPTO-NBER dataset. Firms without USPTO patents are typically deleted or counted as non-innovative firms when using USPTO data and generate a bias in the coverage of patenting. Over 14,000 US firms applied for USPTO patents in the sample period, 9,518 US firms received patents abroad, while 1,676 non-US firms received USPTO patents and 1,758 non-US firms received non-USPTO patents. US firms are granted, on average, 22 patents a year outside the US and 28.2 patents in the US.

As an alternative counterfactual, we use the sample of US firms that patent abroad to investigate the different methods of handling unreported patents, similar to Panel B of Table 6. We impute observations without USPTO patents with zero, industry mean (two-digit SIC code), and multiple imputation. MI is carried out with all the variables in Table 4, by industry (two digits), and the Lasso variables of total assets, stock volume, patent intensity, and R&D stock. Results in Figure 2 show that US patents abroad are not equal to zero, they are different from the USPTO industry mean, but they are not substantially different from MI.

VI. Simulation Analysis

We further consider three simulation studies, one based on the empirical distribution of Compustat (US) data, one on the empirical distribution of patent data, and one on simulated data, to compare different methods of dealing with missing data in various data generating processes (DGPs). In each case, we compare six methods to handle missing values.

First, we consider listwise deletion. Second, we impute the missing R&D expenditure by zeros (ImpZero). Third, we impute the missing R&D by the industry average (ImpMean). Specifically, if an observation of firm *i* at time *t* is missing, we impute the missing observation by the industry average (two-digit SIC code) for the firm in the same year. For both ImpZero and ImpMean, we also include a dummy variable indicating missingness as an explanatory variable. Fourth, we use Heckman's selection procedure (HS) with the selection variables containing all the observed covariates *W*. The Heckman selection procedure first predicts firms' selection probabilities by *W*, then corrects the selection bias by including a transformation of these predicted probabilities as an additional explanatory variable. Next, we consider the inverse probability weighting method (IPW) that weights each *i*-th observed point by the inverse of its conditional selection probability in least square estimation. We use the standard package of the HS and IPW procedures in STATA. Finally, we consider MI. Since the variables generating the missingness are not known a priori, we use all observables including the outcome variable as selection variables in the imputation model. To implement MI, we use 200 imputations based on a Markov chain Monte Carlo (MCMC) procedure and employ a multivariate normal regression for each imputation.

We evaluate the performance of the six methods with two criteria: the bias (B) and root mean squared error (RMSE) of coefficient estimates of the main regression. In particular, let θ be the coefficient vector of the main regression of interest. We calculate the bias and RMSE of the estimate $\hat{\theta}$ respectively, by:

(3a) $$\mathrm{B}(\hat{\theta}) = \frac{1}{R}\sum_{r=1}^{R}\left|\hat{\theta}^r - \theta^0\right| / \theta^0$$

(3b) $$\mathrm{RMSE}(\hat{\theta}) = \left[\left[\frac{1}{R}\sum_{r=1}^{R}\left(\hat{\theta}^r - \theta^0\right)\right]^2 + var\left(\hat{\theta}^r\right)\right]^{1/2},$$

where $\hat{\theta}^r$ is the estimate in the *r*-th replication, $var(\hat{\theta}^r)$ is the estimated robust variance of $\hat{\theta}^r$, θ^0 is the true value of the parameter, R is the number of simulations. Note that we present the bias as a proportion of the benchmark θ^0 to compare across coefficients, thus one cannot use the reported bias (B) to calculate the RMSE. Based on the observed missingness of R&D and patents in the US, we consider two levels of missingness relevant for innovation variables: 50% and 70%. We perform 500 simulations.

A. Empirical Distribution-Based Simulation for Unreported R&D Expenditures

For the empirical distribution-based simulation, we begin with a panel sample of 783 firms in Compustat over the period 1992-2012, where we have non-missing information on all financial variables of interest, except for R&D. The data include (for ease of exposition, we adopt more concise variable notations here): natural log of total assets (A), leverage (L), intangible assets (I), Tobin's Q (Q), return on assets (R), R&D expenditure (RD), stock liquidity (V), industry patent intensity (PI), and sales growth (S). To investigate the effects of R&D missingness on the coefficient estimates of our evaluation model and how different methods of handling missing R&D perform, we generate R&D expenditure with missing observations that incorporate the three types of missingness. The resulting estimated coefficients ($\hat{\theta}^r$) under each condition are used to calculate the bias and RMSE per equations (3a) and (3b). Our baseline regression uses simulated sales growth as the dependent variable and R&D expenditures, the natural log of total assets, Tobin's Q, leverage, and return on assets as explanatory variables. This approach enables us to obtain a clean set of benchmark coefficients that are free from researcher intervention except for the balanced, non-missing data criteria.

1. Generate Unreported R&D Expenditures

To simulate the missing R&D, we employ a subsample of complete balanced panel data, without missingness in R&D, that contains 311 firms over 21 years from 1992 to 2012. A clear advantage of this approach is that we do not need to make assumptions or estimate the conditional distribution of the R&D given that it is not missing, which is typically difficult to obtain. More importantly, it allows us to introduce the three types of missingness more precisely into the data as described below while providing us with "true" values as benchmark cases. We generate a missing indicator for R&D, denoted by *M* that equals 1 if R&D is missing and zero otherwise. Once we model and assign missing R&D observations, we can obtain the simulated R&D, since the data are complete, and the non-missing observations are given by their original values.

To create a missing indicator for R&D, we consider the three missing mechanisms: *missing completely at random, missing at random,* and *missing not at random*. Let α_i be the individual firm effects and denote $\tilde{\eta}_{it}$ as an idiosyncratic error. The three missing patterns can be summarized by:

- Missing completely at random:

(4) $$M_{it} = \tilde{\eta}_{it},$$

- Missing at random:

(5) $$M_{it} = \alpha_i + \beta_O' X_{it}^O + \tilde{\eta}_{it},$$

- Missing not at random:

(6) $$M_{it} = \alpha_i + \beta_O' X_{it}^O + \beta_U' X_{it}^U + \tilde{\eta}_{it},$$

where X_{it}^O contains observed variables by researchers, while X_{it}^U is unobserved and only appears in the DGP but is omitted in imputation models. For MAR, we consider $X_{it}^O = (A_{it}, V_{it}, PI_{it})'$, which are the Lasso derived variables, and for MNAR, we add $X_{it}^U = I_{it}$ to X_{it}^O (where I_{it} represents intangible assets).

To generate the missing indicator M_{it}, we need to know the true values of the parameters α_i, β_O, and β_U. To incorporate the firm fixed effects, we adopt the commonly used assumption that the firm fixed effects are correlated with the time-average of covariates in a linear manner (see

Chamberlain (1984)), i.e. $\alpha_i = c + \gamma_O' \bar{X}_i^O + u_i$ in MAR and $\alpha_i = c + \gamma_O' \bar{X}_i^O + \gamma_U' \bar{X}_i^U + u_i$ in MNAR, where $\bar{X}_i^O = 1/T \sum_{t=1}^{T} X_{it}^O$, $\bar{X}_i^U = 1/T \sum_{t=1}^{T} X_{it}^U$ and u_i is the idiosyncratic noise. This assumption implies that we can incorporate the firm fixed-effects by augmenting equations (5) and (6) by the time-series averages of covariates, respectively, as:

(7) $$M_{it} = c + \beta_O' X_{it}^O + \gamma_O' \bar{X}_i^O + \eta_{it},$$

(8) $$M_{it} = c + \beta_O' X_{it}^O + \beta_U' X_{it}^U + \gamma_O' \bar{X}_i^O + \gamma_U' \bar{X}_i^U + \eta_{it},$$

where $\eta_{it} = \tilde{\eta}_{it} + u_i$. Since there are no fixed effects in equations (7) and (8), we can estimate all parameters in these two models and predict M_{it} based on these estimates. Specifically, we first estimate equations (7) and (8), respectively, by a probit regression of the missing data indicator for R&D using the panel data sample (783 firms). We set the estimates $\hat{c}$, $\hat{\beta}_O$, $\hat{\beta}_U$, $\hat{\gamma}_O$, and $\hat{\gamma}_U$, as the true parameters to generate the missing probability M_{it}^* in the complete subsample of the data:

(9) $$M_{it}^* = \Phi(pm_{it}).$$

Φ is the normal CDF function and pm_{it} is obtained for the three scenarios by:

1. Missing completely at random:

(10) $$pm_{it} = \eta_{it},$$

2. Missing at random:

(11) $$pm_{it} = \hat{c} + \hat{\beta}_O X_{it}^O + \hat{\gamma}_O \bar{X}_i^O + \eta_{it},$$

3. Missing not at random:

(12) $$pm_{it} = \hat{c} + \hat{\beta}_O X_{it}^o + \hat{\beta}_U X_{it}^u + \hat{\gamma}_O \bar{X}_i^O + \hat{\gamma}_U \bar{X}_i^u + \eta_{it},$$

where $\eta_{it} \sim IID\ N(0, \sigma_\eta^2)$ and $\sigma_\eta^2 = 0.15$ based on the empirical distribution of the error term. Once we obtain M_{it}^*, we set the (i,t)-th observation of R&D as missing ($M_{it}^* = 1$) depending on $M_{it}^* > \mathrm{Q}_\tau(M_{it}^*)$, where $\mathrm{Q}_\tau(M_{it}^*)$ is the τ-th quantile of M_{it}^*, and τ controls the percentage of missingness.

2. Generating Sales Growth

We simulate the outcome variable of interest, i.e., sales growth S, because observable growth is potentially influenced by variables omitted from our empirical specification. We want to isolate the impact of missing innovation data from the errors from omitted variables in our regression of sales growth on innovation.[7] We generate S in the complete subsample without any missingness (311 firms over 21 years). The DGP of S is based on the following model:

(13) $$S_{it} = \mu_i + \delta' RD_{it} + \theta' Z_{it} + \varepsilon_{it},$$

where μ_i is firm fixed effects, Z_{it} contains the determinants of sales growth, $Z_{it} = \{A_{it}, Q_{it}, R_{it}, L_{it}\}'$, and $\varepsilon_{it}$is the error term. Note that intangible assets are not observed and thus also not included in the DGP of S. The firm fixed effects μ_i are generated by $\mu_i = 0.1\iota' \bar{Z}_i$, where ι is a 4×1 vector of ones and $\bar{Z}_i = 1/T \ \sum_{t=1}^{T} Z_{it}$, and thus μ_i is correlated with sales growth determinants. To obtain the parameters for δ' and θ', we estimate (13) using the same complete subsample without missingness and fix the estimated values in the simulation. To allow the idiosyncratic error to be correlated with selection instruments, we generate $\varepsilon_{it} = \tilde{\varepsilon}_{it} + \bar{Q}_i$ in MAR and $\varepsilon_{it} = \tilde{\varepsilon}_{it} + 0.5(\bar{Q}_i + \bar{I}_i)$ in MNAR. Here $\bar{Q}_i$ and $\bar{I}_i$ are the time average of Tobin's Q and intangible assets for firm i, respectively, which drive the missingness of R&D as discussed in Section VI.A.1. $\tilde{\varepsilon}_{it} \sim IID\ N(0, \sigma_\varepsilon^2)$ and $\sigma_\varepsilon^2 = 0.18$ based on the empirical distribution of the residual from estimating equation (13).

