# Deleting Unreported Innovation*

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

David M. Reeb: National University of Singapore, Senior Fellow: *ABFER;* dmreeb@nus.edu.sg

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, Erasmus University; wang@ese.eur.nl

April 23, 2020

**Abstract**

Innovation variables exhibit high rates of unobservability, often leading empirical studies to exclude firms that fail to report innovation. We assess the reliability of six methods for dealing with unobserved innovation using future disclosures about historical activity. Our tests reveal that deleting firms without reported innovation or imputing them as zero innovators and including a dummy variable leads to biased parameter estimates for innovation and other *explanatory variables*. We also replicate an influential study, demonstrating the economic significance of the bias introduced by different approaches to handling unreported innovation. Our analysis suggests using both multiple imputation and instrumental variables estimates.

*Keywords: Bias, Listwise Deletion, Innovation, Measuring Innovation, Multiple Imputation, Non-Patenting firms, Unreported R&D, Patents.*

* An early version of this manuscript circulated under the title "Missing Innovation Around the World." We are grateful to seminar participants at City University of Hong Kong, Hong Kong Polytechnic University, Louisiana State University, Maastricht University, the National University of Singapore, Rotterdam School of Management, Temple University, University of Queensland, University of South Carolina, University of Toronto, Virginia Tech, Wilfrid Laurier University, and the University of Technology Sydney. We appreciate discussions and advice from Renee Adams, Sumit Agarwal, Benito Arrunda, Richard Boylan, Lora Dimitrova, Bronwyn Hall, Gilles Hilary, Yael Hochberg, Qin Li, Gustavo Manso, Jiaming Mao, Naci Mocan, Randall Morck, Ivan Png, Wenlan Qian, Amit Seru, Vijay Singal, Xuan Tian and Rosemarie Ziedonis.

## 1. Introduction

Investors and academics exhibit a keen interest in understanding how corporate innovation influences firm growth and performance (Hochberg et al., 2018). Empirical studies typically use patents or R&D expenditures to measure firm innovation, often focusing on innovation as a variable of interest (e.g., Croce et al., 2018) or as a control variable (e.g., Huang, 2018). A well-known complicating factor in this cross-disciplinary body of research is that most firms do not report their R&D spending nor obtain patents (Anton and Yao, 2004; Koh, Reeb, and Zhao, 2018). Figure 1 shows the empirical magnitude of the unobserved innovation problem around the world. Globally, 65% of the observations in listed firms do not provide R&D expenditure data, while more than 85% of them do not obtain patents. Focusing on US firms, slightly over 50% of firms do not report their 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 disclosing R&D or obtaining patents include negligible innovation inputs, unsuccessful innovation projects, or attempts to keep innovation information secret and obfuscate the information environment (Png, 2017).

Recognizing that most firms fail to report R&D expenditures or seek patents, empirical researchers use a variety of methods to handle unreported innovation. The two most common approaches to dealing with this issue are excluding firms without R&D or patents (e.g., Hombert and Matray, 2018) or classifying these firms as zero innovators and including a dummy variable as suggested by Koh and Reeb (2015) (e.g., Masulis and Zhang, 2019). Of course, unreported innovation could arise because the firm does not engage in innovation and has nothing to report. Consequently, any potential method for handling unreported innovation must explicitly capture or allow for these firms to have zero innovation.

Our analysis focuses on how these common methods for dealing with missing innovation data influence the economic conclusions in empirical finance research. To address this issue, we investigate the reliability of different methods for handling unreported innovation in studies that focus on innovation as either an explanatory variable of interest or include it as a control variable. We compare six approaches to handling unreported innovation: Listwise deleting (discarding) firms without R&D or patents, deterministic imputation with either zero or industry average, inverse probability weighting, Heckman selection, and multiple imputation. Multiple imputation 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 (see Internet Appendix I). Our analysis focuses on the assumptions underlying different practices for handling unreported innovation and considers the econometric implications of these common approaches when the assumptions are violated. Preliminary analysis reveals that unreported R&D and firms without patents are predictable by known determinants of innovation and other corporate outcomes of interest, rejecting the hypothesis that unobservable innovation is *missing completely at random*.[1] These results raise the concern that the commonly used methods to handle unreported innovation could lead to biased parameter estimates due to the non-representatives of the population under study (deletion) or distortions of the variance-covariance matrix (deterministic imputation).

Our empirical tests use two different sources and mechanisms of missing innovation data: One on R&D spending and the other on patent counts and citations. We use data on firms that

[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 Section 2 and the Internet Appendix I for more formal discussions of the assumptions underlying the different methods to handle missing innovation and the consequences on inference if these assumptions are violated.

did not report R&D spending in a particular period but reported the amount in subsequent financial statements, to assess the reliability of different methods of handling unreported R&D. As firms that initiate the reporting of R&D expenditures are required to report their R&D expenditures for prior years, we can directly compare several common treatments for missing R&D in the prior years. To the empiricist these firms do not appear to engage in R&D activity in the years R&D spending was not reported, creating a natural laboratory to evaluate different assumptions and methods of handling unreported innovation. We denote this newly reported R&D spending in future financial statements as "*Recovered R&D.*" Using recovered R&D as a baseline, we compare the observed/counterfactual R&D with replacing these firms' missing R&D with zero, the industry average, and multiple imputation. Further tests show that the R&D in firms with unreported R&D significantly differs from zero R&D firms, the average industry R&D, and positive R&D firms in aggregate. Notably, we find that on average multiple imputation gives estimates of R&D for the missing R&D data that is qualitatively similar (not statistically different) to their actual R&D reported in subsequent financial statements.[2]

One potential issue with using recovered R&D to compare treatments for unreported observations is that these firms may differ from other firms that do not report innovation in the Compustat universe. To mitigate this concern, we undertake two simulation studies. In the first simulation analysis, we use the empirical distribution of US Compustat data to evaluate the impact of differing levels of missingness of an innovation variable. This simulation approach approximates the analyses typically found in empirical studies of corporate innovation using panel data. The second simulation analysis uses clearly specified data generating processes, allowing us to gauge

[2] Conceptually, one could also use the R&D reported in tax-based income statements provided to the IRS to gauge the magnitude of missing R&D in audited financial statements. Underlying this approach is the notion that failing to separately report R&D in financial statements does not affect the declarations of R&D in tax-based income statements (Lisowsky, 2010). In addition to the difficulty in obtaining non-public income tax filings from the IRS, classifying expenses as R&D for tax and financial statements purposes often rely on different accounting rules and guidelines.

the impact of unreported innovation in a controlled, cross-sectional setting. This approach mitigates concerns about the comparability of the data on US firms in our first simulation with the data used in other studies. In both simulation exercises, we evaluate the relative performance of deleting unreported R&D firms, deterministic imputation methods to replace missing R&D, inverse probability weighting, Heckman, and multiple imputation in handling 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 (RSME) of the regression coefficient estimates. We evaluate the coefficient estimates on both the innovation variable (e.g. R&D or patents) and other control variables.

In both simulations, we find that deleting or excluding firms with unreported R&D leads to biased coefficient estimates for both R&D and any control variables correlated with R&D. Rather than providing a conservative approach, the deletion of firms without reported R&D is one of the worst methods for handling unreported innovation. For instance, if R&D is *missing at random*, then the average bias from excluding firms without reported R&D is almost three times greater than found using multiple imputation. Moreover, our analyses show that commonly used deterministic imputation models (e.g., replace missing R&D with zero or industry average and include a dummy variable) fare poorly in comparison to multiple imputation. Of course, in a single industry analysis with limited numbers of positive R&D firms (e.g. Real estate renting and leasing, SIC 53) or where upward of 90% of the missing R&D firms arise from zero R&D expenditures, replacing missing with zero will provide a reasonable solution. Yet, multiple imputation still performs well in this scenario as well.

In addition, we find that the RMSEs of the common methods for handling missing innovation are very large in comparison to multiple imputation in both simulation exercises. For instance, in our first simulation, we find that the RMSE for the coefficient estimate on innovation

when deleting firms without observed innovation is 70% larger than found when using multiple imputation (under *missing at random*). We observe similar results on the coefficient estimates for other variables of interest across both simulation exercises. We also find that the bias and RMSE from deleting firms without observable innovation dramatically increase at higher rates of missingness. Prior empirical research highlights the strategic decisions underlying the patent versus trade secret choice, suggesting the high rate of missingness in patent relative to R&D spending makes deletion especially challenging in studies focusing on patents or their citations. One of the most important takeaways from these findings is that commonly used solutions to handle unreported innovation can lead to biased parameter estimates that make prior inferences about corporate innovation difficult to assess.

To illustrate the economic magnitude 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. They use market-to-book and R&D expenditures as measures of investment opportunities, and they classify firms without reported R&D expenditure as zero R&D firms. Their pecking order-based arguments predict positive correlations between leverage and both market-to-book and R&D, but they find a negative relation for R&D. As a first step, we replicate their analysis and find similar results. We then repeat the analysis using multiple imputation as suggested in our benchmarking exercise, to account for unreported R&D. We find that the coefficient estimates and standard errors for R&D and capital structure are significantly different when using multiple imputation to account for unreported innovation relative to the results when classifying these firms as zero innovators. Strikingly, under multiple imputation both R&D and market-to-book have positive coefficients, providing evidence consistent with pecking order theory predictions rather than the conflicting results reported in Fama and French (2002).

So far, our empirical analysis primarily concentrates on unreported R&D spending. Yet, patent-based metrics are another common approach to measuring corporate innovation. Arguably, missing R&D and patents differ substantially from each other as R&D expenditures are a mandatory disclosure, while patents stem from voluntary disclosure choice (Koh and Reeb, 2015). In 1976, the US Supreme Court defined materiality based on subjective managerial or investors views, rather than specific bright-line thresholds. Consequently, *SFAS 2 Accounting for Research and Development Costs* leaves the decision to separately report R&D expenses to managerial discretion. Still, as noted in Figure 1, over 75% of US firms do not seek patents, suggesting the potential for an even more severe missing data problem with patents relative to R&D. US firms without USPTO patents in studies of corporate innovation potentially arise from the decision to keep successful innovation as a trade secret (Png, 2017), from the decision to patent in alternative markets, or from failed research programs. Among US firms, 69% of positive R&D firms never file for patents using USPTO data, while only 43% never file patent applications using the 30 global patent offices. This 26% wedge in unreported patents to the researcher relying on USPTO patents provides another opportunity to examine different methods of handling missing innovation data, arising from a different source of missingness. Although unobserved patents and R&D likely differ in the mechanism of missingness, we again find that multiple imputation provides much closer coefficient estimates to the unreported non-USPTO patents than commonly used replacement methods.

Our patent and patent citation analyses reveal that unobservable patents are predictable with commonly used firm-level variables, providing additional evidence against deleting firms without observable patents. Additional tests show that these missing patents generate sizable but different levels of citations than their domestic counterparts. Thus, similar to unreported R&D, we reject the hypotheses that unreported patents arise completely at random. In a simulation with the

empirical distribution of US financial and patent data, we find that deleting or excluding firms without patents leads to biased coefficient estimates for both patents and any control variables correlated with patents. Coupled with the results in our simulation tests based on simulated data generating processes, the patent results suggest that deleting or discarding firms without observable patents can lead to biased coefficient estimates and economic conclusions.

