JOURNAL OF FINANCIAL AND QUANTITATIVE ANALYSIS Vol. 57, No. 6, Sept. 2022, pp. 2324–2354
© THE AUTHOR(S), 2021. PUBLISHED BY CAMBRIDGE UNIVERSITY PRESS ON BEHALF OF THE MICHAEL G. FOSTER SCHOOL OF BUSINESS, UNIVERSITY OF WASHINGTON
doi:10.1017/S0022109021000764

Deleting Unreported Innovation

Ping-Sheng Koh
ESSEC Business School
kohp@essec.edu

David M. Reeb
National University of Singapore and Asian Bureau of Finance and Economic Research
dmreeb@nus.edu.sg (corresponding author)

Elvira Sojli
University of New South Wales
e.sojli@unsw.edu.au

Wing Wah Tham
University of New South Wales
w.tham@unsw.edu.au

Wendun Wang
Erasmus School of Economics
wang@ese.eur.nl

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.

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),

The simulation code and R&D imputed data are available at www.wendunwang.weebly.com/research.html. An early version of this article 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, the University of Queensland, the University of South Carolina, the University of Toronto, Virginia Tech, Wilfrid Laurier University, University of Technology Sydney, and the 2020 Price College of Business (PCOB) Valuation & Accounting for Intangible Assets Workshop. Koh acknowledges financial support from the ESSEC Foundation and CY Initiative of Excellence.

https://doi.org/10.1017/S0022109021000764 Published online by Cambridge University Press

https://doi.org/10.1017/S0022109021000764 Published online by Cambridge University Press

Koh, Reeb, and Zhao (2018)). For instance, over 50% of U.S. 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, and attempts to keep the innovation information secret (Png (2017)). Recent articles develop additional measures of corporate innovation, ranging from the textual analysis of financial analyst reports (Bellstam, Cookson, and Bhagat (2021)) to firm disclosures of new products (Mukherjee, Singh, and Zaldokas (2017)). These new measures provide meaningful approaches to capturing 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 either focuses on innovation as 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 0 or the industry average, inverse probability weighting (IPW), Heckman selection (HS), 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 because of the nonrepresentativeness 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 U.S. 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

[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 Supplementary Materials for details).

https://doi.org/10.1017/S0022109021000764 Published online by Cambridge University Press

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 0, the industry average, and MI. Our tests show that the R&D in firms with unreported R&D significantly differs from that of zero-R&D firms, the average industry R&D, and that of 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 the actual R&D reported in subsequent financial statements.

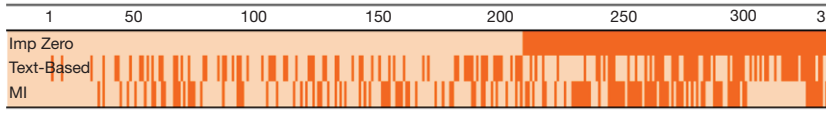
The absence of reported R&D in a firm could arise from a lack of innovation or as a disclosure choice. One potential issue with using recovered R&D is that these firms all engage in innovation. Consequently, we use an alternative counterfactual group that allows for both innovative and non-innovative firms. Specifically, we use the text-based ranking of corporate innovation in the S&P 500 firms from Bellstam et al. (2021). Analysts' discussions that describe corporate innovation arguably arise in firms that engage in R&D, whereas firms without such discussions are likely those without meaningful innovative activity. Again, we find that MI is the best solution for handling 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. (2021). 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, although only available for S&P 500 firms, provides a more complete distribution of corporate innovation. Similarly, the MI approach to handling unreported R&D also mitigates the clumping problem inherent in deleting these firms or classifying them as zero innovators.

Next, we focus on firms without patents. A commonly used metric to measure corporate innovation is patents from the U.S. Patent Office (USPTO). 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

FIGURE 1

Ranking Innovative Firms

Figure 1 shows the rank of innovative firms for the S&P 500 using R&D expenditures and text-based innovation. The figure shows innovation ranks based on the R&D expenditure imputed with 0 for unreported R&D (ImpZero), the text-based innovation measure, and multiple imputation (MI). Firms with reported R&D expenditures are shown in light gray, and those with unobserved R&D in dark gray.



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 U.S. firms, 69% of positive-R&D firms never file for patents using USPTO data, whereas 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 (DPGs), 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 MAR, 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 an HS 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 that 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

https://doi.org/10.1017/S0022109021000764 Published online by Cambridge University Press

https://doi.org/10.1017/S0022109021000764 Published online by Cambridge University Press

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, when researchers decide how to handle missing innovation data, they 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 (LD) 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 article 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 or MI, IPW, or HS 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 article 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 MI provides the most reasonable solution. Surprisingly, and across a wide variety of specifications and approaches, we find that both HS and IPW 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 a set of guidelines for unreported innovation in the conclusion.