3. Simulation Results

Table 7 reports the simulation results under three missing mechanisms and two levels of

[7] We use simulated sales growth rather than observed sales growth in the benchmarking exercise because it allows us to explicitly compare the estimated coefficients to the true values. In contrast, using observed sales growth in our tests allows bias from two sources: imputation bias and misspecification bias (e.g., omitted variables).

missingness (50% and 70%). When R&D is *missing completely at random*, we find that both bias and RMSE increase with increasing missingness in R&D (Panel A). All methods show a relatively small bias under MCAR, except IPW and HS. IPW and HS typically do not include fixed effects due to the difficulty in estimating fixed effects in binary model settings, which potentially explains part of their relatively poor performance. In MCAR, MI has the lowest bias. MI exhibits relatively smaller RMSE than other methods too. Deterministic imputation methods (ImpZero and ImpMean) generate double the bias in multiple imputations and RMSEs that are similar to MI. Still, MI has both the lowest average bias and RMSE under MCAR.

Panel B shows the results for MAR, where the bias of all methods increases from MCAR. Under MAR, all methods lead to biased estimates, not only for R&D (which has missing observations), but also for the other explanatory variables that do not have any missingness. MI on average, produces the lowest bias across all of the six methods followed by ImpMean and ImpZero. The average absolute bias in LD is over 10 times greater than the bias in MI, while bias in IPW and HS are over 170 times and 80 times greater than MI. The common imputation methods, on average, exhibit similar RMSEs, where ImpZero, ImpMean, and MI have the lowest RMSEs. Panel C shows the results when missingness is driven by unobservables (MNAR). Under MNAR, MI continues to produce the lowest bias among all six methods followed by ImpZero and ImpMean. The bias in LD is six times larger than the bias in MI. Focusing on RMSE, once again ImpZero, ImpMean, and MI all exhibit similar magnitudes.

In the Internet Appendix, we also investigate other settings for both the determinants of R&D missingness in equations (11) and (12), as well as the sales growth DGP in equation (13). We introduce a larger set of conditioning variables in equations (11) and (12) in Panel A of Table IA2, a larger set of conditioning variables in equation (13) in Panel B of Table IA2, and Lasso as a variable selection procedure to determine which covariates should be included in the MI models in Panel C of Table IA2. In the last setting, we consider a Double Lasso procedure that applies Lasso to both R&D expenditure and sales growth regressions, which allows us to select variables that are correlated with both R&D and sales growth.[8] The results in Table IA2 are quantitatively similar to those in Table 7. MI consistently gives the lowest average bias across the various methods.

Our Table 7 simulation results focus on two separate levels of R&D missingness, namely 50% and 70%. However, our cross-country sample, which underlies Tables 1 to 4 reveals that the level of missingness varies by country. Consequently, we repeat the simulation analysis across a wide selection of missingness levels in 5% increments. Figure 3 shows the relative bias in the R&D coefficient estimate in using MI and LD as the rate of missing R&D increases from 5% to 85% for MAR. Across the entire range of missing R&D, MI exhibits substantially lower bias in the R&D coefficient estimate relative to LD.

It is worth noting that these results constitute a lower bound on bias generated by LD and deterministic imputation methods for two reasons. First, we include all the missingness determinants (A, V, and PI) as control variables, which implies that even if one knows the missingness mechanism and correctly controls for it, the estimated coefficients are still biased. Second, we have assumed that the errors in the sales growth and the selection regressions are not correlated. In untabulated results, we show that if the errors of the two regressions are correlated, then the bias of deletion and deterministic imputation increases.

[8] We first employ Lasso to select covariates in the model regressing R&D expenditure on all available covariates: Tobin's q, total assets, leverage, ROA, liquidity, industry patent intensity, and their time-series averages for each firm. We denote the selected covariates as X_{RD}. Next, we use Lasso to select the covariates in the model regressing sales growth on all available covariates specified above and obtain the selected covariates denoted as X_{SG}. We use the union of X_{RD} and X_{SG} as variables in multiple imputation.

B. Empirical Data-Based Simulation for the Absence of Patents

Patents and R&D expenditures may have different determinants and missingness levels. To further understand the properties of the different methods for handling missing data in the patent setting, we replicate the empirical distribution-based simulation, with the USPTO patent data distribution. Table 8 presents the results of the simulation based on the patent empirical distribution. Under MAR, IPW and HS generate the highest biasness in coefficient estimates relative to both imputation and deletion. Focusing on MNAR, deterministic imputation and multiple imputation both perform better than LD, IPW and HS approaches. Strikingly, the magnitudes of the bias and RMSE are substantially higher from excluding firms without patents than from excluding firms without R&D. For instance, the bias induced on the coefficient estimates from excluding firms from without patents is roughly eight times higher than that found with R&D. In short, the absence of patent data for so many firms makes this measure of corporate innovation especially sensitive to research design choices.

C. Simulation with Generated Data

1. Data Generating Process

We generate the dependent variable of interest as follows:

(14) $$Y_i = z_{1i}\theta_1 + z_{2i}\theta_2 + \varepsilon_i, \quad i = 1, \dots, N,$$

where $\theta_1 = \theta_2 = 1$, $\varepsilon_i \sim IID\ N(0,1)$, and N=1,000. The two covariates z_{1i} and z_{2i} are generated by a multivariate normal distribution with unit means and variance-covariance matrix specified later. z_{1i} contains missing observations, while z_{2i} is completely observed. Let M_i be the missing indicator of x_{1i} that equals 1 if x_{1i} is missing and zero otherwise, which is determined by $M_i = 1[M_i^* > Q_\tau(M^*)]$, where $1[\cdot]$ is an indicator function, M_i^* is a latent variable, and $Q_\tau(M^*)$ is the

τ-th quantile of M^*. We consider two values of $\mathrm{Q}_\tau(M^*)$, 0.7 and 0.5, which correspond to 70% and 50% of missing observations in x_{1i}, respectively. We consider three missing mechanisms for x_{1i}:

- Missing completely at random: $M_i^* = \eta_i$,
- Missing at random: $M_i^* = x_{1i}\gamma_1 + x_{2i}\gamma_2 \ + \eta_i$,
- Missing not at random: $M_i^* = x_{1i}\gamma_1 + x_{2i}\gamma_2 + x_{3i}\gamma_3 + x_{4i}\gamma_4 + \eta_i$.

We set $\{\gamma_1, \gamma_2, \gamma_3, \gamma_4\} = \{2,1,1,1\}$, and x_{1i} and x_{2i} are observed covariates that drive the missing pattern, while x_{3i} and x_{4i} are unobserved. η_i is the error term, independently generated from N(0,1) in MCAR, but correlated with ε_i in MAR and MNAR. We consider various patterns of correlations between the generated variables. In the benchmark case, we set the covariance matrix for the multivariate normally distributed $\{z_1, z_2, x_1, x_2, x_3, x_4, \varepsilon, \eta\}$ as:

$$\begin{pmatrix} 1 & & & & & & & \\ 0.4 & 1 & & & & & & \\ 0.5 & 0.4 & 1 & & & & & \\ 0.4 & 0.4 & -0.2 & 1 & & & & \\ 0.2 & 0.1 & 0.2 & 0.3 & 1 & & & \\ 0.1 & 0.2 & 0.1 & 0.1 & 0.1 & 1 & & \\ 0 & 0 & 0 & 0 & 0 & 0.4 & 1 & \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.4 & 1 \end{pmatrix}.$$

Note that all covariates $\{z_1, z_2, x_1, x_2, x_3, x_4, \varepsilon, \eta\}$ are correlated with each other. The two error terms are correlated with each other, but they are independent of the observed covariates. In MNAR, the missingness is also driven by two unobservables, which may be correlated with the errors. Hence, the unobserved selection variable x_3 is uncorrelated with both errors, and x_4 correlated with ε.

2. Results

Table IA3 in the Internet Appendix presents the results for the simulation with generated data. As in the previous simulation, both bias and RMSE increase with missingness in z_1 and from MCAR to MNAR. We first focus on the results with 70% missingness. Panel A shows the results for MCAR, where all methods produce negligible bias and small RMSE for both explanatory variables. There are only marginal differences across LD, HS, IPW and MI. However, the two deterministic imputation methods (using zero and industry average) produce the largest biases and RMSE.

Panel B shows the results for MAR, where listwise deletion exhibits a substantial sample selection bias, and both coefficient estimates are downward biased at 15% and 12% respectively for θ_1 and θ_2. Imputation using zeros or industry means increase the bias in θ_1 from -19% to -23% but decreases the bias in θ_2 from 28% to 12%. The Heckman selection procedure exhibits among the smallest bias that is comparable to MI, but at the cost of variance. This is reflected in the large RMSE of the Heckman selection estimates, suggesting that the two-step procedure is rather inefficient. In contrast, MI performs well: despite the increase in biasness, it continues to have among the lowest bias. The bias of MI is around half as large as that of listwise deletion, and almost two times smaller than that of imputation using zeros or means, and MI has the smallest RMSE.

For MNAR, all methods produce biased estimates due to the non-random missing pattern, but the degree of bias differs substantially across methods (Panel C). Imputation using zeros or mean leads to the largest bias and RMSE for θ_1 among the six methods (θ_1 bias is around 28%), while the bias and RMSE for θ_2 are generally in the middle of the six methods (θ_2 bias is 14%). The biasness in LD and IPW methods also deteriorates in comparison to the MAR setting, leading to more than 17% and 15% downward bias in θ_1 and θ_2 respectively. Both the bias and RMSE for the Heckman selection procedure deteriorated by 62.5% in comparison to MAR (the largest deterioration among the six methods). Despite the observed deterioration compared to MAR, MI continues to produce the lowest bias and RMSE among the six methods. Replicating our analysis across different levels of missingness reveals that multiple imputation consistently exhibits the smallest bias. In contrast, the bias from listwise deletion increases substantially with the level of missingness, suggesting that studies relying on patents or citations could exhibit substantial bias with these approaches.

VII. Impact of Bias on Inferences through Replication

Our analysis so far has conceptually demonstrated the problems with the various methods of handling unreported R&D. We use the analysis of Fama and French (2002, FF02 hereafter) to assess the economic relevance of the effect of different methods of handling unreported R&D on economic inference. They test the implications of trade-off and pecking order models for firm dividends and leverage. FF02 find ambiguous results on the relation between investments and leverage, as the two proxies for investment have opposite signs: market-to-book is positively correlated with leverage, and R&D expenditures are negatively correlated with leverage. For expositional simplicity, we focus this analysis on comparing multiple imputation to the two most commonly used solutions to missing R&D, namely deleting firms without R&D and classifying them as zero innovators.

We replicate their sample and note that 60% of the firms in their sample do not report R&D expenditures. FF02 classify all firms with unreported R&D as having zero R&D, and they include a dummy variable equal to 1 to differentiate firms with unreported R&D from firms that report zero R&D. We estimate the leverage regression (equation 15) below to evaluate if leverage differs across firms in the manner predicted by the trade-off or pecking order model using three approaches—listwise deletion, zero imputation with a missing dummy, and multiple imputation—and compare the resulting estimates:

(15) $$\frac{L_t}{A_t} = \beta_0 + \beta_1 \frac{V_t}{A_t} + \beta_2 \frac{ET_t}{A_t} + \beta_3 \frac{Dp_t}{A_t} + \beta_4 RDD_t + \beta_5 \frac{RD_t}{A_t} + \beta_6 \ln(A_t) + e_t.$$

We follow FF02 in the choice of the sample period, variables of interest, and notation. $\frac{ET_t}{A_t}$ the ratio of annual pre-interest pre-tax earnings to end-of-year total assets, is a proxy for the expected profitability of assets in place.[9] $\frac{V_t}{A_t}$, the ratio of a firm's total market value to its book value, is a proxy for expected investment opportunities. $\frac{RD_t}{A_t}$, the ratio of R&D expenditures to assets, is an additional proxy for expected investment. Unreported R&D is imputed with zero. RDD_t is a dummy variable equal to 1 for unreported R&D, and zero otherwise. $\frac{Dp_t}{A_t}$, the ratio of depreciation expense to assets serves as a proxy for non-debt tax shields. $\ln(A_t)$, the natural logarithm of total assets is a proxy for volatility. The sample period is 1965-1999 as in FF02.