The nature of missing innovation data is unknown to the researcher. How we handle this missing data problem ultimately comes down to our assumptions about the mechanisms of missingness. Implicitly, researchers deciding how to handle missing innovation data are making assumptions about whether missingness can or cannot be predicted by observables. In other words, a researcher must decide on the assumptions of MAR versus MNAR, which involves the use of MI and IV respectively. IV relies on the ability to find truly exogenous shocks to overcome the selection bias (see discussion in Jiang, 2017). MI assumes that missingness can be predicted with observables, which implicitly facilitates the estimation of average treatment effect under MAR. Given that the assumptions underlying IV and MI are both likely to be violated to some degree, the choice between assuming MAR or MNAR depends on the bias of the IV and MI estimates. Collins, Schafer, and Kam (2001) demonstrates that in many realistic cases, an erroneous assumption about MAR often has limited impact on estimates and standard errors because covariates included in the imputation models are often correlated with unobservable determinants of missingness. Consequently, we recommend using both MI and IV estimates when confronted with missing innovation data. It is also important to discuss the plausibility of the assumptions about the nature of missing innovation data in a particular sample.

This study provides several insights and contributions to the innovation literature. First, studies on innovation should consider using several different approaches to handling unreported innovation (R&D and patents). In this context, 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 deleting firms with missing patents and R&D, we should attempt to adjust for the non-randomness in missing innovation data. We advise against the common approach of performing the main analysis by replacing missing R&D with zero (or industry average) and then repeating the tests after excluding these non-reporting firms as sensitivity analysis. Both deterministic imputation and listwise deletion of firms with unreported corporate innovation can provide biased coefficient estimates if the missingness is non-random, making it difficult to evaluate how well one biased approach can provide a robustness test for another biased approach.

Second, this study contributes to the burgeoning work on the econometric challenges faced by researchers in finance. Bertrand, Duflo and Mullainathan (2004), Petersen (2009), and Thompson (2011) discuss methods to appropriately compute standard errors in the presence of cross-sectional and time-series dependence across residuals. Koh and Reeb (2015) compare various deterministic imputation methods, excluding listwise deletion, and find that including a dummy variable for missing R&D in the regression analysis improves their results, but they are silent on the relative biasness of these methods. Our analysis shows that this class of solutions can lead to biased estimates and standard errors for both unreported R&D and patents. Importantly, our paper questions the foundations for deleting firms with unreported innovation (widely adopted in economics and finance) and the impact of using deterministic imputation models that classify these firms as zero innovation firms and including a dummy variable (frequently used by accounting and finance scholars). In addition, we show that the use of patents (counts or citations) as alternative innovation measures does not resolve the missing data problem. Instead, the problem of unobservable innovation is arguably more pronounced in studies that use patents to measure innovation than in ones using R&D expenditures, because of the higher rate of missingness in

patents. At a minimum, our results indicate that studies of US firms using USPTO patents should use the larger PATSTAT sample. These missing patent firms comprise over one-quarter of the non-patenting firms in the US using USPTO data. More importantly, deleting firms without patents in the PATSTAT sample, or classifying these firms as zero innovators can lead to biased coefficient estimates.

Third, studies that use R&D or patents as a control variable also suffer from this missing data bias. Best practices for dealing with missing R&D and patents depend on the source or type of missing innovation data. If country, industry, or firm characteristics predict unreported R&D or missing patents (see Lerner and Seru, 2017), then our analysis suggests that using multiple imputation provides the most reasonable solution. Surprisingly, and across a wide variety of specifications and approaches, we find that both Heckman and inverse probability weighting rarely provide the best approaches to handling unreported innovation in our samples. For alternative data sets, researchers could consider undertaking simulations similar to ours to evaluate the various methods of handling unreported innovation. We provide our code for researchers interested in performing simulations on unreported innovation using their own unique data.

## 2. Handling Missing Innovation Data

There are numerous possible reasons for why we observe missing innovation data. Unfortunately, the missingness mechanism cannot be positively identified from examining the observable data. Hence, as empiricists, we make either implicit or explicit assumptions about the missingness mechanism for firms without patents or R&D spending to draw inferences, which are separate from the statistical methods we use for parameter estimation. In general, missing data causes two problems: Bias in the parameter estimates and loss in efficiency (Rubin, 1976). Bias

stems from the non-representativeness of the population under study. Loss of efficiency arises because information loss is a direct consequence of missing data, i.e. smaller samples.

To provide a framework for investigating unreported innovation, we consider the case where only one explanatory variable contains missing observations. Let $y_i$ be the dependent variable and $z_i$ be the innovation variable with missingness. We have the linear relation:

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

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 solutions to this problem depends on the missingness mechanism, thus we first present the three missing mechanisms. Rubin (1976) and Little and Rubin (2002) classify missing data mechanisms into *Missing Completely at Random*, *Missing at Random,* and *Missing Not at Random.*

1. **Missing completely at random (MCAR)**: The probability of missing can be formulated by:

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

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

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

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

   The probability of missingness only depends on the set of *observed* variables $x$ and $y$, but not on the missing variable itself or unobservable characteristics.

3. **Missing not at random (MNAR)**: The missing mechanism depends on the value of $z$ itself or on unobserved variables, e.g., high-income individuals tend to not participate in surveys related to income.

There are several potential mechanisms for missing innovation data, which likely differ between different measures of innovation. Missing R&D data could arise from firms seeking to avoid giving benchmark spending numbers to competitors, managerial decisions to create information asymmetry about their R&D intensity, a simple failure to report zero R&D, or difficulties in estimate R&D spending from the costs of goods sold. Unobservable innovation in patent data could arise from failed innovation projects, managerial decisions to facilitate private trading gains, firm attempts to keep detailed blueprints of their innovation out of the public domain, or because they focus on process rather than product innovation. Understanding the mechanism or the underlying reason for the missing innovation data is an especially important component in determining or assessing methods to handle the missing data. For instance, the Heckman selection approach requires the selection of an instrument based on the nature of the missingness.

*2.1. Common Approaches to Unreported Innovation*

One common approach to missing innovation data is to delete or exclude firms without R&D spending or patents. Listwise deletion only uses 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:

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

where $s_i z_i$ is now the explanatory variable and $s_i\varepsilon_i$ is the error term. The OLS (ordinary least squares) estimator is unbiased if $E(s_i\varepsilon_i z_i) = 0$, which is by $E(\varepsilon_i|z_i, s_i) = 0$. If data is *Missing Completely at Random* and $z_i$ is exogenous, then $E(\varepsilon_i|z_i, s_i) = E(\varepsilon_i|z_i) = 0$. Thus, deletion can lead to consistent estimates in the case of *Missing Completely at Random*. However, if the selection is driven

by observed or even unobserved variables, then $E(\varepsilon_i|z_i, s_i) \neq 0$ in general because $\varepsilon_i$ can be correlated with $s_i$ even if one controls for $z_i$, leading to biased estimates produced by deletion. Thus, a preliminary test to consider the potential costs of deleting firms without observable innovation is to assess whether the missing data is correlated with key variables of interest.

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. Frequently used deterministic imputation methods impute the missing values with zeros (i.e. firms without R&D are considered as having zero innovation), with the industry average level of innovation, or with fitted values based on some pre-specified model. The validity of this method depends on whether the specified imputation models are correct. If the imputation model perfectly coincides with the missing mechanism, then the resulting coefficient estimate using the imputed sample is consistent. The misspecification of a deterministic imputation model can lead to biased estimates because of the distortion of the variance-covariance matrices. This provides testable implications for analyzing missing R&D and patents, namely whether firms with unreported innovation have positive values of R&D or patentable innovations.

### *2.2. Alternative Methods for Handling Missing Innovation Data*

Two other approaches to handling missing data are also viable candidates for unreported innovation. Inverse probability weighting (IPW) relies on assigning different weights to observed points depending on their probability of being observed. As this probability is unknown for unreported innovation, we can estimate it using binary choice models, such as logit or probit, or with a nonparametric model. The second approach is multiple imputation (MI). MI is essentially an iterative version of stochastic imputation, which aims at explicitly modeling the uncertainty/variability ignored by the deterministic imputation procedures. Instead of replacing

with a single value (unrelated to other covariates/observed data), multiple imputation uses the (joint) distribution of the observed data to estimate the parameters of interest multiple times to capture the uncertainty/variability in the imputation procedure (see Internet Appendix I). MI methods and Heckman-type approaches to deal with unreported innovation arise from different assumptions about the nature of the missing data. In empirical studies, researchers face a tradeoff between the assumptions that underpin MI versus the assumptions about the exogeneity of the instruments used in Heckman models. In our analysis, we investigate the relative performance of deleting firms without observable R&D or patents, common deterministic imputation methods to replace unobservable R&D or patents with zero or industry mean, Heckman, inverse probability weighting, and multiple imputation as different approaches to handle unreported innovation.

## 3. The Severity of Unreported Innovation

*3.1 Data and Sample*

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

Our sample selection begins with the October 2013 version of the PATSTAT data. It contains 44,730,405 observations, including patentees who are individuals, governmental institution/universities, and companies for the sample period of 1999–2012.[3] Our analysis relies

[3] Our patent sample end in 2012, because patents post 2012 may be affected by the truncation bias. The truncation bias arises due to patents after 2012 not having enough time to receive citations and result in fewer citations in comparison to earlier patents (Hall et al., 2001).

on the registered names on the original patent applications, rather than the ultimate patent owners, to better capture the entities that performed the innovation activities. We merge the patent data with all publicly-listed firms in the Compustat North America and Compustat Global database for 32 countries. Our matching algorithm consists of two main steps. First, we standardize patent assignee names and firm names, focusing on unifying suffixes and dampening the non-informative parts of firm names. Second, we apply multiple fuzzy string-matching techniques to identify the firm, if any, to which each patent belongs. We randomly selected 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).[4] 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 UK, and the US. There are 30 patent offices in the sample because the EPO is a separate entity from each European country's patent office; European firms sometimes patent in their home patent office and other times with the EPO. Our baseline sample includes 333,920 firm-year observations and 37,272 unique firms, of which 5,374 are cross-listed firms. All accounting variables are from Compustat (North America and Global) and are defined in Panel A of Table A1 in the Appendix.

Panel A in Table 1 reports the basic descriptive statistics of our sample firms. Only 35% of the observations in our sample report any information on R&D. Of those reporting R&D expenditures (118,264), 93% report positive R&D with an average R&D expenditure of 8% of their total assets. 7% of firms report zero R&D. The 75$^{th}$ percentile of R&D expenditures captures

[4] Relaxing this 100-firm constraint or using a 1,000-firm constraint leads to similar inferences (see Table IA1 in Internet Appendix II).

firms where R&D equates to roughly 6% of total assets. In addition, the sample firms invested an average of 6% of total assets in capital expenditure. Firms have an average of 9 patent applications, 4 patents granted, and 23 citations over the sample period.[5] On average, firms are profitable with an average ROA (return on assets) of 1% (median of 5%) and are highly levered with median leverage of 52%. In our analysis, we focus on patent applications, as these capture the R&D activity happening around the firm, but find similar results using patents granted.

*3.2. 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 both reported R&D expenditures *and* patent information. Our counterfactual 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 the full sample (i.e., no deletion based on either reported R&D or patent application), the partially deleted sample with either positive R&D or patent applications, and the sample, with both reported R&D and patents.

Panel B in Table 1 reports the univariate characteristics of the full sample (Column 1), the sample that reports R&D (Column 2), the sample that reports 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

[5] The average time between filing a patent application and a patent being granted across different patent offices ranges between 2 and 4 years.

to the full sample. The samples with reported R&D or patents have less than a third of the observations of the full sample. These subsamples have higher total assets than the full sample, while the rest of the variables are significantly lower (Columns 5 and 6). The R&D and patent-only sample consists of 53,456 observations. Total assets, Tobin's Q, and sales growth are larger than those in the full sample, while the rest of the variables are smaller (Column 7). It is worth pointing out that ROA decreases by 400% from the full sample to the R&D and patenting sample. These results indicate that R&D and patenting are at least not *missing completely at random* and may depend on observables.