[2]Our simulations use the common MI approach of Markov chain Monte Carlo (MCMC), which assumes the underlying data follow 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 simulations, 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 U.S. firms. The code for MI and the imputed R&D values are available at https://bit.ly/3qyd00X.

https://doi.org/10.1017/S0022109021000764 Published online by Cambridge University Press

II. Handling Unreported Innovation Data

There are numerous possible reasons that 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 of efficiency (Rubin (1976)). Bias stems from the nonrepresentativeness 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)), whereas research in machine learning also focuses on training and testing data that suffer 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 1 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 following linear relation:

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

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 the solutions to this problem depends on the nature of the 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. LD 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:

$$y_i = s_i\alpha + \theta s_i z_i + s_i\varepsilon_i, \tag{2}$$

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. IPW relies on assigning different weights to observed points depending on their probability of being observed. Because this probability is unknown for unreported innovation, we can estimate it using binary-choice or nonparametric models. A second approach is MI, which is essentially an iterative version of stochastic imputation and 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 the Supplementary Material for more details on these methods).

III. The Severity of Unreported Innovation

A. Data and Sample

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

Our sample selection begins with the Oct. 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 2 main steps. First, we standardize patent assignee names and firm names, focusing on unifying suffixes and dampening the noninformative 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, Sweden, Switzerland, Taiwan, the United Kingdom, and the United States. 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 of 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, and 7% of firms report R&D of 0. The 75th percentile of R&D-reporting firms has R&D that equates to roughly 6%

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

https://doi.org/10.1017/S0022109021000764 Published online by Cambridge University Press

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 report patent applications in any patent office, and "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 *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

Panel A. Sample Characteristics

Variables	*N*	Mean	Median	Std. Dev.	25th	75th
	1	2	3	4	5	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	Report R&D	Report Patent	Report R&D and Patent	Differences		
	1	2	3	4	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			

of total assets. In addition, the sample firms invested an average of 6% of total assets in capital expenditure. Firms have an average of 9 patent applications, 4 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 median leverage of 52%. In our analysis, we focus on patent applications because these capture the R&D activity happening around the firm, but we 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 2 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.

https://doi.org/10.1017/S0022109021000764 Published online by Cambridge University Press

(2020), the lasso approach identifies total assets, stock liquidity, and industry patent intensity as relevant predictive variables for unreported R&D and identifies 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 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 of 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, whereas 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, whereas 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 MCAR 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 χ^2 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

https://doi.org/10.1017/S0022109021000764 Published online by Cambridge University Press

TABLE 2
Testing MCAR

Table 2 presents the missing completely at random (MCAR) test for the predictability of unreported innovation. The test is based on the Little (1988) test for MCAR. Columns 1–4 ("World") present the results for all countries in the sample. Columns 5–7 ("U.S.") present the results for the United States only. "DoF" is the number of degrees of freedom, and "Prob. $>\chi^{2n}$ is the probability of the null hypothesis that the data is MCAR. Variable definitions are presented in the Appendix.

	World				U.S.		
	1	2	3	4	5	6	7
R&D ($millions)	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
DoF	2	9	87	326	5	23	180
Prob. $>\chi^2$	0.00	0.00	0.00	0.00	0.00	0.00	0.00

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 U.S. samples, 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 U.S. firms. The chosen characteristics are based on innovation theory and the lasso approach described previously. 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 0 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

[5]We report the results using least squares estimation because it allows us to easily incorporate multilevel 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 from the authors.

https://doi.org/10.1017/S0022109021000764 Published online by Cambridge University Press

TABLE 3

Predictability of Unreported R&D

Table 3 presents the OLS regression results for the predictability of unreported R&D. Columns 1–4 ("World") present the results for all countries in the sample. Columns 5–7 ("U.S.") present the results for the United States only. Standard errors are double clustered at the firm and year levels. *t*-statistics are presented in parentheses. Variable definitions are presented in the Appendix. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

	World				U.S.		
	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 fixed effects (FE)	No	No	Yes	No	No	No	No
Industry FE	No	Yes	Yes	No	No	Yes	No
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Firm FE	No	No	No	Yes	No	No	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

year or year and industry fixed effects explains 13%–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, whereas it decreases with ln(TOTAL_ASSETS), STOCK_LIQUIDITY, and PATENT_INTENSITY. For U.S. firms (columns 5–7), industry and year fixed effects explain 53% of the variation in unreported R&D, whereas firm and year fixed effects explain 93% of this variation. Unreported R&D increases with PPE, ROA, and LEVERAGE for U.S. firms, whereas it decreases with PATENT_INTENSITY. Thus, the results are inconsistent with absent R&D arising MCAR.