Table IA4 (Internet Appendix) replicates FF02 using a contemporaneous regression with two-way fixed effects, double clustered standard errors, and five additional treatments for unreported R&D. Panel A presents results for dividend payer firms and Panel B for non-dividend payer firms. We use listwise deletion (*LD*), multiple imputation with only the variables in the regression (*MI*), multiple imputation with Lasso variables volume and patent intensity (*MI Lasso*), pseudo-R&D, and text-based innovation. These estimation techniques lead to very different estimates for β. The estimates based on zero imputation and listwise deletion reported in columns (1) and (2) are negative and differ considerably from estimates using multiple imputation (columns

[9] ET_t, earnings before taxes, preferred dividends, and interest payments is the income that could be sheltered from corporate taxes by interest deductions. Thus, $\frac{ET_t}{A_t}$ is a measure of profitability when we look at tax effects in the trade-off model.

3 and 4), which are positive. This suggests that the inferences made by the researcher in innovation could be driven by how they chose to deal with missing innovation data.

Using alternative measures of innovation like pseudo-R&D and text-based innovation leads to similar results to zero-imputation and listwise deletion. The pseudo-R&D result is heavily driven by the zero imputation of the rest of the observations. The estimates in Table IA4 illustrate that the method used to handle missing R&D can lead to substantially different inferences. Using multiple imputation for missing R&D in this setting potentially explains the puzzling findings in the original FF02 study.

VIII. Conclusions and Recommendations

Most public firms do not report R&D expenditures, do not obtain patents, or receive patent citations. Studies across accounting, economics, and finance typically exclude firms without reported R&D or patent activity or classify them as zero innovators (e.g., Autor, Dorn, Hanson, Pisano, and Shu (2020), Corrado, Martin, and Wu (2020), De Simone, Huang, and Krull (2020), Koch et al., 2021). Instead of arising randomly, we document that unreported innovation is systematically correlated with several firm, industry, and country characteristics. Accordingly, eliminating firms without R&D or patents provides biased results, if a proposed innovation covariate is correlated with any of these predictors (e.g., firm size, leverage, or profits). Because patent prevalence is even lower than the frequency of reported R&D, concerns about biases from deleting firms without patents are especially pronounced.

Our results show that unreported R&D and firms without patents are predictable and that the variables used to predict this missingness are known determinants of both innovation and other corporate outcomes of interest. Consequently, in studies that rely on the traditional methods of handling unobservable innovation, the residual in the regressions will likely be correlated with other explanatory variables. One of the most important takeaways from these findings is that commonly used solutions to handle unreported innovation can lead to biased parameter estimates that make prior inferences about corporate innovation difficult to assess.

It is difficult to give definitive solutions to dealing with missing innovation across different datasets, countries, and research settings. Our results using US data reveal that the two common methods for handling missing innovation data can provide biased coefficient estimates and standard errors. Strikingly, across a wide range of specifications, multiple imputation exhibits the least bias and RMSE among the six methods we investigate. Importantly, MI is consistently the best solution for unreported innovation data in our analysis. Of course, in a single industry analysis with limited numbers of positive R&D firms (e.g., real estate renting and leasing, SIC 53) or where upward of 90% of the missing R&D firms arise from zero R&D expenditures, imputing missing innovation data as zero will provide a reasonable solution. Yet, MI still performs well in this scenario as well.

The results allow us to provide some general guidelines and recommendations for economics and finance scholars confronted with unreported innovation.

1. In studies of innovation, missing R&D and patents can arise from: i) random data collection error by data providers, ii) managers not reporting R&D expenses due to zero (near zero) innovation, iii) strategic disclosure choices in reporting R&D expenses and patenting, iv) unsuccessful R&D, or v) firms filing for patents in alternative patent offices. Consequently, researchers should report both full and partial sample characteristics of the variables of interest. The level or degree of missingness of the innovation variable being used should be noted.
2. Researchers with missing innovation data should test if the missing data are predictable or *missing completely at random*. Little (1988) provides a test to determine if the data are MCAR. For Stata users, the *mcartest* command implements this test.

3. If the missing data are unpredictable or MCAR (maybe because the missing data stems from random collection errors by the data provider), then researchers could potentially delete or exclude the observations with missing data.
4. If the missing data are predictable, then researchers should attempt to predict missing innovation data using economically motivated observable variables. The predictive variables should be included as covariates in the regression and selection model. The researcher could use multiple imputation (for Stata users the MI command) to handle the missing observations.
5. If the missing data are predictable and there are both observable and unobservable characteristics that lead to missing innovation data, the problem is more challenging. Schafer and Graham (2002) show that MI can often be unbiased for MNAR and MAR data even though the researcher assumes the data to be MAR. Conceptually, both Heckman selection correction and MI on remain appealing, with both approaches involving assumptions and tradeoffs. Surprisingly, Heckman selection and inverse probability weighting are the worst performers under MNAR in our simulations, with MI typically performing the best.

In finance studies, a researcher must often decide on the assumptions of MAR versus MNAR, which involves the use of MI versus HS respectively. Yet, the assumptions underlying both HS and MI are likely to be violated. In practice, violating the assumptions of MNAR often has only a minor impact on estimates and standard errors under MI because the covariates included in imputation models are often correlated with the determinants of missingness (Collins, Schafer, and Kam (2001)). Consequently, researchers should analyze the randomness and predictability of the missing innovation data in their sample.

References

Anton, J., and D. Yao. "Little Patents and Big Secrets: Managing Intellectual Property." *Rand Journal of Economics*, 35 (2004), 1-22.

Autor, D.; D. Dorn; G. Hanson; G. Pisano; and P. Shu. "Foreign Competition and Domestic Innovation: Evidence from US Patents." *American Economic Review: Insights*, 2 (2020), 357-74.

Bellstam, G.; A. Cookson; and S. Bhagat. "A Text-based Analysis of Corporate Innovation." *Management Science*, forthcoming (2020).

Bena, J., and K. Li. "Corporate Innovations and Mergers and Acquisitions." *Journal of Finance*, 69 (2014), 1923-1960.

Branstetter, L.; M. Drev; and N. Kwon. "Get with the Program: Software-driven Innovation in Traditional Manufacturing." *Management Science*, 65 (2019), 541-558.

Chamberlain, G. "Panel Data." In *Handbook of Econometrics*, eds. Z. Griliches and M. D. Intrilligator, eds. North-Holland Amsterdam (1984), 1248-1318.

Collins, L.M.; J. L. Schafer; and C. Kam. "A Comparison of Inclusive and Restrictive Strategies in Modern Missing Data Procedures." *Psychological Methods*, 6 (2001), 330-351.

Corrado, C.; D. Martin; and Q. Wu. "Innovation α: What do IP-intensive Stock Price Indexes Tell Us about Innovation?" *American Economic Review Papers and Proceedings*, 110 (2020), 31-35.

De Simone, L.; J. Huang; and L. Krull. "R&D and the Rising Foreign Profitability of U.S. Multinational Corporations." *The Accounting Review*, forthcoming (2020).

Fama, E., and K. French. "Testing the Trade-off and Pecking Order Predictions about Dividends and Debt." *Review of Financial Studies*, 15 (2002), 1-33.

Financial Accounting Standard Board. *Accounting Standards Codification* (ASC 205-10-45). (n.d.).

Grangier, D., and I. Melvin. "Feature Set Embedding for Incomplete Data." *Advances in Neural Information Processing Systems Conference Proceedings* (2010).

Heitjan, F., and R. Little. "Multiple Imputation for the Fatal Accident Reporting System." *Journal of the Royal Statistical Society*, Series C. 40 (2019), 13-29.

Hombert, J., and A. Matray. "Can Innovation Help U.S. Manufacturing Firms Escape Import Competition from China?" *Journal of Finance*, 73 (2018), 2003-2039.

Horton, N., and S. Lipsitz. "Multiple Imputation in Practice: Comparison of Software Packages for Regression Models with Missing Values." *American Statistician*, 55 (2001), 244-254.

Hui, F.; D. Warton; and S. Foster. "Tuning Parameter Selection for the Adaptive Lasso using ERIC." *Journal of the American Statistical Association*, 110 (2015), 262-269.

Koch, A.; M. Panayides; and S. Thomas. "Common Ownership and Competition in Product Markets." *Journal of Financial Economics*, 131 (2021), 109-137.

Koh, P.-S., and D. Reeb. "Missing R&D." *Journal of Accounting and Economics*, 60 (2015), 73-94.

Koh, P.-S.; D. Reeb; and W. Zhao. "CEO Confidence and Unreported R&D." *Management Science*, 64 (2018), 5461-5959.

Lerner, J., and A. Seru. "The Use and Misuse of Patent Data: Issues for Corporate Finance and Beyond." Working Paper, Harvard University (2017).

Little, R. "A Test of Missing Completely at Random for Multivariate Data with Missing Values." *Journal of the American Statistical Association*, 83 (1988), 1198-1202.

Masulis, R., and E. Zhang. "How Valuable are Independent Directors: Evidence from External Distractions." *Journal of Financial Economics*, 132 (2019), 226-256.

Mukherjee, A.; M. Singh; and A. Zaldokas. "Do Corporate Taxes Hinder Innovation?" *Journal of Financial Economics*, 124 (2017), 195–221.

Petersen, M. "Estimating Standard Errors in Finance Panel Data Sets: Comparing Approaches." *Review of Financial Studies*, 22 (2009), 435–480.

Png, I. "Law and Innovation: Evidence from State Trade Secrets Laws." *Review of Economics and Statistics*, 99 (2017), 167-179.

Reeb, D.M., and W. Zhao. "Disregarding the Shoulders of the Giants: Evidence from Innovation Research." Working Paper, National University of Singapore (2021).

Robins, J., and N. Wang. "Inference for Imputation Estimators." *Biometrika*, 87 (2000), 113-124.

Rubin, D. "Inference and Missing Data." *Biometrika*, 63 (1976), 581–592.

Schafer, J., and J. Graham. "Missing Data: Our View of the State of the Art." *Psychological Methods*, 7 (2002), 147-177.

Thompson, S.B. "Simple Formulas for Standard Errors that Cluster by Both Firm and Time." *Journal of Financial Economics*, 99 (2011), 1-10.

TABLE 1
Sample Characteristics and Univariate Comparisons

Table 1 reports the sample characteristics and univariate comparisons. Panel A presents the sample characteristics. The sample period is 1999–2012. Panel B shows the difference in characteristics across different deletion methods. "Full Sample" uses all available observations without deletion based on either reported R&D or patent application information. "Report R&D" includes only observations that report R&D. "Report Patent" includes only observations that patent applications in any patent office, "Report R&D and Patent" includes only observations that have reported R&D and have patent filings in the PATSTAT database. Firm-years represent the maximum number of observations available for each subsample. Variable definitions are presented in the Appendix. *, **, and *** represent significance at the 10%, 5%, and 1% levels, respectively.