*3.3. Tests of the Deletion Assumptions*

Next, we evaluate the validity of the assumptions underlying the common practices of deleting missing R&D and replacement with zero. An example of an MCAR process (when deletion of observations with R&D is valid) is one in which firms decide whether to report R&D 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 R&D is more prevalent across firms with certain firm characteristics, by examining the predictability of unreported R&D through regression analysis.

Whether missing data is MCAR can be tested by investigating if there are significant differences between the means of different missing-value patterns across variables of interest. This is formalized by Little (1988), who implements the Chi-square test of MCAR for multivariate quantitative data. The test statistic takes a form similar to the likelihood-ratio statistic for multivariate normal data and is asymptotically $\chi^2$ 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.

Panel A of 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 $\chi^2$ statistic ranging between 297 and 22,889 for both the global and U.S. sample, providing strong evidence that unreported R&D is not MCAR. This concurs with results in Panel B of Table 1.

Next, we investigate whether the observed variation in unreported R&D at the firm-year level is systematically related to firm characteristics.[6] We assess the existence of identifiable patterns in unreported R&D by conducting a regression analysis of unreported R&D on observable firm characteristics at the firm-year level for international and U.S. firms. Note that these tests do not seek to establish causality, but rather to emphasize association and predictability in the variation in unreported R&D to shed light on the nature of missingness in R&D.

We estimate a panel regression model with year, industry, and country fixed effects, where the dependent variable is unreported R&D. Unreported R&D is equal to 1 when R&D is not reported and zero otherwise.[7] For all firms, firm characteristics with country, industry, and year fixed effects explain up to 38% of the variation in unreported R&D (Panel B Column 3 of Table 2). Firm characteristics with firm and year fixed effects explain 81% of the variation in unreported R&D (Column 4). Furthermore, unreported R&D increases at the firm level with property, plant and equipment (PPE) investment, ROA, and sales growth, while it decreases with total assets. For

[6] Cross-country regressions show that percentage of firms with unreported R&D is predictable with macroeconomic variables, including economic openness, manufacturing intensity, government subsidies, labor regulations, intellectual property rights, university ties, skilled labor, honesty, regulatory efficacy, and Commonwealth countries.

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

U.S. firms (Columns 5-7), industry and year fixed effects explain 53% of the variation in unreported R&D, while firm and year fixed effects explain 93% of this variation. Unreported R&D increases with PPE, ROA, and leverage, while it decreases with capital expenditure (CapEx) and sales growth for U.S. firms (using either contemporaneous or lagged explanatory variables). Collectively, our evidence indicates a significant correlation between missing R&D and firm-specific factors and the results are inconsistent with R&D being *missing completely at random*.

## 4. Unique Setting to Investigate Imputing Unreported R&D

### *4.1. The Setting*

The main challenge to imputation approaches relates to how close are the imputed estimates 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 comparisons by users of general-purpose financial statements (Statement of Financial Accounting Concepts No. 8). This enables us to identify a unique (albeit narrower) 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 R&D expenditure amounts. In this instance, we can identify the previously unreported R&D expenditures. Our unique setting is thus especially appropriate to investigate how close the imputed estimates from various imputation methods are to the "recovered" true values.

Using the sample of US firms for the period 1992 to 2016, we identify firms that switch between reporting and not reporting R&D expenditure.[8] We find 738 unique firms that switch between reporting and not reporting R&D. We then manually collect data from the annual reports (10Ks) of these firms on their prior years' R&D expenditure, collecting information on the reported R&D in the year of the switch and up to two years prior to the switch in reporting. We restrict our analysis to firms without any major corporate events (e.g., merger and acquisitions) over the past two years that would have altered the underlying business operations of the firm (e.g. Bena and Li, 2014).[9] We denote these as "Recovered R&D" firms. This provides us with 763 observations for the switch year (i.e. some firms switch between reporting and not reporting R&D more than once during our sample period) and 1,032 recovered observations (i.e., some firms report amounts for one year, while others for two years before the switch). Figure IA1 in the Internet Appendix shows the distribution of "Recovered R&D" and its cumulative distribution function.

*4.2 Comparing Recovered R&D Firms to Zero*

We begin our analysis by comparing the characteristics of Recovered R&D firms with zero R&D firms and positive R&D firms. Table 3 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 for the comparative years (t-1 and t-2). By default, the R&D expenditure and R&D value of recovered firms are statistically different from the zero R&D firms. In addition, the recovered firms differ from zero R&D firms across several different dimensions like total assets, PPE, and leverage.

[8] Our initial analysis uses a window which covers 1999-2012 due to data limitations for pre-1999 international data. Our PATSTAT sample ends in 2013 and determines the end of the main sample. In tests focusing strictly on US firms, we use a longer sample period (1992-2016).

[9] This is to ensure that the prior year figures disclosed in the switch year reflect only business operations that existed in the prior year 10K filings where R&D spending was not reported.

Panel B of Table 3 compares Recovered R&D firms with their positive R&D counterparts. The R&D absolute investment for recovered firms is significantly lower than positive R&D firms, but the R&D expenditure of the two groups are not distinguishable from each other. Recovered R&D firms also differ from positive R&D firms in total assets and PPE. Untabulated multivariate tests provide similar inferences, showing that recovered R&D is predictable by many common firm characteristics. Overall, results in Panels A and B of Table 3 show that unreported R&D expenditure firms differ from both zero R&D firms and positive R&D firms, suggesting that deleting them or classifying them as zero innovators is problematic. 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.

*4.3. Comparing Different Imputation Methods*

Potential methods of handling missing data are listwise deletion, imputation with zero or industry mean, and multiple imputation.[10] We test the different imputation techniques using the "Recovered R&D" sample as a counterfactual for the true R&D in Panel C of Table 3. In the Compustat data, this recovered R&D appears as missing, and we impute this R&D with zero, with average industry R&D (two-digits), 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 present statistics both for R&D and R&D/Ln(Total Assets).

For multiple imputation, we impute R&D and R&D/Ln(Total Assets) using two multiple imputation settings for the whole US sample for the period 1992 to 2016 (not just the recovered R&D sample). In the first imputation method M1, we impute R&D using the natural log of total

[10] See Internet Appendix I for a detailed exposition of all the methods.

assets, ROA, PPE, sales growth and leverage by industry (two-digit). In the second imputation method, M2, we impute R&D using the same model as M1 with the addition of the lagged R&D expenditure at the firm level, as conditioning information. We use 200 iterations for imputation in the analysis.

Panel C of Table 3 shows that recovered R&D is statistically different from zero, i.e. replacing with zero underestimates the recovered R&D values. In terms of the dollar amount of R&D, the average recovered R&D is $6.69 million, while imputing with the industry average, gives an estimate of $77.35 million. The industry average imputed value is over 10 times their actual R&D spending and significantly different from the recovered value. On the other hand, the two multiple imputation methods generate an average of $15.52 million and $12.17 million which are not statistically different from the recovered R&D values. The relatively large variance in the MI values points to the difficulty in using these as exact point estimates for innovation in firms with missing R&D, even though it could provide less biased coefficient estimates in OLS models. One reason for this high variance is that MI procedures allow for non-positive estimates of missing innovation even though R&D can take only strictly positive values. Figure 2 shows the distribution of imputed R&D in our primary sample, which demonstrates the concerns of using statistics based on conditional quantiles with this approach or with using these values as stand-alone point estimates for a particular firm. In instances where this is of particular concern, standard statistical packages typically offer options for MI with chained predictive mean matching procedure and imputation of categorical variables.

Overall, results in Table 3 show that firms with recovered R&D differ from firms that explicitly report zero R&D, they are not similar to the average firm in the industry, and multiple imputation provides the closest imputation to the true value of their R&D investment. Although this design provides a sharp setting to generate the above-mentioned insights, it may not be fully

representative of the broader set of firms with missing innovation data. As such, in the next section, we turn to two simulation analyses to alleviate this concern.

## 5. Simulation Analysis

We consider two simulation studies, one based on the empirical distribution of Compustat (US) data and one on simulated data, to compare different methods of dealing with missing data in various data generating processes (DGPs). The first approach mimics current empirical exercises involving R&D. The second approach allows us to determine the distribution of all variables and their correlations and to examine the performance of methods in a well-controlled environment.

In both cases, we compare six methods to handle missing values. First, we consider listwise deletion. Second, we impute the missing R&D expenditure by zeros (ImpZero). Third, we impute the missing R&D by the industry average (ImpMean). Specifically, if an observation of firm $i$ at time $t$ is missing, we impute the missing observation by the industry average (two-digit SIC code) for the firm in the same year. For both ImpZero and ImpMean, we also include a dummy variable indicating missingness as an explanatory variable. Fourth, we use Heckman's two-stage procedure with the selection variables containing all the observed covariates $W$. Heckman's procedure first predicts firms' selection probabilities by $W$, then corrects the selection bias by including a transformation of these predicted probabilities as an additional explanatory variable. Next, we consider the inverse probability weighting method (IPW) that weights each $i$-th observed point by the inverse of its conditional selection probability in least square estimation. We use the standard package of Heckman's procedure and IPW in STATA. Finally, we consider multiple imputation (MI). Since the variables generating the missingness are not known a priori, we use all observables

including the outcome variable as selection variables in the imputation model.[11] To implement MI, we use 200 imputations based on a Markov chain Monte Carlo (MCMC) procedure and employ a multivariate normal regression for each imputation.

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

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

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

where $\hat{\theta}^r$ is the estimate in the *r*-th replication, $var(\hat{\theta}^r)$ is the estimated robust variance of $\hat{\theta}^r$, $\theta^0$ is the true value of the parameter, R is the number of simulations. Note that we present the bias as a proportion of the benchmark $\theta^0$ to compare across coefficients, thus one cannot use the reported bias (B) to calculate the RMSE. Based on the observed missingness of R&D and patents in Figure 1, we consider two levels of missingness relevant for innovation variables: 50% and 70%. We perform 500 simulations.

*5.1. Empirical Distribution-Based Simulation*

For the empirical distribution-based simulation, we begin with a panel sample of 783 firms in Compustat over the period 1992-2012, where we have non-missing information on all financial

[11] Multiple imputation draws the missing variable from a joint (predictive) distribution of observables for multiple times, and thus the set of observables should include all variables that are potentially correlated with the missing variable. Since sales growth is correlated with R&D, we also include it in the imputation model. Ignoring sales growth in imputing missing R&D leads to incomplete conditioning of observables and biased estimates in the regression of interest (Moons et al., 2006; Sterne et al., 2009; Bartlett et al., 2011).

variables of interest, except for R&D. The data includes: natural log of total assets (A), leverage (L), intangible assets (I), Tobin's Q (Q), return on assets (R), R&D expenditure (RD), and sales growth (S). To investigate the effects of R&D missingness on the coefficient estimates of our evaluation model and how different methods of handling missing R&D perform, we generate R&D expenditure with missing observations that incorporate the three types of missingness. The resulting estimated coefficients ($\hat{\theta}^r$) under each condition are used to calculate the bias and RMSE per Equations 3a and 3b. Our baseline regression uses simulated sales growth as the dependent variable and R&D expenditures, the natural log of total assets, Tobin's Q, leverage, and return on assets as explanatory variables. This approach enables us to obtain a clean set of benchmark coefficients that are free from researcher intervention except for the balanced, non-missing data criteria. Next, we describe the data generating process (DGP) for i) R&D expenditure with missing observations, and ii) the outcome variable of interest, sales growth.

*5.1.1. Generate Missing R&D Expenditure*

To simulate the missing R&D, we employ a subsample of complete balanced panel data, without missingness in R&D, that contains 311 firms over 21 years from 1992 to 2012. A clear advantage of this approach is that we do not need to make assumptions or estimate the conditional distribution of the R&D given that it is not missing, which is typically difficult to obtain. More importantly, it allows us to introduce the three types of missingness more precisely into the data as described below while providing us with "true" values as benchmark cases. We generate a missing indicator for R&D, denoted by *M* that equals 1 if R&D is missing and 0 otherwise. Once we model and assign missing R&D observations, we can obtain the simulated R&D since the data are complete and the non-missing observations are given by their original values.