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 0 otherwise. USPTO patents are the benchmark in this analysis because they are widely used to measure both U.S. and non-U.S. 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 it decreases with ln(TOTAL_ASSETS) and STOCK_LIQUIDITY. Focusing on just the subset of U.S. firms, firm and industry characteristics explain a substantial amount of the variation in absent patents (32%–77%; columns 6 and 7). Unreported patents increase with ROA while decreasing with ln(TOTAL_ASSETS) and STOCK_LIQUIDITY for U.S. firms. Collectively, the evidence in this section indicates a

https://doi.org/10.1017/S0022109021000764 Published online by Cambridge University Press

TABLE 4

Predicting Non–Patent-Seeking Firms

Table 4 presents OLS regressions of unreported U.S. Patent and Trademark Office (USPTO) patents and explanatory variables. The dependent variable is an indicator variable equal to 1 when a firm does not have USPTO patents, and 0 otherwise. Columns 1–4 ("World") present the regression results for all countries in the sample. Columns 5–7 ("U.S.") present the regression for U.S.-listed firms only. Standard errors are double clustered at the firm and year levels. *t*-statistics are presented in parentheses. Variable definitions are presented in the Appendix. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

	World				U.S.		
	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 fixed effects (FE)	No	No	Yes	No	No	No	No
Industry FE	No	Yes	Yes	No	No	Yes	No
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Firm FE	No	No	No	Yes	No	No	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

significant correlation between the absence of patents and firm-specific factors. Again, the results are inconsistent with innovation MCAR. Taken together, these results suggest that unreported R&D and patents are MNAR.[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.

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

[6]The results in Table IA1 of the Supplementary Material show that focusing only on the lasso inferred variables does not qualitatively affect the inference.

https://doi.org/10.1017/S0022109021000764 Published online by Cambridge University Press

https://doi.org/10.1017/S0022109021000764 Published online by Cambridge University Press

comparisons by users of general-purpose financial statements (Financial Accounting Standards Board (FASB) Accounting Standards Codification (ASC) 205-10-45; FASB (n.d.)). 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 for investigating how close the imputed estimates from various imputation methods are to the "recovered" true values.

Using the sample of U.S. firms for the period 1992–2016, 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 (10-Ks) 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 2 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 2 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 1 year before the switch, whereas others report for 2 years).

B. Comparing Recovered-R&D Firms to Zero-R&D Firms

We begin our analysis by comparing the characteristics of recovered-R&D firms with those of 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 those of the zero-R&D firms. In addition, the recovered firms differ from zero-R&D firms across several different dimensions, such as total assets, PPE, and leverage.

The R&D expenditure of recovered firms is significantly lower than that of positive-R&D firms, but the R&D expenditures of the 2 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, the results in Panel A of Table 5 show that firms with unreported R&D expenditure 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.

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 their comparisons with different imputation methods. Panel A presents a comparison of recovered-R&D, zero-R&D, and positive-R&D firm characteristics. Panel B presents a 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*-statistic presents its difference from 0. IMPUTED_R&D_INDUSTRY_AVG is the average industry expenditure (2-digit SIC codes) for the observations that are recovered; IMPUTED_R&D_MI_FULL_SAMPLE is the multiply imputed R&D using only the lasso variables of ln(TOTAL_ASSETS), STOCK_LIQUIDITY, and PATENT_INTENSITY by industry (2-digit SIC codes) for the complete sample; IMPUTED_R&D_ MI_SUBSAMPLE is the multiple-imputation (MI) R&D using the same MI model on the restated R&D subsample and the industry- and size-matched peers. "Diff." is the difference between Recovered R&D and imputed R&D. "*t*-Stat." represents the *t*-statistic for the difference between Recovered R&D and imputed R&D. Panel C shows the rank correlation among text-based innovation, patent applications (U.S. Patent and Trademark Office (USPTO) only and PATSTAT), R&D_EXPENDITURE, and MI. *, **, and *** indicate 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 ($millions)	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

Mean	Std. 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_SUBSAMPLE	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***

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 R&D, with industry-average R&D (2-digit SIC codes), 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 2 samples for the R&D MI. First, we base our MI estimation on the whole U.S. sample for the period 1992–2016 (not just the recovered-R&D sample): MI_FULL_SAMPLE. Second, we base the MI only on the sample of recovered-R&D firms and firms matched within industry and size quartile: MI_SUBSAMPLE. Our imputation is based on the 3 variables determined with adaptive lasso (total assets, stock liquidity, and industry patent intensity) and estimated by industry (2-digit SIC codes). 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 0; that is, replacing with 0 underestimates the recovered-R&D values. In terms of the dollar amount of R&D, the average

https://doi.org/10.1017/S0022109021000764 Published online by Cambridge University Press

recovered R&D is $6.91 million, whereas imputing with the industry average gives an estimate of $77.86 million. The industry-average imputed value is over 10 times the actual R&D spending and significantly different from the recovered value. On the other hand, the two MI methods generate averages of $6.36 million and $8.66 million, respectively, 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. (2021)). Text-based innovation measures firm innovation using latent Dirichlet allocation (LDA) analysis for analyst reports of the S&P 500 firms. This measure is standardized with a mean of 0 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 R&D of 0, 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 2 sets of counterfactuals: First, we use the product-innovation measure of Mukherjee et al. (2017), and second, we use the non-USPTO patents of U.S. 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

https://doi.org/10.1017/S0022109021000764 Published online by Cambridge University Press