Panel A. Sample Characteristics

Variables	N 1	Mean 2	Median 3	Std. Dev. 4	25th 5	75th 6
R&D_EXPENDITURE	118,264	0.08	0.02	0.60	0.00	0.06
REPORT_R&D	333,920	0.35	0.00	0.48	0.00	1.00
ln(TOTAL_ASSETS)	330,790	6.74	6.64	2.96	4.75	8.61
PPE	328,021	0.28	0.23	0.23	0.01	0.43
TOB_Q	225,349	1.67	0.64	19.97	0.31	1.30
LEVERAGE	330,580	0.95	0.52	63.21	0.32	0.69
CAPEX	311,017	0.06	0.03	0.78	0.01	0.07
ROA	328,801	0.01	0.05	0.22	0.01	0.10
SALES_GROWTH	302,442	0.26	0.07	1.05	-0.04	0.25
#_PATENT_APP	333,920	9.36	0.00	140.78	0.00	0.00
#_PATENTS	333,920	4.50	0.00	69.54	0.00	0.00
CITATIONS	333,920	23.43	0.00	442.67	0.00	0.00

Panel B. Univariate Comparison of Samples

	Full Sample 1	Report R&D 2	Report Patent 3	Report R&D and Patent 4	Differences 5 = (1-2)/1	 6 = (1-3)/1	 7 = (1-4)/1
ln(TOTAL_ASSETS)	6.74	7.25	7.47	7.40	-8%***	-11%***	-10%***
PPE	0.28	0.24	0.23	0.20	14%***	18%***	29%***
TOB_Q	1.67	1.55	1.74	1.86	7%**	-4%	-11%***
LEVERAGE	0.95	0.53	0.57	0.48	44%***	40%***	49%***
CAPEX	0.06	0.05	0.05	0.05	17%***	17%***	17%***
ROA	0.01	0.00	-0.01	-0.03	100%***	200%***	400%***
SALES_GROWTH	0.26	0.23	0.25	0.31	12%***	4%**	-19%***
N (Firm-years)	330,790	118,264	53,456	26,273			

TABLE 2
Testing MCAR

Table 2 presents the missing completely at random test for the predictability of unreported innovation. The test is based on Little (1988) test for MCAR. Columns (1)-(4) *World* present the results for all countries in the sample. Columns (5)-(7) present the results for the *US* only. D.o.F. is the number of degrees of freedom, Prob> χ^2 is the probability of the null hypothesis that the data is MCAR. Variable definitions are presented in the Appendix.

	World				US		
	1	2	3	4	5	6	7
R&D ($m)	**X**	**X**	**X**	**X**	**X**	**X**	**X**
#_PATENT_APP	**X**	**X**	**X**	**X**	**X**	**X**	**X**
ln(TOTAL_ASSETS)		**X**	**X**	**X**		**X**	**X**
PPE			**X**	**X**		**X**	**X**
LEVERAGE			**X**	**X**		**X**	**X**
CAPEX			**X**	**X**			**X**
ROA				**X**			**X**
SALES_GROWTH				**X**			**X**
χ^2 dist.	297	7,431	25,062	42,971	6,359	9,857	22,889
D.o.F.	2	9	87	326	5	23	180
Prob> χ^2	0.00	0.00	0.00	0.00	0.00	0.00	0.00

TABLE 3
Predictability of Unreported R&D

Table 3 presents the OLS regression results for predictability of unreported R&D. Columns (1)-(4) *World* present the results for all countries in the sample. Columns (5)-(7) present the results for the *US* only. Standard errors are double clustered at firm and year level. T-statistics are presented in brackets. Variable definitions are presented in the Appendix. *, **, and *** represent significance at the 10%, 5%, and 1% levels, respectively. Adj. R^2 is the adjusted R^2.

	World				US		
	1	2	3	4	5	6	7
ln(TOTAL_ASSETS)	-0.028***	-0.024***	-0.018***	-0.009**	0.033***	-0.001	-0.009**
	(-12.17)	(-10.29)	(-11.24)	(-2.67)	(8.91)	(-0.25)	(-2.71)
PPE	0.228***	0.206***	0.179***	0.016*	0.237***	0.309***	0.049**
	(13.83)	(14.01)	(17.31)	(1.95)	(6.44)	(9.20)	(2.24)
LEVERAGE	-0.000	-0.000	0.000**	-0.000	0.002**	0.001**	0.001***
	(-1.04)	(-1.06)	(2.31)	(-0.11)	(2.18)	(2.32)	(6.73)
CAPEX	0.004	0.003	-0.000	0.000	0.019	-0.159***	-0.015
	(1.35)	(1.22)	(-0.62)	(0.22)	(0.31)	(-3.02)	(-1.18)
ROA	0.182***	0.172***	0.128***	0.022***	0.178***	0.166***	0.039***
	(15.34)	(14.66)	(10.79)	(3.79)	(8.13)	(10.19)	(3.92)
SALES_GROWTH	0.011***	0.006*	-0.002	0.002**	-0.006*	-0.007**	-0.001
	(3.29)	(1.79)	(-1.11)	(2.32)	(-1.77)	(-3.17)	(-0.88)
STOCK_LIQUIDITY	-0.006***	-0.006***	-0.005***	-0.000	-0.008***	-0.003***	-0.000
	(-7.47)	(-8.54)	(-12.00)	(-1.38)	(-12.37)	(-5.43)	(-1.53)
PATENT_INTENSITY	-556.167***	-6.925	13.555	37.054**	-573.734***	-21.605***	-19.704**
	(-15.05)	(-0.50)	(1.08)	(2.22)	(-12.74)	(-4.42)	(-2.36)
Country FE			Yes				
Industry FE		Yes	Yes			Yes	
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Firm FE				Yes			Yes
N	283,987	283,987	283,987	281,243	64,386	64,386	63,086
Adj. R^2	0.14	0.23	0.38	0.81	0.23	0.53	0.93

TABLE 4
Predicting Non-patent Seeking Firms

Table 4 presents OLS regressions of unreported USPTO patents and explanatory variables. The dependent variable is an indicator variable equal to 1 when a firm does not have USPTO patents, and zero otherwise. Columns (1)-(4) *World* present the regression results for all countries in the sample. Columns (5)-(7) present the regression for *US listed firms* only. Standard errors are double clustered at firm and year level. T-statistics are presented in brackets. Variable definitions are presented in the Appendix. *, **, and *** represent significance at the 10%, 5%, and 1% levels, respectively. Adj. R^2 is the adjusted R^2.

	World				US		
	1	2	3	4	5	6	7
ln(TOTAL_ASSETS)	-0.018***	-0.014***	-0.031***	-0.010***	-0.018***	-0.036***	-0.021***
	(-15.49)	(-13.42)	(-23.36)	(-8.03)	(-4.97)	(-12.75)	(-5.04)
PPE	0.109***	0.149***	0.117***	-0.007*	0.123***	0.225***	-0.034*
	(12.66)	(15.24)	(13.60)	(-1.72)	(5.68)	(8.12)	(-1.71)
LEVERAGE	0.000	0.000	0.000	-0.000	0.001*	0.000	-0.000
	(0.62)	(0.54)	(1.06)	(-1.08)	(1.79)	(0.71)	(-1.23)
CAPEX	0.000	-0.000	-0.001	-0.000	0.071	-0.122**	0.025
	(0.17)	(-0.15)	(-1.36)	(-0.85)	(1.55)	(-2.40)	(1.27)
ROA	0.141***	0.121***	0.145***	0.015***	0.286***	0.221***	0.019*
	(12.08)	(13.15)	(16.50)	(3.17)	(8.77)	(8.25)	(1.88)
SALES_GROWTH	0.003**	0.001	-0.005***	0.002***	-0.004	-0.004**	0.005***
	(2.18)	(0.73)	(-4.62)	(3.78)	(-0.94)	(-1.96)	(3.07)
STOCK_LIQUIDITY	-0.007***	-0.007***	-0.004***	-0.000**	-0.007***	-0.005***	-0.001**
	(-19.71)	(-21.10)	(-14.43)	(-2.55)	(-14.81)	(-11.10)	(-2.04)
PATENT_INTENSITY	-355.432***	-40.196**	-30.734**	-10.067*	-518.481***	-44.799**	-11.023
	(-16.00)	(-3.26)	(-3.38)	(-1.64)	(-14.05)	(-3.11)	(-0.74)
R&D_STOCK	-0.000**	-0.000**	-0.000**	-0.000	-0.000***	-0.000***	0.000*
	(-2.11)	(-2.13)	(-2.25)	(-1.12)	(-6.37)	(-4.71)	(1.78)
Country FE			Yes				
Industry FE		Yes	Yes			Yes	
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Firm FE				Yes			Yes
N	281,763	281,763	281,763	278,999	64,383	64,383	63,086
Adj. R^2	0.11	0.16	0.26	0.76	0.18	0.32	0.77

TABLE 5
Recovered R&D and Imputed Unreported R&D

Table 5 presents the Recovered R&D (in the years t-1 and t-2 from switch year) statistics and its comparison with different imputation methods. Panel A presents the comparison of Recovered R&D, Zero R&D, and positive R&D firm characteristics. Panel B presents the comparison of Recovered R&D for firms with SIC information with different imputation methods. *Recovered R&D* is the recovered R&D expenditure as reported in 10-K filings, its t-stat presents its difference from 0. *Imputed R&D (Industry Avg.)* is the average industry expenditure (two-digit) for the observations that are recovered, *Imputed R&D MI (Full Sample)* is the multiply imputed R&D using only the Lasso variables: ln(TOTAL_ASSETS), STOCK_LIQUIDITY, and PATENT_INTENSITY, by industry (two-digit) for the complete sample, *Imputed R&D MI (Sub Sample)* is the multiply imputed R&D using the same MI model on the restated R&D sub sample and the industry and size matched peers. "Diff." is the difference between Recovered R&D and imputed R&D. t-stat. represent the *t*-statistic for the difference between Recovered R&D an imputed R&D. Panel C shows the rank correlation among text-based innovation, patent applications (USPTO only and PATSTAT), R&D_EXPENDITURE, and multiple imputation. *, **, and *** represent significance at the 10%, 5%, and 1% levels, respectively.

Panel A. Recovered R&D and R&D Firms

	Recovered R&D	Zero R&D	Diff.	Positive R&D	Diff.
R&D ($m)	6.69	0.00	6.69***	112.76	-106.07***
R&D_EXPENDITURE	0.87	0.00	0.87	0.34	0.53
ln(TOTAL_ASSETS)	3.60	4.84	-1.24***	4.69	-1.09***
ROA	-3.11	-2.81	-0.30	-0.70	-2.41
PPE	0.20	0.30	-0.10***	0.18	0.02***
SALES_GROWTH	15.31	1.54	13.77	1.49	13.82
CAPEX	0.06	0.06	-0.01	0.05	0.01
LEVERAGE	2.61	6.74	-4.13***	1.81	0.80

Panel B. Comparison with Imputation

Variable	Mean	St. Dev.	R&D	Diff.	t-stat.
Recovered R&D	6.91	24.24			8.84
Imputed R&D (Industry Avg.)	77.86	92.13	6.91	-70.95	-23.19
Imputed R&D MI (Full Sample)	6.36	242.75	6.91	0.55	0.07
Imputed R&D MI (Sub Sample)	8.66	245.55	6.91	-1.74	-0.22

Panel C. Rank Correlation MI and Text-based Innovation

	Text-based Innovation	Text-based Negative Innovation
Patent USPTO	0.22***	0.17***
Patent PATSTAT	0.21***	0.15***
R&D_EXPENDITURE	0.26***	0.29***
Imputed R&D MI Full Sample	0.30***	0.27***

TABLE 6
New Products, Patents, and Imputation Methods

Table 6 presents an analysis of new product announcements, patents, and imputation methods for firms with different combinations of products and patents. Panel A presents the innovation (R&D and patents) and new product characteristics. New product announcement data is from Mukherjee et al. (2017) and patents are based on USPTO data. NEW_PRODUCTS include the average returns for all new product announcements and MAJ_NEW_PRODUCTS includes the number of new products in the 75th percentile of returns. Panel B presents the comparison of single and multiple imputation methods for patents with different product announcements. Single imputation includes: imputation with zero (*Impute Zero*), and imputation with the two-digit industry average (*Impute Industry Average*). MI M1 presents the multiply imputed USPTO patents using the Lasso variables: ln(TOTAL_ASSETS), STOCK_LIQUIDITY, R&D_STOCK, and PATENT_INTENSITY from PATSTAT patents and MI M2 is the same as M1 with the addition of ROA, PPE, CAPEX, SALES_GROWTH, and LEVERAGE. Column (1) presents the information for the subsample with no USPTO patents and with product announcements; column (2) presents the information for the subsample with no USPTO patents and no product announcements; column (3) presents the information for firms with USPTO patents and product announcements; and column (4) presents the information for firms with USPTO patents and no new product announcements. *Diff.* presents the difference between columns (1) and (4), and *t-stat.* present the t-statistic for the difference.