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

- Missing completely at random: $M_{it} = \tilde{\eta}_{it}$, (4)
- Missing at random: $M_{it} = \alpha_i + \beta_O' X_{it}^O + \tilde{\eta}_{it}$, (5)
- Missing not at random: $M_{it} = \alpha_i + \beta_O' X_{it}^O + \beta_U' X_{it}^U + \tilde{\eta}_{it}$, (6)

where $X_{it}^O$ contains observed variables by researchers, while $X_{it}^U$ is unobserved and only appears in the DGP but is omitted in imputation models. In the case of MAR, we consider $X_{it}^O = (Q_{it}, A_{it})'$, 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 $\alpha_i$, $\beta_O$, and $\beta_U$. However, it is well recognized that modeling binary variables is difficult in econometrics, and this is even more complicated in panel data models due to the difficulty of estimating individual fixed effects; see Lahiri and Yang (2013) for a review. To incorporate the firm fixed effects, we adopt the commonly used assumption that the firm fixed effects are correlated with the time-average of covariates in a linear manner (see Chamberlain, 1984), i.e. $\alpha_i = c + \gamma_O' \bar{X}_i^O + u_i$ in MAR and $\alpha_i = c + \gamma_O' \bar{X}_i^O + \gamma_U' \bar{X}_i^U + u_i$ in MNAR, where $\bar{X}_i^O = 1/T \sum_{t=1}^T X_{it}^O$, $\bar{X}_i^U = 1/T \sum_{t=1}^T X_{it}^U$ and $u_i$ is the idiosyncratic noise. This assumption implies that we can incorporate the firm fixed-effects by augmenting regressions (5) and (6) by the time-series averages of covariates, respectively, as:

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

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

where $\eta_{it} = \tilde{\eta}_{it} + u_i$. Since there are no fixed effects in (7) and (8), we can estimate all parameters in these two models and predict $M_{it}$ based on these estimates. Specifically, we first estimate (7) and (8), respectively, by a probit regression of the missing data indicator for R&D using the panel data

sample (783 firms). We set the estimates $\hat{c}$, $\hat{\beta}_O$, $\hat{\beta}_U$, $\hat{\gamma}_O$, and $\hat{\gamma}_U$, as the true parameters to generate the missing probability $M_{it}^*$ in the *complete subsample* of the data:

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

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

1. Missing completely at random: $pm_{it} = \eta_{it}$, (10)
2. Missing at random: $pm_{it} = \hat{c} + \hat{\beta}_O X_{it}^O + \hat{\gamma}_O \bar{X}_i^O + \eta_{it}$, (11)
3. Missing not at random: $pm_{it} = \hat{c} + \hat{\beta}_O X_{it}^o + \hat{\beta}_U X_{it}^u + \hat{\gamma}_O \bar{X}_i^O + \hat{\gamma}_U \bar{X}_i^u + \eta_{it}$, (12)

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

*5.1.2 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.[12] 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:

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

[12] 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), rendering the comparison between various imputation methods less clear.

where $\mu_i$ is firm fixed effects, $Z_{it}$ contains the determinants of sales growth, $Z_{it} = \{A_{it}, Q_{it}, R_{it}, L_{it}\}'$, and $\varepsilon_{it}$is the error term. Note that intangible assets are not observed and thus also not included in the DGP of $S$. The firm fixed effects $\mu_i$ are generated by $\mu_i = 0.1\iota'\bar{Z}_i$, where $\iota$ is a 4×1 vector of ones and $\bar{Z}_i = 1/T \sum_{t=1}^{T} Z_{it}$, and thus $\mu_i$ is correlated with sales growth determinants. To obtain the parameters for $\delta'$ and $\theta'$, we estimate (13) using the same complete subsample without missingness and fix the estimated values in the simulation. To allow the idiosyncratic error to be correlated with selection instruments, we generate $\varepsilon_{it} = \tilde{\varepsilon}_{it} + \bar{Q}_i$ in MAR and $\varepsilon_{it} = \tilde{\varepsilon}_{it} + 0.5(\bar{Q}_i + \bar{I}_i)$ in MNAR. Here $\bar{Q}_i$ and $\bar{I}_i$ are the time average of Tobin's Q and intangible assets for firm $i$, respectively, which drive the missingness of R&D as discussed in Section 5.1.1. $\tilde{\varepsilon}_{it} \sim IID\ N(0, \sigma_\varepsilon^2)$ and $\sigma_\varepsilon^2 = 0.18$ based on the empirical distribution of the residual from estimating equation (13).

*5.1.3 Simulation Results*

Table 4 reports the simulation results under three missing mechanisms and two levels of missingness (50% and 70%). When R&D is missing completely at random, we find that both bias and RMSE increase with increasing missingness in R&D (Panel A). All methods show a relatively small bias under MCAR, except IPW and Heckman. IPW and Heckman, typically do not include fixed effects due to the difficulty in estimating fixed effects in binary model settings, which potentially explains part of their relatively poor performance (we use the standard packages in STATA for these two methods). In *missing completely at random*, deletion and multiple imputation have the lowest bias. Multiple imputation exhibits relatively smaller RMSE than deletion. Deterministic imputation methods (ImpZero and ImpMean) generate RMSE that are similar to MI. Still, MI has both the lowest average bias and RMSE under MCAR.

Panel B shows the results for MAR, where the bias of all methods increases from MCAR. Under MAR, all methods lead to biased estimates, not only for R&D (which has missing observations), but also for the other explanatory variables that do not have any missingness. MI on average produces the lowest bias across all of six methods follows by ImpZero and ImpMean. The average absolute bias in listwise deletion is over nine times greater than the bias in multiple imputation, while bias in IPW and Heckman are over 319 times and 180 times greater than MI. The common imputation methods, on average, exhibit similar RMSEs, where ImpZero, ImpMean, and MI have the lowest RMSEs. Panel C shows the results when missingness is driven by unobservables (MNAR). Under MNAR, MI continues to produce the lowest bias among all six methods follows by ImpZero and ImpMean. The bias in LD is six times larger than the bias in MI. Focusing on RMSE, once again ImpZero, ImpMean, and MI all exhibit similar magnitudes.

Our simulations focus on two separate levels of R&D missingness, namely 50% and 70%. However, our cross-country sample, which underlies Tables 1 and 2 reveals that the level of missingness varies by country. Specifically, the rate of missing R&D data ranges from 5% missing in Japan to 85% missing in Italy. Consequently, we repeat the simulation analysis across a wide selection of missingness levels in 5% increments. Figure 3 shows the relative bias in the R&D coefficient estimate in using multiple imputation and listwise deletion as the rate of missing R&D increases from 5% to 85%. Across the entire range of missing R&D, multiple imputation exhibits substantially lower bias in the R&D coefficient estimate relative to listwise deletion.

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 both the missingness determinants (A, Q) as control variables, which implies that even if one knows the missingness mechanism and correctly controls for it, the estimated coefficients are still biased. Second, we have assumed that the errors in the sales growth and the selection regressions are not correlated, which

is most likely not the case in reality. In untabulated results, we show that if the errors of the two regressions are correlated, then the bias of deletion and deterministic imputation increases.

The analysis of simulations based on the empirical distribution, albeit realistic and informative, does not allow us to clearly infer how the correlation between variables, which might differ across data samples, influences the performance of methods. Therefore, in the next subsection, we conduct simulation analysis using generated data, where we can precisely specify the correlation among errors and compare the magnitude of the effects of various methods in a well-controlled environment.

*5.2 Simulation with Generated Data*

*5.2.1 Data Generating Process*

We generate the dependent variable of interest as follows:

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

where $\theta_1 = \theta_2 = 1$, $\varepsilon_i \sim IID\ N(0,1)$, and N=1,000. The two covariates $z_{1i}$ and $z_{2i}$ are generated by a multivariate normal distribution with unit means and variance-covariance matrix specified later. $z_{1i}$ contains missing observations, while $z_{2i}$ is completely observed. Let $M_i$ be the missing indicator of $x_{1i}$ that equals 1 if $x_{1i}$ is missing and 0 otherwise, which is determined by $M_i = 1[M_i^* > \mathrm{Q}_\tau(M^*)]$, where $1[\cdot]$ is an indicator function, $M_i^*$ is a latent variable, $\mathrm{Q}_\tau(M^*)$ is the $\tau$-th quantile of $M^*$. We consider two values of $\mathrm{Q}_\tau(M^*)$, 0.7 and 0.5, which correspond to 70% and 50% of missing observations in $x_{1i}$, respectively. We consider three missing mechanisms for $x_{1i}$:

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

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

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

Note that all covariates $\{z_1, z_2, x_1, x_2, x_3, x_4, \varepsilon, \eta\}$ are correlated with each other. The two error terms are correlated with each other, but they are independent of the observed covariates. In MNAR, the missingness is also driven by two unobservables, which may be correlated with the errors. Hence, the unobserved selection variable $x_3$ is uncorrelated with both errors, and $x_4$ correlated with $\varepsilon$.[13]

*5.2.2 Results*

Table 5 presents the results for the simulation with generated data. As in the previous simulation, both bias and RMSE increase with missingness in $z_1$ and from MCAR to MNAR. We first focus on the results with 70% missingness. Panel A shows the results for MCAR, where all methods produce negligible bias and small RMSE for both explanatory variables. There are only

[13] We tried different parameters and generated different densities, reaching quantitatively similar conclusions. We also considered alternative specifications of the covariance matrix to investigate how the correlation between variables affects the performance of different methods. The results are available upon request from the authors.

marginal differences across LD, Heckman, IPW and MI. However, the two deterministic imputation methods (using zero and industry average) produce the largest biases and RMSE.

Panel B shows the results for MAR, where listwise deletion exhibits a substantial sample selection bias, and both coefficient estimates are downward biased at 15% and 12% respectively for $\theta_1$ and $\theta_2$. Imputation using zeros or industry means increase the bias in $\theta_1$ from -19% to -23% but decreases the bias in $\theta_2$ from 28% to 12%. The Heckman procedure exhibits among the smallest bias that is comparable to MI, but at the cost of variance. This is reflected in the large RMSE of the Heckman estimates, suggesting that the two-step procedure is rather inefficient. On the contrary, MI performs fairly well, despite the increase in biasness, it continues to have among the lowest bias. The bias of MI is around half as large as that of listwise deletion, and almost two times smaller than that of imputation using zeros or means, and MI has the smallest RMSE.

For MNAR, all methods produce biased estimates due to the non-random missing pattern, but the degree of bias differs substantially across methods (Panel C). Imputation using zeros or mean leads to the largest bias and RMSE for $\theta_1$ among the six methods ($\theta_1$ bias is around 28%); while the bias and RMSE for $\theta_2$ are generally in the middle of the six methods ($\theta_2$ bias is 14%). The biasness in LD and IPW methods also deteriorates in comparison to the MAR setting, leading to more than 17% and 15% downward bias in $\theta_1$ and $\theta_2$ respectively. Both the bias and RMSE for the Heckman procedure deteriorated by 62.5% in comparison to MAR (the largest deterioration among the six methods). Despite the observed deterioration compared to MAR, MI continues to produce the lowest bias and RMSE among the six methods.

In general, all six methods exhibit lower bias and RMSE at the lower level of missingness (50%) than at the higher level of missingness (70%). As the level of missingness varies across different data sets, the relative efficacy of the methods we investigated could differ. Replicating our

analysis across different levels of missingness reveals that multiple imputation consistently exhibits the smallest bias. In contrast, the bias from listwise deletion increases substantially with the level of missingness.