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 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 75th percentile in the respective calendar year for each firm. Table 6 presents a comparison between two different MI methods. MI M1 includes only the lasso variables of total assets, stock liquidity, R&D stock, and industry patent intensity using PATSTAT patents; MI M2 is the MI 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 of Table 6 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. Approximately 6% of the sample has no USPTO patents but announces new

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 are from Mukherjee et al. (2017), and patents are based on U.S. Patent and Trademark Office (USPTO) data. NEW_PRODUCTS includes 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 a comparison of single- and multiple-imputation (MI) methods for patents with different product announcements. Single imputation includes imputation with 0 ("Impute 0") and imputation with the 2-digit industry average ("Impute industry average"). MI M1 presents the multiply imputed USPTO patents using the lasso variables of ln(TOTAL_ASSETS), STOCK_LIQUIDITY, R&D_STOCK, and PATENT_INTENSITY from PATSTAT patents, and MI M2 is the same as M1 but 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." presents the *t*-statistics for the difference.

	No Patents + New Products	No Patents + No New Products	Patents + New Products	Patents + No New Products	Diff.	
	1	2	3	4	5 = 1 − 4	*t*-Stat.
Panel A. Patents and New Products						
No. of obs. (% of sample)	5.71%	75.58%	3.10%	13.83%		
R&D ($millions)	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 0	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

https://doi.org/10.1017/S0022109021000764 Published online by Cambridge University Press

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 $\geq$ 40.18).

Panel B of Table 6 presents the single imputation and MIs for the four categories described previously 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 0 or the industry-average results in values statistically different from the counterfactuals of firms with patents and no new products ($|t$-statistics$| \geq 4.00$). In contrast, imputing with MI results in patent numbers that are not statistically different from those of firms with patents and no new products ($|t$-statistics$| \leq 1.26$).

B. Firms Without Patents: Global Patents as Counterfactuals

Innovation-related studies across accounting, economics, and finance mainly use data from the USPTO–NBER data set. 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 U.S. firms applied for USPTO patents in the sample period, and 9,518 U.S. firms received patents abroad, whereas 1,676 non-U.S. firms received USPTO patents, and 1,758 non-U.S. firms received non-USPTO patents. U.S. firms are granted, on average, 22 patents a year outside the United States and 28.2 patents a year in the United States.

As an alternative counterfactual, we use the sample of U.S. 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 0, industry mean (2-digit SIC code), and MI. MI is carried out with all the variables in Table 4, by industry (2 digits), and the lasso variables of total assets, stock volume, patent intensity, and R&D stock. The results in Figure 2 show that U.S. patents abroad are not equal to 0, 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 (U.S.) 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 DGPs. In each case, we compare six methods to handle missing values.

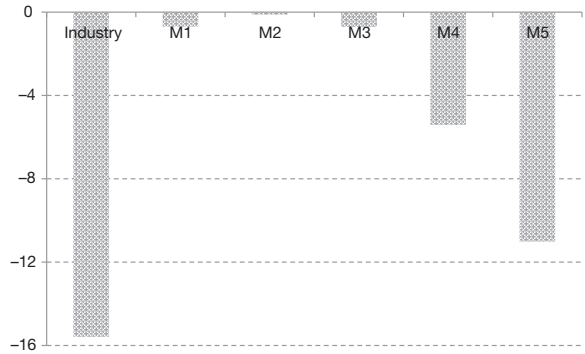
First, we consider LD. Second, we impute the missing R&D expenditure with 0 (ImpZero). Third, we impute the missing R&D by the industry average (ImpMean). Specifically, if an observation of firm i at time t is missing, we impute the missing observation by the industry average (2-digit SIC code) for the 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 the HS procedure with the selection variables containing all the observed covariates W. The HS procedure first predicts firms' selection probabilities by W, then corrects the selection bias by including a transformation of these predicted probabilities as an

https://doi.org/10.1017/S0022109021000764 Published online by Cambridge University Press

FIGURE 2

Imputation of Missing Patents

Figure 2 presents the *t*-statistics of the comparison of years with only non–U.S. Patent and Trademark Office (USPTO) patents for U.S. 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 that 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 (2-digit); M2 is the multiply imputed non-USPTO patents using the same model as M1 but without sales growth and leverage as conditioning information; M3 is the multiply imputed non-USPTO patents using the same model as M1 but with the addition of R&D expenditure as conditioning information; M4 is the multiply imputed non-USPTO patents using the lasso variables of 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.