	No Patents + New Products	No Patents + No New Products	Patents + New Products	Patents + No New Products	Diff.	t-stat.
	1	2	3	4	5 = 1 - 4	
Panel A. Patents and New Products						
Obs. (% of Sample)	5.71%	75.58%	3.10%	13.83%		
R&D ($m)	85.72	32.05	285.20	148.39		
R&D (% of report)	56.79%	39.12%	91.38%	85.31%		
#_PATENT_APP	0.00	0.00	47.87	32.07		
NEW_PRODUCTS	0.07	0.00	0.13	0.00	0.07	40.18
MAJ_NEW_PRODUCTS	0.79	0.00	1.60	0.00	0.79	42.02
Panel B: Compare Imputation Method: Imputed Patents						
Single Imputation:						
Impute Zero	0.00	0.00	47.87	32.07	-32.07	-21.87
Impute Industry Average	24.54	28.74	47.87	32.07	-7.53	-4.00
Multiple Imputation:						
Imputed Patents MI M1	29.38	11.98	47.87	32.07	-2.68	-1.26
Imputed Patents MI M2	30.50	10.28	47.87	32.07	-1.57	-0.73

TABLE 7
Simulation Based on the Empirical Distribution from Compustat Data

Table 7 provides the evaluation statistics, bias (relative bias over true parameter) and root mean squared error (RMSE) for the simulation based on the empirical distribution from Compustat (US) data, as described in Section V.A. The empirical distribution is from the panel of 783 firms with non-missing information for all variables except R&D. The methods evaluated are listwise deletion (LD), imputation with zero (ImpZero), imputation with industry mean, two-digit SIC code (ImpMean), inverse probability weighting (IPW), Heckman selection procedure (HS), and multiple imputation (MI). MI uses total assets, stock liquidity, industry patent intensity identified using Lasso analysis in the regression and is estimated using MCMC with 200 iterations for convergence. The regressions for imputation with zero and industry mean include a dummy variable for the imputed observations. Absolute average represents the average of the absolute bias across all variables. We present results for three missingness mechanisms: missing completely at random (MCAR) in Panel A, missing at random (MAR) in Panel B, and missing not at random (MNAR) in Panel C. Variable definitions are presented in the Appendix. We generate missingness R&D for 50 and 70% of the sample. We conduct 500 simulations.

		Missing 70%						Missing 50%					
		LD	Imp Zero	Imp Mean	IPW	HS	MI	LD	Imp Zero	Imp Mean	IPW	HS	MI
						Panel A. MCAR							
Bias	R&D_EXPENDITURE	0.84	-0.58	-0.48	0.76	0.80	-0.04	0.67	-0.52	-0.47	0.73	0.68	-0.19
	ln(TOTAL_ASSETS)	1.42	-0.23	-0.22	18.00	18.11	0.01	0.86	-0.34	-0.33	18.00	18.11	-0.18
	TOB_Q	1.18	0.15	0.13	0.40	0.10	0.07	0.73	0.10	0.09	0.35	0.08	0.03
	LEVERAGE	0.24	-0.02	-0.02	0.82	0.75	0.03	0.20	-0.02	-0.01	0.81	0.69	0.02
	ROA	0.44	-0.01	-0.01	1.23	1.19	0.15	0.39	-0.01	-0.01	1.20	1.04	0.09
	Avg. Abs. Bias	***0.83***	***0.20***	***0.17***	***4.24***	***4.19***	***0.06***	***0.57***	***0.19***	***0.18***	***4.22***	***4.12***	***0.10***
RMSE	R&D_EXPENDITURE	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	ln(TOTAL_ASSETS)	0.02	0.01	0.01	0.19	0.19	0.01	0.02	0.01	0.01	0.19	0.19	0.01
	TOB_Q	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
	LEVERAGE	0.08	0.03	0.03	0.15	0.13	0.03	0.06	0.03	0.03	0.14	0.12	0.03
	ROA	0.13	0.04	0.04	0.22	0.20	0.05	0.10	0.04	0.04	0.19	0.17	0.05

		Missing 70%						Missing 50%					
		LD	Imp Zero	Imp Mean	IPW	HS	MI	LD	Imp Zero	Imp Mean	IPW	HS	MI
						Panel B. MAR							
Bias	R&D_EXPENDITURE	1.09	-0.66	-0.54	4.60	-19.52	-0.17	0.84	-0.38	-0.26	3.63	-23.51	0.00
	ln(TOTAL_ASSETS)	1.43	-0.32	-0.28	62.02	12.88	-0.03	1.00	-0.18	-0.16	62.21	26.46	0.05
	TOB_Q	1.25	0.24	0.24	2.68	-4.33	0.12	0.93	0.06	0.04	1.40	-17.88	0.02
	LEVERAGE	0.42	-0.04	-0.04	3.57	-0.67	-0.06	0.22	-0.03	-0.03	2.67	-0.71	0.00
	ROA	0.86	-0.04	-0.04	5.10	1.89	0.09	0.43	-0.04	-0.04	4.49	5.53	0.06
	Avg. Abs. Bias	***1.01***	***0.26***	***0.23***	***15.59***	***7.86***	***0.09***	***0.69***	***0.14***	***0.11***	***14.88***	***14.82***	***0.03***
RMSE	R&D_EXPENDITURE	0.01	0.00	0.00	0.02	0.33	0.00	0.00	0.00	0.00	0.01	0.30	0.00
	ln(TOTAL_ASSETS)	0.02	0.01	0.01	0.64	0.94	0.01	0.02	0.01	0.01	0.64	0.90	0.01
	TOB_Q	0.02	0.00	0.00	0.02	1.19	0.01	0.01	0.01	0.01	0.02	1.06	0.01
	LEVERAGE	0.12	0.04	0.04	0.61	2.97	0.04	0.06	0.04	0.04	0.46	2.90	0.03
	ROA	0.22	0.05	0.05	0.78	4.25	0.06	0.11	0.05	0.05	0.68	4.32	0.05
						Panel C. MNAR							
Bias	R&D_EXPENDITURE	0.89	-0.55	-0.49	4.81	-30.91	-0.21	0.64	-0.55	-0.52	3.61	-16.35	-0.15
	ln(TOTAL_ASSETS)	1.31	-0.18	-0.18	61.87	-2.98	0.09	0.96	-0.26	-0.25	62.06	1.27	-0.04
	TOB_Q	1.57	0.40	0.40	3.17	-33.98	0.29	0.89	0.31	0.30	2.07	-7.09	0.26
	LEVERAGE	0.36	-0.08	-0.08	3.51	-5.10	-0.10	0.20	-0.06	-0.06	2.63	-1.56	-0.03
	ROA	0.83	-0.07	-0.07	5.06	-0.46	0.05	0.45	-0.06	-0.06	4.21	-0.17	0.05
	Avg. Abs. Bias	***0.99***	***0.25***	***0.24***	***15.68***	***14.69***	***0.15***	***0.63***	***0.25***	***0.24***	***14.92***	***5.29***	***0.10***
RMSE	R&D_EXPENDITURE	0.01	0.00	0.00	0.02	0.38	0.00	0.00	0.00	0.00	0.01	0.35	0.00
	ln(TOTAL_ASSETS)	0.02	0.01	0.01	0.64	0.96	0.01	0.02	0.01	0.01	0.64	0.92	0.01
	TOB_Q	0.02	0.01	0.01	0.03	1.29	0.01	0.01	0.01	0.01	0.02	1.17	0.01
	LEVERAGE	0.11	0.04	0.04	0.61	2.87	0.04	0.06	0.04	0.04	0.45	2.62	0.04
	ROA	0.21	0.04	0.04	0.78	4.34	0.05	0.11	0.05	0.05	0.64	4.84	0.05

TABLE 8

Patent Simulation Based on the Empirical Distribution of Data

Table 8 provides the evaluation statistics, bias (relative bias over true parameter) and root mean squared error (RMSE) for the simulation based on the empirical distribution from Compustat (US) and USPTO data. The empirical distribution comes from the panel of 783 firms with non-missing information for all variables except USPTO patents. The methods evaluated are listwise deletion (LD), imputation with zero (ImpZero), imputation with industry mean, two-digit SIC code (ImpMean), inverse probability weighting (IPW), Heckman selection procedure (HS), and multiple imputation (MI). MI uses all the variables in sample and is estimated using MCMC with 200 iterations for convergence. The regressions for imputation with zero and industry mean include a dummy variable for the imputed observations. Absolute average represents the average of the absolute bias across all variables. Variable definitions are presented in the Appendix. We present results for two missingness mechanisms: missing at random (MAR) in Panel A and missing not at random (MNAR) in Panel B. We generate missingness in patents for 70% of the sample. We conduct 500 simulations.

		LD	Imp Zero	Imp Mean	IPW	HS	MI
		Panel A. MAR					
Bias	#_PATENT_APP	-14.77	-2.38	-2.88	16.86	-48.13	0.34
	ln(TOTAL_ASSETS)	21.85	1.64	3.33	62.00	-15.40	2.51
	TOB_Q	-7.95	1.47	1.13	-190.58	-238.28	1.23
	LEVERAGE	-1.94	0.02	0.04	9.08	9.84	0.07
	ROA	0.41	0.11	0.10	3.95	2.66	0.11
	Avg. Abs. Bias	***9.38***	***1.13***	***1.50***	***56.50***	***62.86***	***0.85***
RMSE	#_PATENT_APP	0.00	0.00	0.00	0.00	0.00	0.00
	ln(TOTAL_ASSETS)	0.06	0.02	0.03	0.12	0.09	0.03
	TOB_Q	0.03	0.02	0.02	0.41	0.52	0.02
	LEVERAGE	0.31	0.12	0.13	0.91	1.07	0.12
	ROA	0.51	0.32	0.33	2.96	2.42	0.32
		Panel A. MNAR					
Bias	#_PATENT_APP	-15.14	-0.22	0.05	10.25	-23.99	1.26
	ln(TOTAL_ASSETS)	19.45	-1.00	-1.22	52.67	2.04	0.06
	TOB_Q	-7.58	-0.34	-0.23	-106.94	-136.07	-0.39
	LEVERAGE	-1.84	0.03	0.02	4.07	3.65	0.06
	ROA	0.43	0.01	-0.01	2.23	1.37	0.01
	Avg. Abs. Bias	***8.89***	***0.32***	***0.31***	***35.23***	***33.42***	***0.36***
RMSE	#_PATENT_APP	0.00	0.00	0.00	0.00	0.00	0.00
	ln(TOTAL_ASSETS)	0.06	0.02	0.03	0.10	0.07	0.03
	TOB_Q	0.03	0.02	0.02	0.23	0.30	0.02
	LEVERAGE	0.29	0.13	0.13	0.42	0.57	0.13
	ROA	0.51	0.30	0.33	1.69	1.67	0.30

FIGURE 1
Ranking Innovative Firms

Figure 1 shows the rank of innovative firms for S&P 500 using R&D expenditures and Text-Based innovation. The figure shows innovation ranks based on the R&D expenditure imputed with zero for unreported R&D (*Imp Zero*), Text-based innovation measure, and Multiple Imputation (*MI*). Firms with reported R&D expenditures are shown in light orange, and those with unobserved R&D as red.