*5.2.3 Extensions*

We extend the benchmark simulation design by considering alternative specifications of the covariance matrix for the generated variables. We report results for the case of MAR with 70% missing observations in Table IA2 in Internet Appendix II, to conserve space. The ranking of the methods remains similar when considering MNAR and different levels of missingness.

First, we consider how the correlation between errors influences the performance of the methods. We increase the correlation between $\eta$ and $\varepsilon$ to 0.6. When the correlation is higher, the bias and RMSE in $\theta_1$ under all six methods deteriorates (relative to benchmark case reported in Table 5 Panel B). Heckman and MI continue to have among the lowest biasness and RMSE in $\theta_1$, while deterministic imputation and MI exhibit the lowest biasness and RMSE in $\theta_2$. Next, we consider how the correlation between the selection variables and variables of interest influences the estimates. We increase the correlation between $x_1$ and $z_1$ to 0.6, and we find that an increase in correlation increases the biasness in $\theta_1$ under LD and IPW but not Heckman; while that of ImpZero, ImpMean and MI improved. On the other hand, the biasness in $\theta_2$ improved for all six methods.

Finally, we allow for correlation between the observed selection variables and the error in the main regression. We set the correlation between $x_1$ and $\varepsilon$ to 0.4, but we generate $\eta$ independently from $\varepsilon$ to avoid direct endogeneity in the selection equation. In this case, even though the two errors are uncorrelated, the correlation between selection variables and the error

term in the main regression also significantly biases $\theta_1$ estimates under LD, ImpZero, ImpMean, and IPW. Heckman also performs poorly in this setting, because it is derived based on the joint distribution of $\varepsilon$ and $\eta$, but does not consider the correlation between $\varepsilon$ and other observables. The $\theta_2$biasness deteriorates significantly under LD, IPW and Heckman methods, while ImpZero and ImpMean exhibit smaller bias than under the baseline simulation. Overall, MI's performance improves and produces the lowest bias in $\theta_1$ and $\theta_2$ estimates among all methods. Its performance is even better than in the benchmark case because the correlation between errors is zero, leading to more accurate stochastic imputation based on the joint distribution with the selection variables.

## 6. 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. Yet, we do not know the economic significance of the effect of the treatment of unreported R&D for empirical studies. This is a cross disciplinary problem and we use the analysis of Fama and French (2002, FF02 hereafter) to assess the economic significance of the effect of different methods of handling unreported R&D on economic inference. One of the most important issues in corporate finance is understanding how firms chose their capital structure. The two prevailing models are the trade-off and pecking order models. Fama and French (2002) test the implications of these models for firm dividends and leverage. They report a positive relation between leverage and profitability for dividend and non-dividend paying firms. FF02 find ambiguous results on the relation between investments and leverage, as the two proxies for investment have opposite signs: market-to-book is positively correlated with leverage and R&D expenditures are negatively correlated with leverage. For expositional simplicity, we focus this

analysis on comparing multiple imputation to the two most commonly used solutions to missing R&D, namely deleting firms without R&D and classifying them as zero innovators.

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

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

We follow FF02 in the choice of the sample period, variables of interest, and notation. $\frac{ET_t}{A_t}$ the ratio of annual pre-interest pre-tax earnings to end-of-year total assets, is a proxy for the expected profitability of assets in place.[14] $\frac{V_t}{A_t}$, the ratio of a firm's total market value to its book value, is a proxy for expected investment opportunities. $\frac{RD_t}{A_t}$, the ratio of R&D expenditures to assets, is an additional proxy for expected investment. Unreported R&D is imputed with zero. $RDD_t$ is a dummy variable equal to 1 for unreported R&D, and zero otherwise. $\frac{Dp_t}{A_t}$, the ratio of depreciation expense to assets serves as a proxy for non-debt tax shields. $\ln(A_t)$, the natural logarithm of total assets is a proxy for volatility. The sample period is 1965-1999.

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

Table 6 replicates FF02 using a contemporaneous regression with two-way fixed effects, double clustered standard errors, and two additional treatments for unreported R&D.[15] The various estimation techniques lead to very different estimates for β, confirming the importance of how violations of the MCAR assumption and the distortion of the variance-covariance matrix with zero imputation that cause listwise deletion and zero imputation to yield inconsistent estimates. The estimates based on zero imputation and listwise deletion reported in Column 1-2 and 4-5 of Table 6, i.e. they are negative, differ considerably from estimates using multiple imputation (Column 3), which is positive; this suggests that the inferences made by researcher in innovation could be driven by how they chose to deal with missing innovation data.

The results using zero imputation and a dummy variable show no relation between market-to-book ratio and leverage, and a marginal effect of profitability on leverage for dividend-paying firms (Column 1). Listwise deletion leads to an insignificant relation between market-to-book ratio and leverage and between depreciation and leverage for dividend-paying firms (Column 2). Columns 3 and 6 of Table 6 presents the results for unreported R&D imputed using multiple imputation. In this case, there is a substantial change in the magnitude of the coefficients of all the explanatory variables. Most importantly, the relation between both investment variables and leverage is positive and internally consistent. Now R&D expenditure has a positive impact on leverage and not negative, as in Columns 1-2 and 4-5, which is congruent with pecking order theory. The estimates in Table 6 illustrate that the method used to handle missing R&D can lead

[15] The original Fama and French (2002) paper uses lagged variables and a Fama-Macbeth style regression in order to deal with autocorrelation of residuals at the firm level and heteroscedasticity. Panel A of Table IA3 in Internet Appendix II replicates the results of Table 3 Panels A and B in FF02, using lagged explanatory variables instead of contemporaneous ones, and Fama-Macbeth regressions instead of panel regressions. Panel B of Table IA3 in Internet Appendix II presents the results for the FF02 specification with lagged explanatory variables and two-way fixed effects. The results are qualitatively and quantitatively similar to FF02.

to substantially different inferences. Using multiple imputation for missing R&D in this setting potentially explains the puzzling findings in the original FF02 study.

## 7. Patents

So far, we have presented analyses using unreported R&D. Yet, 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. Unlike disclosure requirements for R&D expenditures, firms do not face an affirmative duty to seek patents, which potentially explains why the vast majority of US firms do not submit patent applications. Consequently, we conduct a similar analysis for patents as for unreported R&D. First, we investigate whether firms' decision to file for patents are *missing completely at random*, through the Little's test and regression analysis. Then we investigate the quality of patents that are not captured by the USPTO data and the performance of different imputation methods using non-USPTO patents of US firms as counterfactuals.

### *7.1 Predictability of Patent Applications (or non-patent firms)*

First, we investigate whether the observed variation in USPTO patent applications is systematically related to firm characteristics. The results of the Little (1988) test are presented in Panel A of Table 2 jointly for R&D and patents. The test results reject the null hypothesis that non-patenting firms are unpredictable. Second, we assess the existence of identifiable patterns in patent applications by conducting a regression analysis of missing patents on observable firm characteristics at the firm-year level for international and U.S. firms. We estimate a panel regression model with year, industry, and country fixed effects, where the dependent variable is missing patents. Missing patents is set equal to 1 when a firm does not file for a USPTO patent in a given year and zero otherwise.

Table 7 shows the effects of firm-specific characteristics on the filing of USPTO patents. Firm, industry, and country characteristics explain up to 25% of the variation in missing patents (Columns 1-3). For instance, missing patents increase at the firm level with PPE and ROA and decrease with total assets. Focusing on just the subset of US firms, these firm and industry characteristics explain a substantial amount of the variation in missing patents (31% to 77%; Columns 4-7). Missing patents increase with PPE and ROA, while decreasing with total assets and capital expenditure for U.S. firms. Collectively, our evidence indicates a significant correlation between the missing patents and firm-specific factors. Thus, the result appears to be inconsistent with patents *missing completely at random*.

*7.2 Relevance of Unreported Innovation via Patents*

Innovation-related studies, across accounting, economics, and finance, focus on patenting as the most important outcome of the research and development process. These studies mainly use data from the USPTO-NBER dataset. This dataset includes all firms that have applied for USPTO patents and the NBER has conducted extensive disambiguation of firm data. While this dataset has been instrumental in conducting the first pieces of research, it provides a partial view. For researchers interested in understanding and/or capturing a fuller extent of firm innovation activities, investigating patenting only through the USPTO will underestimate the innovative activity of many firms. Next, we investigate the properties of patents filed outside USPTO jurisdictions to understand the importance of non-USPTO patents.

Table 8 shows statistics related to patents filed with USPTO and other jurisdictions by US and non-US firms. Over 14,000 US-firms applied for USPTO patents in the sample period, 9,518 US-firms received patens abroad, while 1,676 non-US firms received USPTO patents and 1,758 non-US firms received non-USPTO patents. The total number of patents granted in non-USPTO

jurisdictions per year is substantial for both US and foreign firms. For instance, foreign firms are granted 20% more patents in non-USPTO jurisdictions than in the US. US firms also are granted, on average, 22 patents a year outside the US and 28.2 patents in the US. 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.

We use the sample of U.S. firms that patent abroad to investigate the different methods of handling unreported patents, similar to Panel C of Table 2. We impute observations without USPTO patents with zero, industry mean (two-digit SIC code), and multiple imputation. Multiple imputation is carried out with all the variables in Table 7, within the industry. Results in Panel B of Table 8 show that U.S. patents abroad are not equal to zero, they are different from the USPTO industry mean, but they are not statistically different from MI.

*7.3 Empirical Data-Based Simulation*

Patents and R&D expenditures may have different determinants and missingness levels. To understand the properties of the different methods for handling missing data in the patent setting, we replicate the empirical distribution-based simulation, with the USPTO patent data distribution. The empirical distribution is derived from a panel of 783 firms with non-missing information for all variables except USPTO patents for the period 1992 to 2012. We follow the same simulation procedure as described in Section 5.1. We analyze the case of 70% missing data, as Table 1 shows that patents exhibit very large levels of missingness. In addition, we only show the results for MAR and MNAR, since the analyses in Table 2 and Table 7 show that patent data is not *missing completely at random*. Table A2 in the Appendix presents the results of the simulation based on the patent empirical distribution. Under MAR, IPW and Heckman generate the highest biasness in coefficient estimates relative to both imputation and deletion. Focusing on MNAR, deterministic

imputation and multiple imputation both perform better than listwise deletion, IPW and Heckman approaches.

## 8. Conclusions and Recommendations

Most public firms do not report R&D expenditures, do not obtain patents, nor 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 (with a dummy variable). We study how these methods of handling unreported innovation affect our inferences about corporate research and development. More specifically, we explore the assumptions underlying different methods of handling unreported innovation, assess the biases that each of these methods introduces, and provide guidance for future research.

Instead of arising randomly, we document that unreported innovation is systematically correlated with several firm, industry, and country characteristics. Accordingly, eliminating firms without R&D or patents provides biased results, if a proposed innovation covariate is correlated with any of these predictor variables (e.g., firm size, leverage, profits, etc.). Because patent prevalence is even lower than the frequency of reported R&D, concerns about biases from deleting firms without patents is especially pronounced.

Using recovered R&D, which allows us to accurately measure unreported R&D expenditures in prior unreported years, we compare different methods of handling firms without reported R&D. These recovered R&D firms do not look like zero R&D firms nor do they appear similar to positive reporting R&D firms. The recovered R&D is also statistically different from the industry average. This finding is problematic for the common methods for handling unreported innovation, namely deleting the firms, classifying them as zero innovators or setting their R&D to the industry average and including a dummy term. Simulation results allow us to rank the biases

created from different methods of coping with firms without patents or reported R&D. For instance, replacing missing innovation with zeros underestimates true innovation and leads to biased R&D coefficient estimates (e.g. Table 4). To demonstrate the economic impact of these findings, we replicate an influential finance study (Fama and French, 2002) and explicitly show how different approaches to unreported innovation affect empirical inferences.