additional explanatory variable. Next, we consider the IPW method that weights each *i*th observed point by the inverse of its conditional selection probability in least squares estimation. We use the standard package of the HS and IPW procedures in Stata. Finally, we consider MI. Because 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 an 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 RMSE of the 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 $\widehat{\theta}$, respectively, as follows:

$$\text{B}\left(\widehat{\theta}\right) = \frac{\frac{1}{R}\sum_{r=1}^{R}\left|\widehat{\theta}^{r} - \theta^{0}\right|}{\theta^{0}}, \tag{3a}$$

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

where $\widehat{\theta}^{r}$ is the estimate in the *r*th replication, $\text{var}\left(\widehat{\theta}^{r}\right)$ is the estimated robust variance of $\widehat{\theta}^{r}$, θ^{0} is the true value of the parameter, and 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

https://doi.org/10.1017/S0022109021000764 Published online by Cambridge University Press

the RMSE. Based on the observed missingness of R&D and patents in the United States, 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 nonmissing 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 logarithm 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 ($\widehat{\theta}^{r}$) 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 logarithm of total assets, Tobin's Q, leverage, and ROA as explanatory variables. This approach enables us to obtain a clean set of benchmark coefficients that is free from researcher intervention, except for the balanced, nonmissing 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 in the following discussion, 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 0 otherwise. Once we model and assign missing R&D observations, we can obtain the simulated R&D because the data are complete, and the nonmissing observations are given by their original values.

To create a missing indicator for R&D, we consider the three missing mechanisms: MCAR, MAR, and MNAR. Let α_i be the individual firm effects, and let $\widetilde{\eta}_{it}$ denote an idiosyncratic error. The three missing patterns can be summarized as follows:

- MCAR:

$$M_{it} = \widetilde{\eta}_{it}; \tag{4}$$

https://doi.org/10.1017/S0022109021000764 Published online by Cambridge University Press

https://doi.org/10.1017/S0022109021000764 Published online by Cambridge University Press

• MAR:

$$M_{it} = \alpha_i + \beta_O' X_{it}^O + \widetilde{\eta}_{it}; \tag{5}$$

• MNAR:

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

where X_{it}^O contains variables observed by researchers, whereas X_{it}^U is unobserved and only appears in the DGP but is omitted in imputation models. For MAR, we consider $X_{it}^O = (A_{it}, V_{it}, \mathrm{PI}_{it})'$, 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)); that is, $\alpha_i = c + \gamma_O' \overline{X}_i^O + u_i$ in MAR, and $\alpha_i = c + \gamma_O' \overline{X}_i^O + \gamma_U' \overline{X}_i^U + u_i$ in MNAR, where $\overline{X}_i^O = 1/T \sum_{t=1}^T X_{it}^O, \overline{X}_i^U = 1/T \sum_{t=1}^T X_{it}^U$, 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 follows:

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

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

where $\eta_{it} = \widetilde{\eta}_{it} + u_i$. Because 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 $\widehat{c}$, $\widehat{\beta}_O$, $\widehat{\beta}_U$, $\widehat{\gamma}_O$, and $\widehat{\gamma}_U$ as the true parameters to generate the missing probability M_{it}^* in the complete subsample of the data:

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

Φ is the normal cumulative distribution function (CDF), and pm_{it} is obtained for the three scenarios as follows:

1. MCAR:

$$pm_{it} = \eta_{it}; \tag{10}$$

2. MAR:

$$pm_{it} = \widehat{c} + \widehat{\beta}_O X_{it}^O + \widehat{\gamma}_O \overline{X}_i^O + \eta_{it}; \tag{11}$$

3. MNAR:

$$pm_{it} = \widehat{c} + \widehat{\beta}_O X_{it}^o + \widehat{\beta}_U X_{it}^u + \widehat{\gamma}_O \overline{X}_i^O + \widehat{\gamma}_U \overline{X}_i^u + \eta_{it}, \tag{12}$$

where $\eta_{it} \sim \text{IID}N\left(0,\sigma_{\eta}^{2}\right)$, 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 $M_{it}^{*}>\text{Q}_{\tau}\left(M_{it}^{*}\right)$, where $\text{Q}_{\tau}\left(M_{it}^{*}\right)$ is the τth 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:

$$S_{it}=\mu_{i}+\delta'\text{RD}_{it}+\theta' Z_{it}+\varepsilon_{it}, \tag{13}$$

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 ε_{it} is the error term. Note that intangible assets are not observed and thus are also not included in the DGP of S. The firm fixed effects μ_i are generated by $\mu_i=0.1\iota'\overline{Z}_i$, where ι is a 4×1 vector of 1s. and $\overline{Z}_i=1/T\sum_{t=1}^{T}Z_{it}$, and thus μ_i is correlated with sales-growth determinants. To obtain the parameters for δ' and θ', we estimate equation (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 $\varepsilon_{it}=\widetilde{\varepsilon}_{it}+\overline{Q}_i$ in MAR and $\varepsilon_{it}=\widetilde{\varepsilon}_{it}+0.5\left(\overline{Q}_i+\overline{I}_i\right)$ in MNAR. Here, $\overline{Q}_i$ and $\overline{I}_i$ 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. $\widetilde{\varepsilon}_{it}\sim\text{IID}N\left(0,\sigma_{\varepsilon}^{2}\right)$, 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 missingness (50% and 70%). When R&D is MCAR, 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 because of 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 also exhibits relatively smaller RMSEs than the other methods. Deterministic imputation methods (ImpZero and ImpMean) generate double the bias in MIs and RMSEs that are similar to those of MI. Still, MI has both the lowest average bias and lowest RMSE under MCAR.