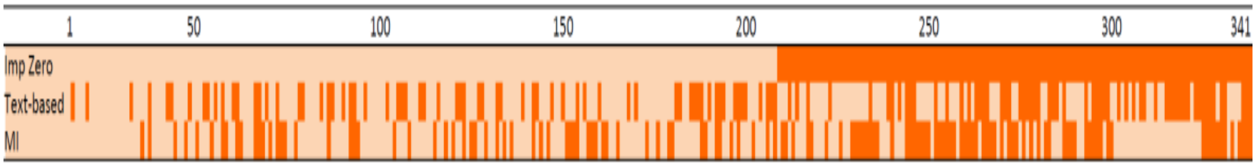


FIGURE 2
Imputation of Missing Patents

Figure 2 presents the t-statistics of the comparison of years with only non-USPTO patents for US firms with different imputation methods. Industry is the difference between non-USPTO patents and industry patents defined as the average industry expenditure for the observations the year with non-USPTO patents only, M1 is the multiply imputed non-USPTO patents using ln(total assets), ROA, PPE, capital expenditure, sales growth, and leverage by industry (two-digit), M2 is the multiply imputed non-USPTO patents using the same model as M1 without sales growth and leverage as conditioning information, M3 is the multiply imputed non-USPTO patents using the same model as M1 with the addition of R&D expenditure as conditioning information, M4 is the multiply imputed non-USPTO patents using the Lasso variables ln(total assets), stock liquidity, industry patent intensity, and stock R&D, M5 is the multiply imputed non-USPTO patents combining models M1 and M5.

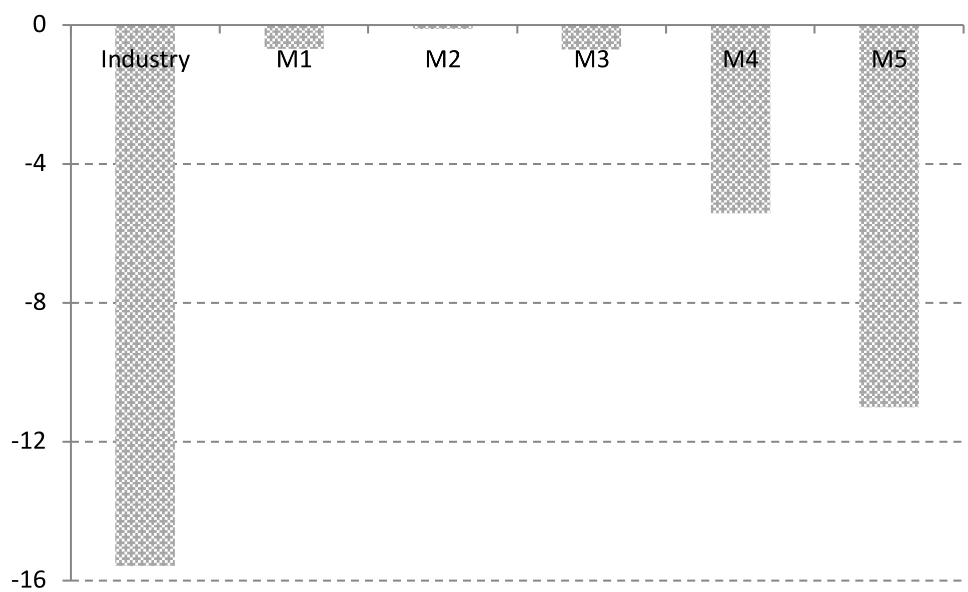
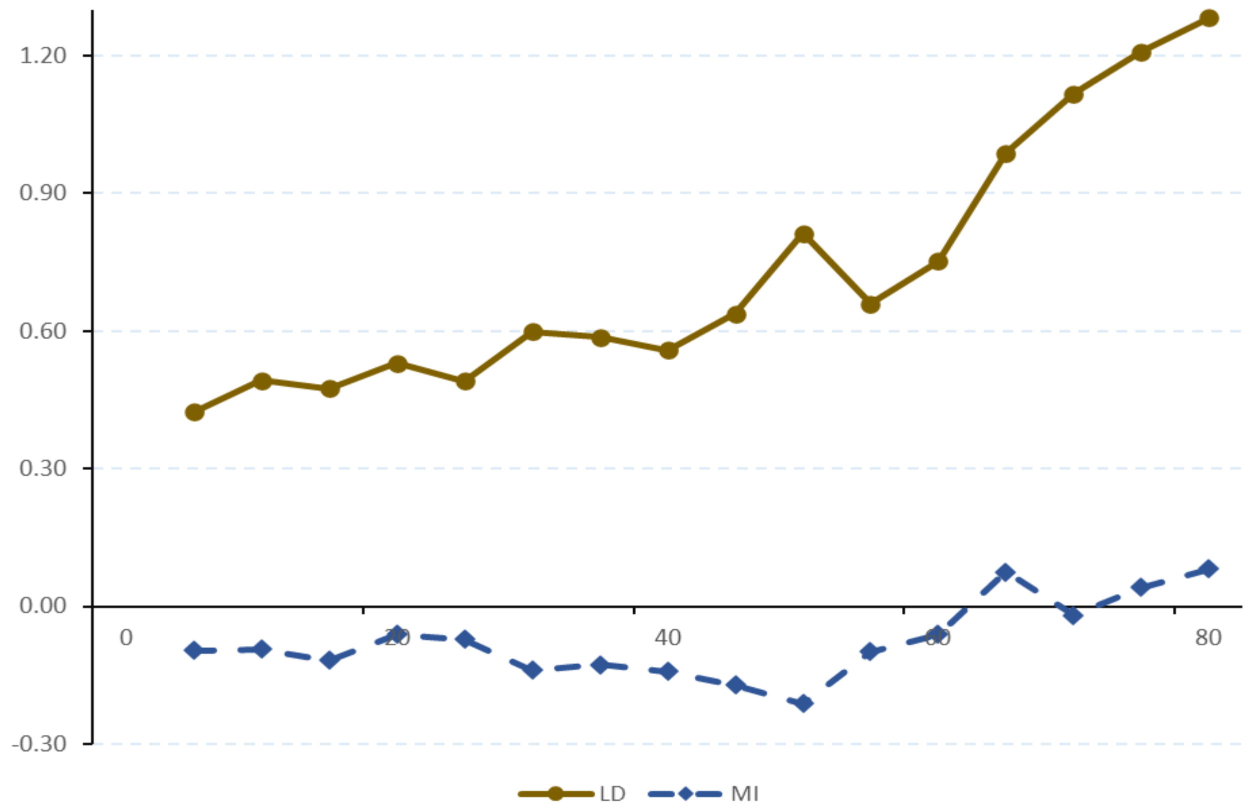


FIGURE 3
Bias from Deleting Firms without Reported Innovation

Figure 3 presents the bias of the R&D coefficient for listwise deletion (LD) and multiple imputation (MI) across different missingness levels. The simulation is based on the empirical distribution of the panel of 783 firms with non-missing information for all variables except R&D. MI uses all the variables in the regression in Section V.A and is estimated using MCMC with 200 iterations for convergence. We present results for data missing at random (MAR). We conduct 500 simulations.



Appendix. Variable Definitions

Variable Names	Notation	Variable Definitions	Code
R&D_EXPENDITURE	RD	R&D expenditure divided by total assets	XRD/AT
R&D ($m)		R&D expenditure in million USD	XRD
REPORT_R&D		Indicator variable: 1 if a firm reported zero or positive R&D expenditure; 0 otherwise	
PPE		Net property, plant, and equipment divided by total assets	PPENT/AT
TOB_Q	Q	Tobin's Q, measured as market value of equity divided by total assets	MKTVAL/AT
LEVERAGE	L	Total liabilities divided by total assets	LT/AT
ln(TOTAL_ASSETS)	A	Natural log of total assets	Ln(AT)
CAPEX		Capital expenditure divided by total assets	CAPX/AT
ROA	R	EBIT divided by total assets	EBIT/AT
SALES_GROWTH	S	Annual sales growth	$(Sale_t\text{-}Sale_{t-1})/Sale_{t-1}$
INTANGIBLES	I	Intangible assets	INTANO/AT
#_PATENT_APP		Total number of patent applications	
#_PATENTS		Total number of patents granted	
CITATIONS		Total number of citations per patent	
STOCK_LIQUIDITY	V	Yearly sum of daily trading volume in USD	PRC*VOL (for US-Stocks), PRCCD*CSHTRD*Exchange Rate (for non-US stocks)
PATENT_INTENSITY	PI	Number of PATSTAT patents per total assets for industry, using two-digit SIC across countries, unless specified otherwise	
R&D_STOCK		R&D accumulated over 10 years and 15% yearly depreciation	

Internet Appendix

This appendix provides a summary of the missing data problem and discusses several popular econometric approaches to handling missing data that are considered in this paper. With partially observed data, we can rarely be sure of the mechanism leading to such missing data. Therefore, we highlight some approaches to analyzing missing data under different mechanisms, which helps to establish inference robustness in the face of uncertainty about the missingness mechanism. In particular, we consider listwise deletion, deterministic imputation, inverse probability weighting, Heckman selection, and multiple imputation. For exposition simplicity (as in the main body of the paper), we consider the case where only one explanatory variable contains missing observations. Let y_i be the dependent variable and z_i be the explanatory variables with missingness. We have the linear relation:

$$y_i = \alpha + \theta z_i + \varepsilon_i, \quad i = 1, \ldots, N. \tag{IA1}$$

Let s_i be a selection indicator where $s_i = 1$ when z_i is not missing and firm i is included in the regression. Otherwise, when $s_i = 0$ firm i is deleted from the data. The validity of solutions to this problem depends on the missing mechanism, thus we first present the three missing mechanisms.

1. Missing completely at random (MCAR):

   $$P(s = 0|y, z, x) = P(s = 0).$$

   This means that the missing probability does not depend on any random variables.

2. Missing at random (MAR): The probability of missing can be formulated by:

   $$P(s = 0|y, z, x) = P(s = 0|x).$$

   In other words, the probability of missingness only depends on the set of *observed* variables x, but not on the missing variable itself nor on unobservables.

3. Missing not at random (MNAR): the missing mechanism is neither MAR nor MCAR. For example, the missing mechanism depends on the value of z itself, or on unobserved variables, e.g., high-income individuals do not participate in surveys related to income.

Effects of Listwise Deletion

Listwise deletion only uses a subsample of observations, deleting those that contain missing values in the z-variable.[10] This leads to estimating the following regression using the subsample of the data:

$$y_i = s_i\alpha + \theta s_i z_i + s_i\varepsilon_i, \tag{IA2}$$

[10] We consider the univariate setup for simplicity. There might be other covariates of interest that drive the outcome variable but including them in the regression does not change the problem of deletion.

where $s_i z_i$ is now the explanatory variable and $s_i \varepsilon_i$ is the error term. The OLS (ordinary least squares) estimator is unbiased if $E(s_i \varepsilon_i z_i) = 0$, which can be implied by $E(\varepsilon_i | z_i, s_i) = 0$. If MCAR holds and z_i is exogenous, then $E(\varepsilon_i | z_i, s_i) = E(\varepsilon_i | z_i) = 0$. Thus, deletion can lead to consistent estimates in the case of MCAR. However, if selection is driven by observed or even unobserved variables as in MAR and MNAR cases, $E(\varepsilon_i | z_i, s_i) \neq 0$ in general because ε_i can be correlated with s_i even if one controls for z_i, leading to biased estimates produced by deletion.