Innovation variables exhibit very high rates of unobservability. The most common methods to handle firms without observable innovation (R&D or patents) are excluding them (listwise deletion) and deterministic imputation (with zero or the industry mean). 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. The deletion of firms with unobservable innovation and their classification as non-innovators, even after including a dummy variable, can lead to biased coefficients of not only innovation, but also other explanatory variables. These traditional methods of handling unreported innovation do not work well in addressing unreported innovation when the selection is correlated with outcome variables of interest.

It is difficult to give definitive solutions to dealing with missing innovation across different datasets, countries and research settings. Our results using US data reveal that the two common methods to handling missing innovation data provide biased coefficient estimates and standard errors. Strikingly, across a wide range of specifications, multiple imputation exhibits the least bias and RMSE among the six methods we investigate. Importantly, MI is the solution to unreported innovation data that is used the least in finance and economics studies.

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 collection error from data providers, ii) managers not reporting R&D expenses due to zero (near zero) innovation, iii) strategic disclosure choices in reporting R&D expenses and patenting, iv) unsuccessful R&D, or v) firms filing for patents in alternative patent offices. Consequently, researchers should report both full and partial sample characteristics of the variables of interest. The level or degree of missingness of the innovation variable being used should be noted.
2. Researchers with missing innovation data should test if the missing data is predictable or MCAR. Little (1988) provides a test to determine if the data is *missing completely at random*. For Stata users, the *mcartest* command implements this test.
3. If the missing data is unpredictable or MCAR (maybe because the missing data stems from random collection errors by the data provider), then researchers could potentially delete or exclude the observations with missing data.
4. If the missing data is predictable, then researchers should attempt to predict missing innovation data using economically motivated observable variables. The predictive variables should be included as covariates in the regression and selection model. The researcher could use multiple imputation (for Stata users the MI command) to handle the missing observations.
5. If the missing data is 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 multiple imputation can often be unbiased for MNAR + MAR data even though the researcher assumes the data to be MAR.

> Conceptually, both Heckman and multiple imputation remain appealing, with both approaches involving assumptions and tradeoffs. Surprisingly, Heckman and Inverse Probability Weighting are the worst performers under MNAR in our simulations, with MI typically performing the best.

Overall, when missingness is beyond the researcher's control and its distribution is unknown, handling this missing data ultimately boils down to the assumptions and mechanisms of missingness. In finance studies, a researcher must often decide on the assumptions of MAR versus MNAR, which involves the use of MI versus IV respectively. Yet, the assumptions underlying both IV 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 because the covariates included in imputation models are often correlated with the determinants of missingness (Collins et al., 2001). Consequently, one might tilt towards the use of MI for missing innovation data. Yet, we recommend reporting both MI and IV estimates, coupled with a discussion of the plausibility of the underlying assumptions in the spirit of partial identification.

In summary, the proportion and distribution of missing innovation data in the sample should be reported. Researchers should conduct an analysis of the randomness and predictability of the missing innovation data in their sample. Researchers should consider performing simulations similar to ours, based on their own data, to choose between the various methods of handling unreported innovation.

## References


Anton, J. and D. Yao, 2004. Little patents and big secrets: Managing intellectual property, *Rand Journal of Economics* 35(1), 1-22.

Bartlett, J.W., C. Frost, and J.R. Carpenter, 2011. Multiple imputation models should incorporate the outcome in the model of interest, *Brain* 134(11), 189.

Bena, J., and K. Li, 2014. Corporate innovations and mergers and acquisitions, *Journal of Finance* 69(5), 1923-1960.

Bertrand, M., E. Duflo, and S. Mullainathan, 2004. How much should we trust differences-in-differences estimates? *The Quarterly Journal of Economics* 119(1), 249–275.

Chamberlain, G., 1984. Panel data, in *Handbook of Econometrics*, eds. Z. Griliches and M. D. Intrilligator, North-Holland Amsterdam, 1248-1318.

Collins, L.M., J. L. Schafer, and C. Kam, 2001. A comparison of inclusive and restrictive strategies in modern missing data procedures, *Psychological Methods* 6(4), 330-351.

Croce, M., T. Nguyen, S. Raymond, and L. Schmid, 2018. Government debt and the returns to innovation, *Journal of Financial Economics* 132(2), 205-225.

Fama, E. and K. French, 2002. Testing the trade-off and pecking order predictions about dividends and debt, *Review of Financial Studies* 15(1), 1-33.

Gow, I., D. Larker, and P. Reiss, 2016. Causal inference in accounting research, *Journal of Accounting Research* 54(2), 477-523.

Gow, I., G. Ormazabal, and D. Taylor, 2010, Correcting for cross-sectional and time-series dependence in accounting research, *The Accounting Review* 85(2), 483-512.

Hall, B. H., Jaffe, A. B., and Trajtenberg, M., 2001, The NBER patent citation data file: lessons, insights and methodological tools, Working Paper 8498, National Bureau of Economic Research.

Hochberg, Y., C. Serrano, and R. Ziedonis, 2018. Patent collateral, investor commitment, and the market for venture lending, *Journal of Financial Economics* 130(1), 74-94.

Hombert, J. and A. Matray, 2018. Can innovation help U.S. manufacturing firms escape import competition from China? *Journal of Finance* 73(5), 2003-2039.

Huang, J., 2018. The customer knows best: The investment value of consumer opinions, *Journal of Financial Economics* 128(1), 164-187.

Jiang, W., 2017. Have instrumental variables brought us closer to the truth, *The Review of Corporate Finance Studies* 6(2), 127–140.

Koh, P. S. and D. Reeb, 2015. Missing R&D, *Journal of Accounting and Economics* 60(1), 73-94.

Koh, P.S, D. Reeb, and W. Zhao, 2018. CEO confidence and unreported R&D, *Management Science* 64(12), 5461-5959.

Lahiri, K., and L. Yang, 2013. Forecasting binary outcomes, in *Handbook of Economic Forecasting* (Vol. 2B), eds. A. Timmermann and G. Elliott, Amsterdam: North-Holland, 1025-1106.

Larcker, D. and T. Rusticus, 2010. On the use of instrumental variables in accounting research, *Journal of Accounting and Economics* 49(3), 186-205.

Lerner, J. and A. Seru, 2015. The use and misuse of patent data: Issues for corporate finance and beyond, Working Paper, Harvard University.

Lisowsky, P., 2010. Seeking shelter: Empirically modelling tax shelters using financial statement information, *The Accounting Review* 85(5), 1693-1720.

Little, R.J.A., 1988. A test of missing completely at random for multivariate data with missing values, *Journal of the American Statistical Association* 83(404), 1198-1202.

Little, R.J.A. and D.B. Rubin, 2002. Statistical analysis with missing data, 2nd Edition. New York, NY: John Wiley & Sons, Inc.

Masulis, R. and E. Zhang, 2019. How valuable are independent directors: Evidence from external distractions, *Journal of Financial Economics* 132(3), 226-256.

Moons, K., R. Donders, T. Stijnen, and Jr F. Harrel, 2006. Using the outcome for imputation of missing predictor values was preferred, *Journal of Clinical Epidemiology* 59(10), 1092–1101.

Petersen, M.A., 2009. Estimating standard errors in finance panel data sets: Comparing approaches, *Review of Financial Studies* 22(1), 435–480.

Png, I., 2017. Law and innovation: Evidence from state trade secrets laws, *Review of Economics and Statistics* 99(1), 167-179.

Rubin, D.B. 1976. Inference and missing data, *Biometrika* 63(3), 581–592.

Rubin, D.B. 1987. Multiple imputation for nonresponse in surveys. New York, NY: John Wiley.

Schafer, J. and J. Graham, 2002. Missing data: Our view of the state of the art, *Psychological Methods* 7, 147-177.

Sterne, J., I. White, J. Carlin, M. Spratt, P. Royston, M. Kenward, A. Wood, and J. Carpenter, 2009. Multiple imputation for missing data in epidemiological and clinical research: Potential and pitfalls, BMJ 338, b2393.

Thompson, S.B., 2011. Simple formulas for standard errors that cluster by both firm and time, *Journal of Financial Economics* 99(1), 1-10.

**Table 1**
**Sample Characteristics and Univariate Comparisons**

This table shows the sample characteristics and univariate comparisons. Panel A presents the sample characteristics. The sample period is 1999–2012. Panel B shows the difference in characteristics across different deletion methods. "Full Sample" uses all available observations without deletion based on either reported R&D or patent application information. "Report R&D" includes only observations that report R&D. "Report Patent" includes only observations that patent applications in any patent office, "R&D and Patent" includes only observations that have positive R&D and patent filings in the PATSTAT. Firm-years represent the maximum number of observations available for each subsample. Variable definitions are presented in Table A1. *, **, and *** represent 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 |
| Tobin's 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 |
| Capital Expenditure | 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 |
| No. of Patent Applications | 333,920 | 9.36 | 0.00 | 140.78 | 0.00 | 0.00 |
| No. of Patents Granted | 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 | 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%*** |
| Tobin's Q | 1.67 | 1.55 | 1.74 | 1.86 | 7%** | -4% | -11%*** |
| Leverage | 0.95 | 0.53 | 0.57 | 0.48 | 44%*** | 40%*** | 49%*** |
| Capital Expenditure | 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 | 122,546 | 118,264 | 53,456 | | | |

**Table 2**
**Predictability of Unreported R&D**

The table presents results on the predictability of unreported R&D. Columns (1)-(4) *World* present the results for all countries in the sample. Columns (5)-(7) present the results for the *US* only. Panel A present the results of the Little (1988) test for MCAR. Panel B presents OLS regressions of unreported R&D on financial accounting variables. Standard errors are double clustered at firm and time level. T-statistics are presented in brackets. Variable definitions are presented in Table A1. *, **, and *** represent significance at the 10%, 5%, and 1% levels, respectively. Adj. $R^2$ is the adjusted $R^2$.

*Panel A. Little's MCAR Test*

| | World | | | | US | | |
|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) |
| R&D($value) | **X** | **X** | **X** | **X** | **X** | **X** | **X** |
| Num. Patent Appl. | **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** |
| Chi-square dist. | 297 | 7,431 | 25,062 | 42,971 | 6,359 | 9,857 | 22,889 |
| D.o.F. | 2 | 9 | 87 | 326 | 5 | 23 | 180 |
| Prob>chi-square | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |

*Panel B. Panel Regression*

| | World | | | | US | | |
|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) |
| Ln(Total Assets) | -0.028*** | -0.030*** | -0.026*** | -0.011** | 0.046*** | -0.004 | -0.010** |
| | (-15.04) | (-13.80) | (-16.43) | (-2.87) | (10.98) | (-1.57) | (-2.95) |
| PPE | 0.374*** | 0.227*** | 0.189*** | 0.016* | 0.432*** | 0.321*** | 0.050** |
| | (27.17) | (15.64) | (18.44) | (1.98) | (13.69) | (9.74) | (2.29) |
| Leverage | 0.000 | -0.000 | 0.000 | -0.000 | 0.003 | 0.001** | 0.001*** |
| | (0.04) | (-1.15) | (2.03) | (-0.20) | (1.75) | (2.19) | (6.87) |
| CapEx | 0.003* | 0.003 | -0.001 | 0.000 | -0.261*** | -0.176*** | -0.015 |
| | (1.94) | (1.32) | (-0.72) | (0.17) | (-3.97) | (-3.45) | (-1.15) |
| ROA | 0.211*** | 0.180*** | 0.137*** | 0.023*** | 0.180*** | 0.160*** | 0.039*** |
| | (16.01) | (15.34) | (11.86) | (3.92) | (7.76) | (9.81) | (3.93) |
| Sales Growth | 0.011** | 0.008** | -0.002 | 0.002** | -0.008** | -0.007** | -0.001 |
| | (2.48) | (2.30) | (-0.75) | (2.57) | (-2.25) | (-2.79) | (-0.86) |
| Country FE | | | Yes | | | | |
| Industry FE | | Yes | Yes | | | Yes | |
| Year FE | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Firm FE | | | | Yes | | | Yes |
| N | 283,987 | 283,987 | 283,987 | 281,243 | 64,386 | 64,386 | 63,086 |
| Adj. $R^2$ | 0.05 | 0.23 | 0.38 | 0.81 | 0.11 | 0.53 | 0.93 |