Panel B of Table 7 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

[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).

https://doi.org/10.1017/S0022109021000764 Published online by Cambridge University Press

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 (U.S.) data, as described in Section V.A. The empirical distribution is from the panel of 783 firms with nonmissing information for all variables except R&D. The methods evaluated are listwise deletion (LD), imputation with 0 (ImpZero), imputation with industry mean (2-digit SIC code) (ImpMean), inverse probability weighting (IPW), Heckman selection procedure (HS), and multiple imputation (MI). MI uses total assets, stock liquidity, and industry patent intensity identified using lasso analysis in the regression and is estimated using Markov chain Monte Carlo (MCMC) with 200 iterations for convergence. The regressions for imputation with 0 and industry mean include a dummy variable for the imputed observations. "Avg. absolute bias" 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	ImpZero	ImpMean	IPW	HS	MI	LD	ImpZero	ImpMean	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. absolute 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
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. absolute 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. absolute 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

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, whereas the biases in IPW and HS are over 170 times and 80 times greater, respectively, than that of 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 6 times larger than the bias in MI. Focusing on RMSE, once again, ImpZero, ImpMean, and MI all exhibit similar magnitudes.

https://doi.org/10.1017/S0022109021000764 Published online by Cambridge University Press

https://doi.org/10.1017/S0022109021000764 Published online by Cambridge University Press

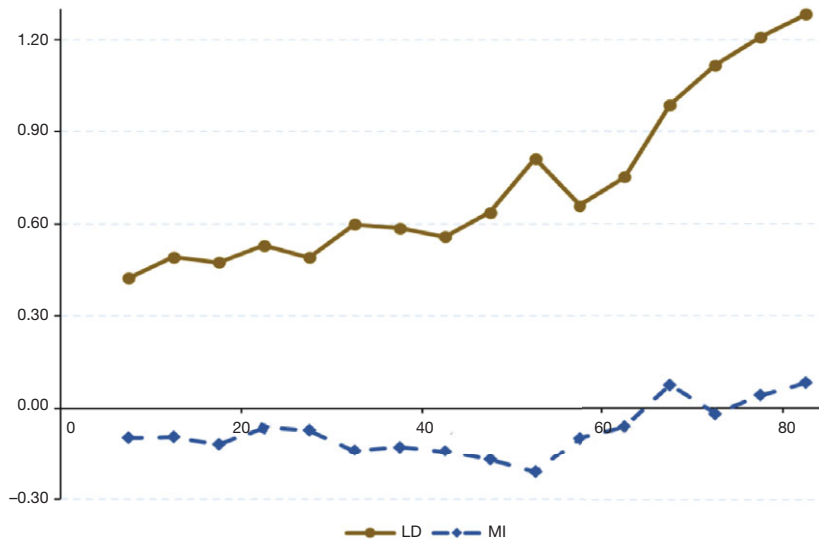
In the Supplementary Material, we also investigate other settings for both the determinants of R&D missingness in equations (11) and (12) and the sales-growth DGP in equation (13). We introduce a larger set of conditioning variables in equations (11) and (12) in Panel A of Table IA2, a larger set of 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 1–4, 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

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 nonmissing information for all variables except R&D. MI uses all the variables in the regression in Section V.A and is estimated using Markov chain Monte Carlo (MCMC) with 200 iterations for convergence. We present results for data missing at random (MAR). We conduct 500 simulations.



[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 previously and obtain the selected covariates denoted as X_{SG}. We use the union of X_{RD} and X_{SG} as variables in MI.

when 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 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.

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 using the USPTO patent data distribution. Table 8 presents the results of the simulation based on the patent empirical distribution. Under MAR,

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 (U.S.) and U.S. Patent and Trademark Office (USPTO) data. The empirical distribution comes from the panel of 783 firms with nonmissing information for all variables except USPTO patents. The methods evaluated are listwise deletion (LD), imputation with 0 (ImpZero), imputation with industry mean (2-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 Markov chain Monte Carlo (MCMC) with 200 iterations for convergence. The regressions for imputation with 0 and industry mean include a dummy variable for the imputed observations. "Avg. absolute bias" 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	ImpZero	ImpMean	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. absolute 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 B. 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. absolute 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

https://doi.org/10.1017/S0022109021000764 Published online by Cambridge University Press