Deterministic Imputation

Another popular approach used in empirical studies is to impute the missing observations using various methods, and then treat the resulting data as given for further analysis. Frequently used deterministic imputation employs, e.g., zero, overall average, average from "similar" observations, or fitted values based on some pre-specified models. The validity of this method depends on whether the specified imputation models are correct. If the imputation model perfectly coincides with the missing mechanism, then the resulting estimate using the imputed sample is consistent. On the contrary, misspecification of the imputation models can lead to potentially biased estimates because of the distortion of the variance-covariance matrices.

Inverse Probability Weighting

Inverse probability weighting assigns different weights to observed data points depending on their probability of being observed. Thus, the computation of IPW requires researchers to know the probability of being observed. Consider the case of MAR, where the probability of missing (or equivalently being observed) only depends on a set of observed variables x. Denote $p(x) \equiv P(s = 1|x) = P(s = 1|y, x, z)$, then we can solve the missing data problem by:

$$\min_{\alpha,\theta} \sum_{i=1}^{N} \left(\frac{s_i}{p(x_i)}\right) (y_i - \alpha - \theta z_i)^2.$$

In practice, $p(x)$ is often unknown except in some special cases, and thus we need to estimate it. To this end, we can regress the selection indicator s on x using flexible binary choice models, such as logit or probit, or even nonparametric models, and obtain the estimated selection probability (or alternatively called the propensity score) $\hat{p}(x)$.

Heckman's Correction for Selection Bias

We know from (IA2) that the OLS estimator $\hat{\theta}$ is biased because $E(y_i | s_i = 1, z_i) = \alpha_i + \theta z_i + E(\varepsilon_i | z_i, s_i = 1)$, and $E(\varepsilon_i | z_i, s_i = 1) \neq 0$ in general. Heckman's method assumes that the missing mechanism is determined by the following model:

$$s_i^* = \beta x_i + \eta_i, \quad i = 1, \dots, N, \tag{IA3}$$

where s_i^* is the latent variable associated with s_i, i.e., $s_i = 1$ if $s_i^* > 0$ and $s_i = 0$ if $s_i^* \leq 0$. Further, assume that the error terms in (IA3) is normally distributed with variance σ_η^2 and correlated with ε_i in (IA1), and their covariance is ρ; x and z are both exogeneous. The Heckman selection procedure approximates the "omitted variable" $(\varepsilon_i|z_i, s_i = 1)$ by its consistent estimate and includes this proxy in the regression to correct for the bias. In particular, based on the joint distribution of η_i and ε_i, one could write $E(\varepsilon_i|z_i, s_i = 1) = \sigma_\eta \rho \lambda(x_i\beta) = \gamma\lambda(x_i\beta)$, where $\lambda(x_i\beta)$ is the inverse Mills ratio defined by:

$$\lambda(x_i\beta) = \frac{\phi(-x_i\beta)}{1-\Phi(-x_i\beta)} = \frac{\phi(x_i\beta)}{\Phi(x_i\beta)}.$$

Then we can rewrite the conditional expectation of y_i given x_i and selection into the sample as:

$$E(y_i|x_i, s_i = 1) = \alpha + \theta z_i + \gamma\lambda(x_i\beta).$$

This leads to Heckman's two-step procedure.

Step 1: Estimate a probit regression $\text{P}(s_i = 1|x_i) = \Phi(x_i\beta)$ using all N observations and obtain the estimate $\hat{\beta}$. Then compute the inverse Mills ratio $\lambda\big(x_i\hat{\beta}\big)$.
Step 2: Estimate the regression $y_i = \alpha + \theta z_i + \gamma\lambda(x_i\hat{\beta})$ using OLS.

The estimates $\hat{\alpha}$, $\hat{\theta}$, and $\hat{\gamma}$ are consistent when x correctly includes all of the selection variables. The validity of Heckman's procedure also heavily relies on the distributional assumptions of the two errors, η_i and ε_i. For example, the deviation from the normality assumption of η_i may negatively affect the performance of the Heckman's procedure. Since γ captures the covariance between η_i and ε_i and a nonzero correlation implies selection bias, we can test whether selection is exogenous (or equivalently MCAR) by testing whether $\hat{\gamma} = 0$. For more extensions of Heckman's procedure, see Wooldridge (2002, Chapter 17).

Multiple Imputation

Multiple imputation (MI) is essentially an iterative version of stochastic imputation, which aims at explicitly modeling the uncertainty/variability ignored by the deterministic imputation procedures. Instead of imputing in a single value, multiple imputation uses the (joint) distribution of the observed data to estimate the parameters of interest multiple times to capture the uncertainty/variability in this imputation procedure. A general multiple imputation procedure consists of three steps:

Step 1. Imputation: Impute the missing data with their estimates and create a complete sample. Repeat this process multiple times.

Step 2. Estimation: For each complete sample, estimate the parameters of interest.

Step 3. Pooling: Combine the parameter estimates obtained from each completed data set.

The imputation method should be chosen depending on the type of variables with missing observations and the pattern of missingness. For example, MI with multivariate normal regressions can be applied to impute one or more continuous variables of arbitrary missing-value patterns; MI with chained equations employs a separate conditional distribution for each imputed variable and is often used to impute a variable with finite and discrete support (e.g., binary, multinomial, or count variable). We illustrate the MI with multivariate normal regressions (MI_MVN). As all MI methods, MI with multivariate normal regressions analyses the data in three steps: imputation, estimation, and pooling. We discuss the three steps in turn.

First, MI_MVN imputes the missing observations using data augmentation. In this case, we assume that the variable containing missing observations z is related with a set of (completely) observed variables x by:

$$z_i = \delta' x_i + v_i, \quad i = 1, \dots, N,$$

where $v_i \sim N(0, \sigma_v^2)$. Denote $w_i = (z_i, x_i)$. Data augmentation in this case is essentially an iterative Markov chain Monte Carlo (MCMC) procedure that iterates between two (sub-)steps, a replacement step and posterior step.

- Replacement step: We replace the missing values of z_i with draws from the conditional posterior distribution of z_i given observed variables and the values of model parameters in this iteration. Particularly, for each iteration t, we can replace the missing observations by:

$$z_i^{(t)} \sim P\left(z_i \middle| x_i, \delta^{(t-1)}, \sigma_v^{(t-1)}\right), \text{ for } \ i \in \{i | s_i = 1\}.$$

- Posterior step: We draw the new values of model parameters from their conditional posterior distribution given the observed data and imputed data from the previous replacement step.

$$\sigma_v^{(t)} \sim P\left(\sigma_v \middle| x_i, z_i^{(t)}\right), \text{ and } \ \delta^{(t)} \sim P\left(\delta \middle| x_i, z_i^{(t)}, \sigma_v^{(t)}\right),$$

  where $z_i^{(t)}$ is the imputed value from iteration t if it is missing and the original value if non-missing.

The conditional posterior distributions above are jointly determined from the prior distribution for the model parameter $P(\delta, \sigma_v)$, e.g., uniform, Jeffreys, or ridge, and the assumed normal distribution of the data. These two steps (replacement and posterior) are iterated until a specified number of iterations or there is numerical convergence.

Second, we estimate the regression of interest (IA1) with the imputed (pseudo-complete) data set using various approaches, e.g., OLS, HS. Since the imputation is conducted for multiple times, D times, we obtain multiple estimates for the same regression parameter θ.

Third, we combine/pool the estimates (coefficients and standard errors) across all imputed datasets and obtain a single statistic for each parameter. The final estimated slope coefficient $\hat{\theta}$ is simply an arithmetic mean of the corresponding estimate obtained from each of the imputed data. The variance of $\hat{\theta}$ is obtained by the total variance formula and is written by the average estimated variance of coefficient estimates across D imputed datasets plus the sample variance of coefficient estimates based on D imputations.

A major advantage of multiple imputation over deterministic imputation is that the final statistics appropriately reflect the uncertainty caused by imputation. If the joint normality is a reasonable assumption and the specification of x is correct (i.e., MAR), MI_MVN produces consistent estimates. In practice, a safe strategy is to include all observables in x to better approximate the posterior distribution.

TABLE IA1
Predictability of Unreported Innovation with Lasso

Table IA1 presents the OLS regression results for predictability of unreported innovation using only Lasso variables. Columns (1)-(4) *World* present the results for all countries in the sample. Columns (5)-(7) present the results for the *US* only. Panel A presents the results for unreported R&D, and Panel B presents the results for non-USPTO patent seeking firms. Standard errors are double clustered at firm and time level. T-statistics are presented in brackets. Variable definitions are presented in the Appendix. *, **, and *** represent significance at the 10%, 5%, and 1% levels, respectively. Adj. R^2 is the adjusted R^2.

Panel A. Unreported R&D

	World				US		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
ln(TOTAL_ASSETS)	-0.020***	-0.017***	-0.012***	-0.009***	0.040***	0.004*	-0.014***
	(-10.20)	(-8.31)	(-7.86)	(-3.54)	(13.28)	(1.73)	(-4.40)
STOCK_LIQUIDITY	-0.007***	-0.007***	-0.005***	-0.001***	-0.008***	-0.003***	-0.001***
	(-8.92)	(-9.99)	(-14.33)	(-3.47)	(-13.50)	(-6.23)	(-3.27)
PATENT_INTENSITY	-604.300***	-1.310	13.370	38.122**	-700.176***	-19.498***	-17.045*
	(-17.42)	(-0.11)	(1.06)	(2.02)	(-21.33)	(-3.08)	(-1.77)
Country FE			Yes				
Industry FE		Yes	Yes			Yes	
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Firm FE				Yes			Yes
N	300,634	300,634	300,634	328,734.0	77,982	77,982	76,944
Adj. R^2	0.13	0.23	0.38	0.80	0.23	0.53	0.93

Panel B. Non-USPTO Patent Seeking Firms

ln(TOTAL_ASSETS)	-0.012***	-0.009***	-0.020***	-0.006***	-0.000	-0.023***	-0.013***
	(-11.62)	(-9.41)	(-19.28)	(-6.88)	(-0.08)	(-9.67)	(-4.03)
STOCK_LIQUIDITY	-0.006***	-0.007***	-0.004***	-0.001***	-0.007***	-0.005***	-0.001***
	(-20.77)	(-21.86)	(-17.74)	(-5.11)	(-14.73)	(-12.24)	(-3.55)
PATENT_INTENSITY	-0.000**	-0.000**	-0.000**	-0.000	-0.000***	-0.000**	0.000
	(-2.09)	(-2.10)	(-2.22)	(-1.14)	(-3.22)	(-2.60)	(1.01)
R&D_STOCK	-368.647***	-30.191**	-25.181***	-4.871	-554.522***	-47.131***	-0.239
	(-17.34)	(-2.57)	(-3.23)	(-0.90)	(-18.37)	(-3.72)	(-0.02)
Country FE			Yes				
Industry FE		Yes	Yes			Yes	
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Firm FE				Yes			Yes
N	327,997	327,997	327,997	326,067	77,958	77,958	76,926
R2	0.09	0.15	0.24	0.76	0.15	0.32	0.78

.

TABLE IA2
Simulation Based on the Empirical Distribution from Compustat Data

Table IA2 provides the evaluation statistics, bias (relative bias over true parameter) and root mean squared error (RMSE) for the simulation based on the empirical distribution from Compustat (US) data, as described in Section VI.A of the paper. Bias presents the average of the absolute bias across all five variables and RMSE presents the average RMSE across the five variables. The empirical distribution is from the panel of 783 firms with non-missing information for all variables except R&D. The methods evaluated are listwise deletion (LD), imputation with zero (ImpZero), imputation with industry mean, two-digit SIC code (ImpMean), inverse probability weighting (IPW), Heckman selection procedure (HS), and multiple imputation (MI). The regressions for imputation with zero and industry mean include a dummy variable for the imputed observations. MI is spec uses all the variables in the regression and is estimated using MCMC with 200 iterations for convergence. We present results for three missingness mechanisms: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). Panel A presents the results for the missingness regression which includes the lasso variables. Panel B presents the results with the MI specification in Panel A as well as includes the Lasso variables in the Sales growth regression. Panel C presents the Double Lasso results. Variable definitions are presented in the Appendix. We generate missingness R&D for 50 and 70% of the sample. We conduct 500 simulations.