**Table 3**
**Recovered R&D and Imputed Unreported R&D**

The table presents the Recovered R&D (in the years t-1 and t-2 from switch year) statistics and its comparison with different imputation techniques. Panel A presents the comparison of Recovered R&D and Zero R&D firm characteristics. Panel B presents the comparison of Recovered R&D and positive R&D firm characteristics. Panel C presents the comparison of Recovered R&D and Recovered R&D/Ln(Total Assets) with different imputation methods. *Recovered R&D* is the recovered R&D expenditure as reported in 10-K filings, its t-stat presents its difference from 0. *Industry R&D* is the average industry expenditure (two-digit) for the observations that are recovered, *MI R&D M1* is the multiply imputed R&D using ln(total assets), ROA, PPE, sales growth and leverage, by industry (two-digit), *MI R&D M2* is the multiply imputed R&D using the same model as M1 with the addition of the lagged R&D as conditioning information. "Diff." is the difference between Recovered R&D and imputed R&D. T-stats represent the t-statistic for the difference between Recovered R&D an imputed R&D.

| | N | MEAN | STD | N | MEAN | STD | Diff. | t-stat |
|---|---|---|---|---|---|---|---|---|
| *Panel A. Comparison with Zero R&D* | | | | | | | | |
| | Zero R&D | | | Recovered R&D | | | | |
| R&D ($ value) | 21,721 | 0.00 | 0.00 | 1,032 | 6.69 | 23.56 | 6.69 | 9.13 |
| R&D Expenditure | 21,721 | 0.00 | 0.00 | 1,022 | 0.87 | 20.11 | 0.87 | 1.38 |
| Ln(Total Assets) | 21,721 | 4.84 | 2.89 | 1,022 | 3.60 | 3.03 | -1.24 | -10.84 |
| ROA | 21,692 | -2.81 | 167.99 | 964 | -3.11 | 47.15 | -0.30 | -0.11 |
| PPE | 18,586 | 0.30 | 0.27 | 1,021 | 0.20 | 0.19 | -0.10 | -12.62 |
| Sales Growth | 19,505 | 1.54 | 84.27 | 804 | 15.31 | 410.93 | 13.77 | 0.91 |
| Capex | 21,367 | 0.06 | 0.46 | 956 | 0.06 | 0.09 | -0.01 | -0.83 |
| Leverage | 21,700 | 6.74 | 150.99 | 1,020 | 2.61 | 20.08 | -4.13 | -2.50 |
| *Panel B. Comparison with Positive R&D* | | | | | | | | |
| | Positive R&D | | | Recovered R&D | | | | |
| R&D ($ value) | 68,622 | 112.76 | 572.81 | 1,032 | 6.69 | 23.56 | -106.07 | -36.32 |
| R&D Expenditure | 68,622 | 0.34 | 10.54 | 1,022 | 0.87 | 20.11 | 0.53 | 0.79 |
| Ln(Total Assets) | 68,622 | 4.69 | 2.69 | 1,022 | 3.60 | 3.03 | -1.09 | -10.38 |
| ROA | 68,618 | -0.70 | 18.79 | 964 | -3.11 | 47.15 | -2.41 | -1.52 |
| PPE | 68,592 | 0.18 | 0.17 | 1,021 | 0.20 | 0.19 | 0.02 | 3.03 |
| Sales Growth | 60,879 | 1.49 | 69.65 | 804 | 15.31 | 410.93 | 13.82 | 0.94 |
| Capex | 67,945 | 0.05 | 0.07 | 956 | 0.06 | 0.09 | 0.01 | 1.57 |
| Leverage | 68,436 | 1.81 | 64.87 | 1,020 | 2.61 | 20.08 | 0.80 | 0.91 |

*Panel C. Comparison with Imputation*

| Variable | N | MEAN | STD | RD | Diff. | t-stat |
|---|---|---|---|---|---|---|
| Recovered R&D | 1032 | 6.69 | 23.56 | | | 9.13 |
| Industry R&D | 1028 | 77.35 | 90.62 | 6.69 | -70.66 | -24.36 |
| MI R&D M1 | 1024 | 15.52 | 328.32 | 6.69 | -8.83 | -0.87 |
| MI R&D M2 | 1024 | 12.17 | 242.56 | 6.69 | -5.48 | -0.73 |
| Recovered R&D/Ln(TA) | 1023 | 1.21 | 7.62 | | | 5.08 |
| Industry R&D/Ln(TA) | 1018 | 8.27 | 8.29 | 1.21 | -7.06 | -20.02 |
| MI R&D/Ln(TA) M1 | 1022 | 2.23 | 31.54 | 1.21 | -1.02 | -1.02 |
| MI R&D/Ln(TA) M2 | 1022 | 1.89 | 23.18 | 1.21 | -0.68 | -0.91 |

**Table 4**
**Simulation Based on the Empirical Distribution from Compustat Data**

This table 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 5.1. The empirical distribution is from the panel of 783 firms with non-missing information for all variables except R&D. The methods evaluated are listwise deletion (LD), imputation with zero (ImpZero), imputation with industry mean, two-digit SIC code (ImpMean), inverse probability weighting (IPW), Heckman procedure (Heckman), and multiple imputation (MI). MI uses all the variables in the regression and is estimated using MCMC with 200 iterations for convergence. The regressions for imputation with zero and industry mean include a dummy variable for the imputed observations. Absolute average represents the average of the absolute bias across all variables. We present results for three missingness mechanisms: missing completely at random (MCAR) in Panel A, missing at random (MAR) in Panel B, and missing not at random (MNAR) in Panel C. Variable definitions are presented in Table A1. We generate missingness R&D for 50 and 70% of the sample. We conduct 500 simulations.

| | | Missing 70% | | | | | | Missing 50% | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | LD | Imp Zero | Imp Mean | IPW | Heckman | MI | LD | Imp Zero | Imp Mean | IPW | Heckman | MI |
| | | | | | | *Panel A. MCAR* | | | | | | | |
| Bias | R&D | -0.04 | -0.65 | -0.56 | 2.55 | 2.12 | -0.14 | 0.03 | -0.60 | -0.45 | 2.52 | 1.95 | -0.10 |
| | Ln(Total Assets) | -0.29 | -0.25 | -0.24 | 9.20 | 9.00 | -0.03 | -0.05 | -0.19 | -0.17 | 9.23 | 9.05 | -0.02 |
| | Tobin's Q | 0.06 | 0.28 | 0.27 | 6.05 | 6.22 | 0.19 | -0.11 | 0.23 | 0.20 | 4.82 | 4.86* | 0.16 |
| | Leverage | -0.03 | -0.04 | -0.04 | 0.43 | 0.41 | 0.00 | -0.01 | -0.03 | -0.03 | 0.28 | 0.28 | 0.00 |
| | ROA | -0.01 | -0.02 | -0.02 | 0.64 | 0.77 | 0.13 | -0.06 | -0.01 | -0.01 | 0.67 | 0.66 | 0.07 |
| | ***Avg. Abs. Bias*** | ***0.09*** | ***0.25*** | ***0.23*** | ***3.77*** | ***3.70*** | ***0.10*** | ***0.05*** | ***0.21*** | ***0.17*** | ***3.51*** | ***3.36*** | ***0.07*** |
| RMSE | R&D | 0.01 | 0.00 | 0.00 | 0.01 | 0.01 | 0.00 | 0.01 | 0.00 | 0.00 | 0.01 | 0.01 | 0.00 |
| | Ln(Total Assets) | 0.03 | 0.01 | 0.01 | 0.10 | 0.09 | 0.01 | 0.02 | 0.01 | 0.01 | 0.10 | 0.09 | 0.01 |
| | Tobin's Q | 0.02 | 0.01 | 0.01 | 0.04 | 0.04 | 0.01 | 0.01 | 0.01 | 0.01 | 0.03 | 0.03 | 0.01 |
| | Leverage | 0.11 | 0.05 | 0.05 | 0.11 | 0.10 | 0.05 | 0.07 | 0.04 | 0.04 | 0.09 | 0.09 | 0.04 |
| | ROA | 0.21 | 0.06 | 0.06 | 0.23 | 0.21 | 0.08 | 0.12 | 0.06 | 0.06 | 0.17 | 0.16 | 0.06 |

*Panel B. MAR*

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Bias | R&D | 0.71 | -0.56 | -0.46 | 25.58 | 14.42 | -0.17 | 0.52 | -0.47 | -0.39 | 24.80 | 14.20 | -0.14 |
| | Ln(Total Assets) | 1.31 | -0.26 | -0.26 | 12.60 | 5.82 | -0.08 | 0.98 | -0.18 | -0.17 | 12.36 | 5.69 | -0.03 |
| | Tobin's Q | 0.77 | 0.25 | 0.23 | 31.83 | 17.66 | 0.00 | 0.64 | 0.13 | 0.11 | 34.42 | 17.45 | -0.05 |
| | Leverage | 0.19 | -0.06 | -0.05 | -4.79 | -2.62 | -0.01 | 0.17 | -0.04 | -0.04 | -4.48 | -2.69 | -0.01 |
| | ROA | 0.31 | -0.05 | -0.05 | -5.23 | -4.78 | -0.01 | 0.24 | -0.06 | -0.06 | -5.11 | -5.12 | -0.03 |
| | ***Avg. Abs. Bias*** | ***0.66*** | ***0.24*** | ***0.21*** | ***16.01*** | ***9.06*** | ***0.05*** | ***0.51*** | ***0.18*** | ***0.15*** | ***16.23*** | ***9.03*** | ***0.05*** |
| RMSE | R&D | 0.00 | 0.00 | 0.00 | 0.10 | 0.06 | 0.00 | 0.00 | 0.00 | 0.00 | 0.10 | 0.06 | 0.00 |
| | Ln(Total Assets) | 0.02 | 0.01 | 0.01 | 0.13 | 0.07 | 0.01 | 0.02 | 0.01 | 0.01 | 0.13 | 0.07 | 0.01 |
| | Tobin's Q | 0.01 | 0.01 | 0.01 | 0.21 | 0.14 | 0.01 | 0.01 | 0.01 | 0.01 | 0.22 | 0.14 | 0.01 |
| | Leverage | 0.05 | 0.03 | 0.03 | 0.83 | 0.76 | 0.03 | 0.04 | 0.03 | 0.03 | 0.77 | 0.72 | 0.03 |
| | ROA | 0.07 | 0.04 | 0.04 | 1.10 | 1.06 | 0.05 | 0.05 | 0.04 | 0.04 | 0.84 | 0.96 | 0.04 |