IPW and HS generate the highest biasness in the coefficient estimates relative to both imputation and deletion. Focusing on MNAR, deterministic imputation and MI both perform better than the 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 without patents is roughly 8 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:

$$Y_i = z_{1i}\theta_1 + z_{2i}\theta_2 + \varepsilon_i,\ i = 1,\ldots,N, \tag{14}$$

where $\theta_1 = \theta_2 = 1$, $\varepsilon_i \sim \text{IID}N(0,1)$, and $N = 1{,}000$. The two covariates z_{1i} and z_{2i} are generated by a multivariate normal distribution with the unit means and variance–covariance matrix specified later. z_{1i} contains missing observations, whereas z_{2i} is completely observed. Let M_i be the missing indicator of x_{1i} that equals 1 if x_{1i} is missing, and 0 otherwise, which is determined by $M_i = 1\left[M_i^* > \text{Q}_\tau(M^*)\right]$, where $1[\cdot]$ is an indicator function, M_i^* is a latent variable, and $\text{Q}_\tau(M^*)$ is the τth quantile of M^*. We consider two values of $\text{Q}_\tau(M^*)$, 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: $M_i^* = x_{1i}\gamma_1 + x_{2i}\gamma_2 + \eta_i$, and
- 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, whereas 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 covariates $\{z_1,z_2,x_1,x_2,x_3,x_4,\varepsilon,\eta\}$ as follows:

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

https://doi.org/10.1017/S0022109021000764 Published online by Cambridge University Press

https://doi.org/10.1017/S0022109021000764 Published online by Cambridge University Press

Note that all covariates $\{z_1, z_2, x_1, x_2, x_3, x_4, \varepsilon, \eta\}$ are correlated with each other. The 2 error terms are correlated with each other, but they are independent of the observed covariates. In MNAR, the missingness is also driven by 2 unobservables, which may be correlated with the errors. Hence, the unobserved selection variable x_3 is uncorrelated with both errors, and x_4 is correlated with ε.

2. Results

Table IA3 in the Supplementary Material 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 biases and small RMSEs for both explanatory variables. There are only marginal differences across LD, HS, IPW, and MI. However, the two deterministic imputation methods (using 0 and the industry average) produce the largest biases and RMSEs.

Panel B of Table IA3 shows the results for MAR, where LD 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 0s or industry means increases the bias in θ_1 from –19% to –23% but decreases the bias in θ_2 from 28% to 12%. The HS procedure exhibits one of the smallest biases, which is comparable to that of MI, but at the cost of variance. This is reflected in the large RMSE of the HS estimates, suggesting that the 2-step procedure is rather inefficient. In contrast, MI performs well: Despite the increase in biasness, its bias continues to be among the lowest. The bias of MI is around half as large as that of LD and almost 2 times smaller than that of imputation using 0s or means, and MI has the smallest RMSE.

For MNAR, all methods produce biased estimates because of the non-random missing pattern, but the degree of bias differs substantially across methods (Panel C of Table IA3). Imputation using 0s or means leads to the largest bias and RMSE for θ_1 among the six methods (θ_1 bias is approximately 28%), whereas the bias and RMSE for θ_2 are generally in the middle of the six methods (θ_2 bias is 14%). The biasness in the LD and IPW methods also deteriorates in comparison to the MAR setting, leading to a downward bias of more than 17% and 15% in θ_1 and θ_2, respectively. Both the bias and RMSE for the HS procedure deteriorate by 62.5% in comparison to MAR (the largest deterioration among the six methods). Despite the observed deterioration compared with MAR, MI continues to produce the lowest biases and RMSEs among the six methods. Replicating our analysis across different levels of missingness reveals that MI consistently exhibits the smallest bias. In contrast, the bias from LD 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 (FF02) (2002) 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 because the 2 proxies for investment have opposite signs: The market-to-book ratio is positively correlated with leverage, and R&D expenditures are negatively correlated with leverage. For expositional simplicity, we focus this analysis on comparing MI 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 the FF02 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 an R&D of 0, and they include a dummy variable equal to 1 to differentiate firms with unreported R&D from firms that report R&D of 0. We estimate the leverage regression (equation (15)) to evaluate if leverage differs across firms in the manner predicted by the trade-off or pecking-order model using three approaches (LD, zero imputation with a missing dummy, and MI) and compare the resulting estimates:

$$(15) \qquad \frac{L_t}{A_t} = \beta_0 + \beta_1 \frac{V_t}{A_t} + \beta_2 \frac{\mathrm{ET}_t}{A_t} + \beta_3 \frac{Dp_t}{A_t} + \beta_4 \mathrm{RDD}_t + \beta_5 \frac{\mathrm{RD}_t}{A_t} + \beta_6 \ln(A_t) + e_t.$$

We follow FF02 in the choice of the sample period, variables of interest, and notation. ET_t/A_t, the ratio of annual pre-interest pretax earnings to end-of-year total assets, is a proxy for the expected profitability of assets in place.[9] V_t/A_t, the ratio of a firm's total market value to its book value, is a proxy for expected investment opportunities. 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 0. RDD_t is a dummy variable equal to 1 for unreported R&D, and 0 otherwise. Dp_t/A_t, the ratio of depreciation expense to assets, serves as a proxy for nondebt 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 in the Supplementary Material replicates FF02 using a contemporaneous regression with 2-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 presents results for non–dividend-payer firms. We use listwise deletion (LD), multiple imputation with only the variables in the regression (MI), multiple imputation with the lasso variables of 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 LD reported in columns 1 and 2 are negative and differ considerably from the estimates using MI (columns 3 and 4), which are positive. This suggests that the inferences made by the researcher in innovation could be driven by how the researcher chooses to deal with missing innovation data.