		Missing 70%						Missing 50%					
		LD	Imp. Zero	Imp. Mean	IPW	HS	MI	LD	Imp. Zero	Imp. Mean	IPW	HS	MI
Panel A. Missingness Regression with Q, A, V, and PI													
MCAR	Bias	0.80	0.24	0.22	3.69	3.67	0.11	0.63	0.16	0.13	3.42	3.36	0.05
	RMSE	0.05	0.02	0.02	0.09	0.09	0.02	0.03	0.02	0.02	0.07	0.07	0.02
MAR	Bias	1.02	0.13	0.12	20.22	118.27	0.10	0.63	0.17	0.15	17.43	100.29	0.07
	RMSE	0.07	0.02	0.02	0.53	2.94	0.02	0.04	0.02	0.02	0.45	2.87	0.02
MNAR	Bias	0.98	0.13	0.12	11.96	67.10	0.11	0.58	0.18	0.15	10.18	59.42	0.08
	RMSE	0.07	0.02	0.02	0.26	2.27	0.02	0.03	0.02	0.02	0.25	2.05	0.02
Panel B. Missingness Regression with Q, A, V, and PI and Sales growth regression with V and PI													
MCAR	Bias	0.86	0.26	0.24	3.74	3.75	0.17	0.60	0.23	0.15	3.52	3.48	0.08
	RMSE	0.05	0.02	0.02	0.09	0.09	0.02	0.04	0.02	0.02	0.07	0.08	0.02
MAR	Bias	1.10	0.25	0.23	18.96	100.69	0.09	0.61	0.18	0.17	16.26	98.45	0.06
	RMSE	0.08	0.02	0.02	0.48	2.73	0.02	0.04	0.02	0.02	0.39	2.50	0.02
MNAR	Bias	0.99	0.13	0.11	11.49	61.56	0.14	0.60	0.17	0.14	9.93	52.77	0.05
	RMSE	0.07	0.02	0.02	0.29	1.86	0.02	0.03	0.02	0.02	0.24	1.76	0.02

		Missing 70%						Missing 50%					
		LD	Imp. Zero	Imp. Mean	IPW	HS	MI	LD	Imp. Zero	Imp. Mean	IPW	HS	MI
Panel C. Double Lasso													
MCAR	Bias	0.77	0.24	0.22	3.77	3.74	0.08	0.60	0.21	0.19	3.59	3.47	0.09
	RMSE	0.05	0.02	0.02	0.09	0.09	0.02	0.04	0.02	0.02	0.07	0.07	0.02
MAR	Bias	0.66	0.18	0.15	15.51	5.76	0.05	0.48	0.19	0.17	16.46	4.50	0.09
	RMSE	0.03	0.02	0.02	0.45	0.40	0.02	0.02	0.02	0.02	0.43	0.35	0.02
MNAR	Bias	0.63	0.24	0.20	10.00	3.58	0.07	0.49	0.20	0.18	10.08	3.34	0.06
	RMSE	0.03	0.02	0.02	0.29	0.24	0.02	0.02	0.02	0.02	0.25	0.18	0.02

TABLE IA3
Simulation Based on Simulated Data

Table IA3 provides the evaluation statistics, bias (relative bias over true parameter) and root mean squared error (RMSE) for the simulation based on simulated data, as described in Section VI.C of the paper. The methods evaluated are listwise deletion (LD), imputation with zero (ImpZero), imputation with industry mean, two-digit SIC code (ImpMean), inverse probability weighting (IPW), Heckman selection (HS), and multiple imputation (MI). MI is estimated using MCMC with 200 iterations for convergence. The regressions for imputation with zero and industry mean include a dummy variable for the imputed observations. We present results for three missingness mechanisms: missing completely at random (MCAR) in Panel A, missing at random (MAR) in Panel B, and missing not at random (MNAR) in Panel C. We generate missingness in x1 for 50 and 70% of the sample. We conduct 500 simulations.

		Missing 70%						Missing 50%					
		LD	Imp Zero	Imp Mean	IPW	HS	MI	LD	Imp Zero	Imp Mean	IPW	HS	MI
Panel A. MCAR													
Bias	θ_1	0.00	-0.19	-0.19	0.00	0.00	-0.01	0.00	-0.13	-0.13	0.00	0.00	-0.01
	θ_2	0.01	0.28	0.28	0.01	0.01	0.01	0.00	0.19	0.19	0.00	0.00	0.00
RMSE	θ_1	0.11	0.21	0.21	0.08	0.11	0.09	0.06	0.07	0.07	0.06	0.06	0.05
	θ_2	0.11	0.29	0.29	0.08	0.11	0.09	0.06	0.10	0.10	0.07	0.06	0.05
Panel B. MAR													
Bias	θ_1	-0.15	-0.23	-0.23	-0.16	-0.08	-0.08	-0.11	-0.16	-0.16	-0.11	-0.09	-0.05
	θ_2	-0.12	0.12	0.12	-0.12	-0.08	-0.05	-0.08	0.04	0.04	-0.07	-0.06	-0.04
RMSE	θ_1	0.17	0.24	0.24	0.18	0.17	0.10	0.13	0.17	0.17	0.13	0.12	0.08
	θ_2	0.15	0.13	0.13	0.15	0.16	0.09	0.07	0.06	0.06	0.07	0.07	0.05
Panel C. MNAR													
Bias	θ_1	-0.17	-0.28	-0.28	-0.19	-0.13	-0.10	-0.13	-0.19	-0.19	-0.13	-0.11	-0.05
	θ_2	-0.16	0.14	0.14	-0.15	-0.13	-0.08	-0.11	0.04	0.04	-0.11	-0.10	-0.07
RMSE	θ_1	0.19	0.29	0.29	0.20	0.17	0.12	0.14	0.20	0.20	0.14	0.13	0.07
	θ_2	0.17	0.12	0.12	0.17	0.16	0.10	0.13	0.06	0.06	0.13	0.12	0.08

TABLE IA4
Imputation Effect on Empirical Inference

Table IA4 replicates the results in Fama and French (2002) using different imputation methods and two-way fixed effects. We present the results of a contemporaneous regression with two-way fixed effects: $\frac{L_t}{A_t} = \beta_0 + \beta_1 \frac{V_t}{A_t} + \beta_2 \frac{ET_t}{A_t} + \beta_3 \frac{Dp_t}{A_t} + \beta_4 RDD_t + \beta_5 \frac{RD_t}{A_t} + \beta_6 \ln(A_t) + e_t$. "ImpZero" presents the result for the sample with imputation with zero and an indicator variable, "LD" presents the results for listwise deletion, "MI" presents the results for multiple imputation implemented using all the variables in the regression in the imputation, "MI Lasso" presents the results for multiple imputation implemented using all the variables in the regression and the Lasso variables stock liquidity and industry patent intensity in the imputation, "Pseudo RD" presents the result using pseudo R&D as an explanatory variable, and "Text-based Innovation" presents the results for the analyst coverage based innovation variable (Bellstam et al., 2020). The dependent variable is book leverage $\frac{L_t}{A_t}$ at time *T*. $\frac{V_t}{A_t}$ is the market to book ratio, $\frac{ET_t}{A_t}$ is earnings before interest and taxes as a proportion of total assets, $\frac{Dp_t}{A_t}$ is depreciation as a proportion of total assets, $\frac{RD_t}{A_t}$ is the R&D expenses as a proportion of total assets, RDD_t is an indicator variable equal to 1 if R&D expenditure is missing and has been imputed with zero, and zero otherwise, *Pseudo R&D*$_t$ is an indicator variable equal to 1 if a firm applies for a patent in PATSTAT and has no reported R&D, and zero otherwise, *Text-based Innovation*$_t$ is the firm analyst-based innovation measure from (Bellstam et al., 2020), and $\ln(A_t)$ is the natural logarithm of total assets. Non-dividend payers include firms that do not pay dividend in year *T-1*. Panel A presents the results for the dividend paying firms and Panel B for the non-dividend paying firms. The sample period is 1965-1999. Standard errors are double clustered.

Panel A. Dividend Payer Firms

Variable	Imp Zero (1)	LD (2)	MI (3)	MI Lasso (4)	Pseudo R&D (5)	Text-based Innovation (6)
Intercept	0.305***	0.344***	0.366***	0.368***	0.300***	0.246***
	(22.62)	(19.83)	(56.52)	(55.24)	(22.13)	(3.94)
$\frac{V_t}{A_t}$	-0.001	-0.001	0.001	0.001	0.000	-0.006
	(-0.15)	(-0.47)	(0.40)	(0.60)	(-0.10)	(-1.29)
$\frac{ET_t}{A_t}$	-0.158**	-0.215**	-0.184**	-0.192	-0.157	-0.628
	(-1.99)	(-2.66)	(-2.07)	(-2.17)	(-1.99)	(-3.91)
$\frac{Dp_t}{A_t}$	-1.076***	-0.059	-1.057***	-1.048***	-1.049***	-0.797***
	(-6.12)	(-0.30)	(-10.67)	(-10.56)	(-6.05)	(-2.77)
RDD$_t$	0.070***				0.075***	
	(11.96)				(12.35)	
$\frac{RD_t}{A_t}$	-0.290***	-0.435***	0.081***	0.033	-0.290***	
	(-2.71)	(-4.54)	(3.91)	(1.48)	(-2.72)	
Pseudo R&D$_t$					-0.098***	
					(-12.43)	
Text-based Innovation$_t$						-0.009*
						(-1.78)
$\ln(A_t)$	0.041***	0.029***	0.038***	0.038***	0.042***	0.048***
	(29.95)	(13.61)	(96.36)	(95.10)	(30.14)	(6.89)

Panel B. Non-dividend Payer Firms

Variable	Imp Zero (1)	LD (2)	MI (3)	MI Lasso (4)	Pseudo R&D (5)	Text-based Innovation (6)
Intercept	0.325***	0.394***	0.376***	0.381***	0.323***	0.242
	(4.70)	(20.07)	(6.74)	(7.05)	(4.66)	(1.11)
$\frac{V_t}{A_t}$	0.027	-0.004***	0.028**	0.029***	0.027	-0.008
	(1.32)	(-3.24)	(2.24)	(2.30)	(1.32)	-1.40
$\frac{ET_t}{A_t}$	-0.517***	-0.301***	-0.139	-0.136	-0.517***	-0.404
	(-3.15)	(-4.77)	(-0.54)	(-0.52)	(-3.14)	-1.56
$\frac{Dp_t}{A_t}$	0.691	1.984***	0.636*	0.651*	0.692	1.725*
	(1.29)	(7.96)	(1.66)	(1.70)	(1.29)	1.84
RDD_t	0.079***				0.082***	
	(4.61)				(4.73)	
$\frac{RD_t}{A_t}$	-0.702***	-0.335***	0.955***	0.962***	-0.701***	
	(-2.83)	(-3.33)	(3.45)	(3.43)	(-2.83)	
Pseudo R&D$_t$					-0.134***	
					(-6.03)	
Text-based Innovation$_t$						-0.095***
						(-5.50)
$\ln(A_t)$	0.032***	0.013**	0.022***	0.024***	0.033***	0.042
	(4.54)	(2.60)	(4.98)	(5.35)	(4.62)	(1.60)