[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, ET_t/A_t is a measure of profitability when we look at tax effects in the trade-off model.

https://doi.org/10.1017/S0022109021000764 Published online by Cambridge University Press

https://doi.org/10.1017/S0022109021000764 Published online by Cambridge University Press

Using alternative measures of innovation like pseudo-R&D and text-based innovation leads to results similar to those of zero imputation and LD. 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 MI 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, and do not 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), and Koch et al. (2021)). Instead of unreported innovation arising randomly, we document that it 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 for handling 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 for dealing with missing innovation across different data sets, countries, and research settings. Our results using U.S. 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, MI 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 R&D expenditures of 0, imputing missing innovation data as 0 will provide a reasonable solution. Yet MI still performs well in this scenario.

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 because of 0 (near 0) 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 the 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 MCAR. 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 stem 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 MI (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 HS correction and MI remain appealing, with both approaches involving assumptions and trade-offs. Surprisingly, HS and IPW 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 samples.

https://doi.org/10.1017/S0022109021000764 Published online by Cambridge University Press

https://doi.org/10.1017/S0022109021000764 Published online by Cambridge University Press

Appendix. Variable Definitions

Variable Names	Notation	Variable Definitions	Code
R&D_EXPENDITURE	RD	R&D expenditure divided by total assets	XRD/AT
R&D ($millions)		R&D expenditure in million USD	XRD
REPORT_R&D		Indicator variable: 1 if a firm reported 0 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 logarithm of total assets	ln(AT)
CAPEX		Capital expenditure divided by total assets	CAPX/AT
ROA	*R*	Earnings before interest and taxes (EBIT) divided by total assets	EBIT/AT
SALES_GROWTH	*S*	Annual sales growth	$(Sale_t - Sale_{t-1})/ Sale_{t-1}$
INTANGIBLE_ASSETS	*I*	Assets that are not physical in nature	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 2-digit SIC codes across countries, unless specified otherwise	
R&D_STOCK		R&D accumulated over 10 years and 15% yearly depreciation	
RECOVERED_R&D		The recovered R&D expenditure as reported in 10-K filings for firms that switch from not reporting R&D to reporting R&D	
IMPUTED_R&D_INDUSTRY_AVG		The average industry R&D expenditure (2-digit) for recovered R&D observations	
IMPUTED_R&D_MI_FULL_SAMPLE		The R&D imputed using multiple imputation with only the Lasso variables: ln(TOTAL_ASSETS), STOCK_LIQUIDITY, and PATENT_INTENSITY, by industry (2-digit) for the complete sample	
IMPUTED_R&D_MI_SUBSAMPLE		R&D imputed using multiple imputation mode l in IMPUTED_R&D_MI_FULL_SAMPLE for the sample of recovered R&D firms and their industry and size matched peers	
NEW_PRODUCTS		The sum of all positive cumulative abnormal returns of material firm's product announcements over the year captured through Factiva, from Mukherjee et al. (2017)	
MAJ_NEW_PRODUCTS		The number of material firm's product announcements (as captured through Factiva) with the cumulative abnormal returns above the 75th percentile year by year after adjusting for firm size and book-to-market ratio, from Mukherjee et al. (2017)	
PATENT_USPTO		Patents applied by firm through USPTO	
PATENT_PATSAT		Patents applied by firm through all patent offices collected by PATSTAT	

Supplementary Material

To view supplementary material for this article, please visit http://doi.org/10.1017/S0022109021000764.

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), 357–374.

Bellstam, G.; A. Cookson; and S. Bhagat. "A Text-Based Analysis of Corporate Innovation." *Management Science*, 67 (2021), 4004–4031.

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*, Z. Griliches and M. D. Intrilligator, eds. Amsterdam, Netherlands: North-Holland (1984).

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." *Accounting Review*, 95 (2020), 177–204.

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. ASC 205-10-45. *Accounting Standards Codification.* Available at http://asc.fasb.org/ (n.d.).

Grangier, D., and I. Melvin. "Feature Set Embedding for Incomplete Data." In *Advances in Neural Information Processing Systems Conference Proceedings*, J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, eds. Red Hook, NY: Curran Associates (2010).

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 (2020).

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.

https://doi.org/10.1017/S0022109021000764 Published online by Cambridge University Press