*Panel C. MNAR*

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Bias | R&D | 0.64 | -0.55 | -0.46 | 15.07 | 11.04 | -0.18 | 0.58 | -0.50 | -0.45 | 14.31 | 10.88 | -0.13 |
| | Ln(Total Assets) | 1.13 | -0.38 | -0.37 | 11.48 | 6.55 | -0.19 | 0.83 | -0.17 | -0.16 | 11.38 | 7.24 | -0.01 |
| | Tobin's Q | 0.74 | 0.25 | 0.23 | 16.80 | 8.10 | 0.01 | 0.65 | 0.12 | 0.10 | 19.12 | 10.67 | -0.06 |
| | Leverage | 0.18 | -0.04 | -0.04 | -2.96 | -1.13 | 0.00 | 0.17 | -0.03 | -0.03 | -2.74 | -1.55 | -0.01 |
| | ROA | 0.28 | -0.03 | -0.03 | -3.78 | -2.35 | 0.01 | 0.26 | -0.02 | -0.02 | -3.15 | -2.77 | 0.00 |
| | ***Avg. Abs. Bias*** | ***0.59*** | ***0.25*** | ***0.22*** | ***10.02*** | ***5.84*** | ***0.08*** | ***0.50*** | ***0.17*** | ***0.15*** | ***10.14*** | ***6.62*** | ***0.04*** |
| RMSE | R&D | 0.00 | 0.00 | 0.00 | 0.06 | 0.04 | 0.00 | 0.00 | 0.00 | 0.00 | 0.06 | 0.04 | 0.00 |
| | Ln(Total Assets) | 0.02 | 0.01 | 0.01 | 0.12 | 0.07 | 0.01 | 0.02 | 0.01 | 0.01 | 0.12 | 0.08 | 0.01 |
| | Tobin's Q | 0.01 | 0.01 | 0.01 | 0.11 | 0.07 | 0.01 | 0.01 | 0.01 | 0.01 | 0.12 | 0.08 | 0.01 |
| | Leverage | 0.05 | 0.04 | 0.04 | 0.51 | 0.39 | 0.04 | 0.04 | 0.03 | 0.03 | 0.47 | 0.40 | 0.03 |
| | ROA | 0.06 | 0.05 | 0.05 | 0.65 | 0.53 | 0.05 | 0.06 | 0.04 | 0.04 | 0.55 | 0.54 | 0.04 |

**Table 5**
**Simulation Based on Simulated Data**

This table provides the evaluation statistics, bias (relative bias over true parameter) and root mean squared error (RMSE) for the simulation based on simulated data, as described in section 5.2. The methods evaluated are listwise deletion (LD), imputation with zero (ImpZero), imputation with industry mean, two-digit SIC code (ImpMean), inverse probability weighting (IPW), Heckman procedure (Heckman), and multiple imputation (MI). MI is estimated using MCMC with 200 iterations for convergence. The regressions for imputation with zero and industry mean include a dummy variable for the imputed observations. We present results for three missingness mechanisms: missing completely at random (MCAR) in Panel A, missing at random (MAR) in Panel B, and missing not at random (MNAR) in Panel C. We generate missingness in x1 for 50 and 70% of the sample. We conduct 500 simulations.

| | | Missing 70% | | | | | | Missing 50% | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | LD | Imp Zero | Imp Mean | IPW | Heck man | MI | LD | Imp Zero | Imp Mean | IPW | Heck man | MI |
| *Panel A. MCAR* | | | | | | | | | | | | | |
| Bias | $\theta_1$ | 0.00 | -0.19 | -0.19 | 0.00 | 0.00 | -0.01 | 0.00 | -0.13 | -0.13 | 0.00 | 0.00 | -0.01 |
| | $\theta_2$ | 0.01 | 0.28 | 0.28 | 0.01 | 0.01 | 0.01 | 0.00 | 0.19 | 0.19 | 0.00 | 0.00 | 0.00 |
| RMSE | $\theta_1$ | 0.11 | 0.21 | 0.21 | 0.08 | 0.11 | 0.09 | 0.06 | 0.07 | 0.07 | 0.06 | 0.06 | 0.05 |
| | $\theta_2$ | 0.11 | 0.29 | 0.29 | 0.08 | 0.11 | 0.09 | 0.06 | 0.10 | 0.10 | 0.07 | 0.06 | 0.05 |
| *Panel B. MAR* | | | | | | | | | | | | | |
| Bias | $\theta_1$ | -0.15 | -0.23 | -0.23 | -0.16 | -0.08 | -0.08 | -0.11 | -0.16 | -0.16 | -0.11 | -0.09 | -0.05 |
| | $\theta_2$ | -0.12 | 0.12 | 0.12 | -0.12 | -0.08 | -0.05 | -0.08 | 0.04 | 0.04 | -0.07 | -0.06 | -0.04 |
| RMSE | $\theta_1$ | 0.17 | 0.24 | 0.24 | 0.18 | 0.17 | 0.10 | 0.13 | 0.17 | 0.17 | 0.13 | 0.12 | 0.08 |
| | $\theta_2$ | 0.15 | 0.13 | 0.13 | 0.15 | 0.16 | 0.09 | 0.07 | 0.06 | 0.06 | 0.07 | 0.07 | 0.05 |
| *Panel C. MNAR* | | | | | | | | | | | | | |
| Bias | $\theta_1$ | -0.17 | -0.28 | -0.28 | -0.19 | -0.13 | -0.10 | -0.13 | -0.19 | -0.19 | -0.13 | -0.11 | -0.05 |

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $\theta_2$ | -0.16 | 0.14 | 0.14 | -0.15 | -0.13 | -0.08 | -0.11 | 0.04 | 0.04 | -0.11 | -0.10 | -0.07 |
| RMSE | $\theta_1$ | 0.19 | 0.29 | 0.29 | 0.20 | 0.17 | 0.12 | 0.14 | 0.20 | 0.20 | 0.14 | 0.13 | 0.07 |
| | $\theta_2$ | 0.17 | 0.12 | 0.12 | 0.17 | 0.16 | 0.10 | 0.13 | 0.06 | 0.06 | 0.13 | 0.12 | 0.08 |

**Table 6**
**Imputation Effect on Empirical Inference**

This table replicates the results in Fama and French (2002) using different imputation methods and two-way fixed effects. We present the results of a contemporaneous regression with two-way fixed effects: $\frac{L_t}{A_t} = \beta_0 + \beta_1 \frac{V_t}{A_t} + \beta_2 \frac{ET_t}{A_t} + \beta_3 \frac{Dp_t}{A_t} + \beta_4 RDD_t + \beta_5 \frac{RD_t}{A_t} + \beta_6 \ln(A_t) + e_t$. The exact replication of Table 3 Panels A and B of Fama and French (2002) is provided in Table IA3 in the Internet Appendix II. "Zero" is the result for the sample with imputation with zero and a dummy variable, "Delete" presents the results for listwise deletion, and "MI" presents the results for multiple imputation. Multiple imputation is implemented using all the variables in the regression in the imputation. The dependent variable is book leverage $\frac{L_t}{A_t}$ at time *T*. $\frac{V_t}{A_t}$ is the market to book ratio, $\frac{ET_t}{A_t}$ is earnings before interest and taxes as a proportion of total assets, $\frac{Dp_t}{A_t}$ is depreciation as a proportion of total assets, $\frac{RD_t}{A_t}$ is the R&D expenses as a proportion of total assets, $RDD_t$ is a dummy variable equal to 1 if R&D expenditure is missing and has been imputed with zero, and zero otherwise, and $\ln(A_t)$ is the natural logarithm of total assets. Non-dividend payers include firms that do not pay dividend in year *T-1*. The sample period is 1965-1999. Standard errors are double clustered.

| | Dividend Payer | | | Non-dividend Payer | | |
|---|---|---|---|---|---|---|
| Variable | Zero | Delete | MI | Zero | Delete | MI |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Intercept | 0.305*** | 0.344*** | 0.366*** | 0.325*** | 0.394*** | 0.376*** |
| | (22.62) | (19.83) | (56.52) | (4.70) | (20.07) | (6.74) |
| $\frac{V_t}{A_t}$ | -0.001 | -0.001 | 0.001 | 0.027 | -0.004*** | 0.028** |
| | (-0.15) | (-0.47) | (0.40) | (1.32) | (-3.24) | (2.24) |
| $\frac{ET_t}{A_t}$ | -0.158** | -0.215** | -0.184** | -0.517*** | -0.301*** | -0.139 |
| | (-1.99) | (-2.66) | (-2.07) | (-3.15) | (-4.77) | (-0.54) |
| $\frac{Dp_t}{A_t}$ | -1.076*** | -0.059*** | -1.057*** | 0.691 | 1.984*** | 0.636* |
| | (-6.12) | (-0.30) | (-10.67) | (1.29) | (7.96) | (1.66) |
| $RDD_t$ | 0.070*** | | | 0.079*** | | |
| | (11.96) | | | (4.61) | | |
| $\frac{RD_t}{A_t}$ | -0.290*** | -0.435*** | 0.081*** | -0.702*** | -0.335*** | 0.955*** |
| | (-2.71) | (-4.54) | (3.91) | (-2.83) | (-3.33) | (3.45) |
| $\ln(A_t)$ | 0.041*** | 0.029*** | 0.038*** | 0.032*** | 0.013** | 0.022*** |
| | (29.95) | (13.61) | (96.36) | (4.54) | (2.60) | (4.98) |

**Table 7**
**Predicting Non-patent Seeking Firms**

The table presents OLS regressions of unreported patents and explanatory variables. Columns (1)-(4) *World* present the regression results for all countries in the sample. Columns (5)-(7) present the regression for *US listed firms* only. Standard errors are double clustered at firm and time level. T-statistics are presented in brackets. Variable definitions are presented in Table A1. *, **, and *** represent significance at the 10%, 5%, and 1% levels, respectively. Adj. $R^2$ is the adjusted $R^2$.

| | World | | | | US | | |
|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) |
| Ln(Total Assets) | -0.021*** | -0.021*** | -0.037*** | -0.011*** | -0.011** | -0.045*** | -0.023*** |
| | (-19.56) | (-17.99) | (-26.38) | (-8.02) | (-3.22) | (-15.81) | (-5.07) |
| PPE | 0.213*** | 0.171*** | 0.124*** | -0.006 | 0.322*** | 0.251*** | -0.032 |
| | (24.22) | (17.61) | (14.82) | (-1.54) | (15.94) | (9.31) | (-1.62) |
| Leverage | 0.000 | 0.000 | 0.000 | -0.000 | 0.002 | 0.000 | -0.000 |
| | (1.47) | (0.41) | (0.93) | (-1.18) | (1.64) | (1.16) | (-1.27) |
| CapEX | -0.001 | -0.000 | -0.002 | -0.000 | -0.216*** | -0.160*** | 0.024 |
| | (-0.67) | (-0.10) | (-1.30) | (-0.85) | (-4.15) | (-3.03) | (1.19) |
| ROA | 0.164*** | 0.131*** | 0.151*** | 0.015*** | 0.293*** | 0.210*** | 0.018* |
| | (15.14) | (14.05) | (16.94) | (3.34) | (9.28) | (7.79) | (1.82) |
| Sales Growth | 0.003* | 0.003* | -0.004*** | 0.002*** | -0.006 | -0.003 | 0.005*** |
| | (2.02) | (2.02) | (-3.80) | (4.12) | (-1.39) | (-1.39) | (3.09) |
| Country FE | | | Yes | | | | |
| Industry FE | | Yes | Yes | | | Yes | |
| Year FE | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Firm FE | | | | Yes | | | Yes |
| N | 283,987 | 283,987 | 283,987 | 281,243 | 64,386 | 64,386 | 63,086 |
| R2 | 0.04 | 0.14 | 0.25 | 0.76 | 0.05 | 0.31 | 0.77 |

**Table 8**
**Yearly Foreign Patents**

The table presents analysis using non-USPTO patents. Panel A presents the average characteristics of USPTO and non-USPTO patents per firm. We report the average total number of citations per year, the average total number of patents per year, and average the number of citations per patent per firm for four groups of firms. US firms patenting with USPTO, Foreign firms patenting with the USPTO, US firms patenting abroad and foreign firms patenting in non-USPTO. Panel B presents the comparison of years with only non-USPTO patents for US firms with different imputation methods. Non-USPTO is number of non-USPTO patents of US firms in years with only non-USPTO patents; its t-stat presents its difference from 0. Industry Patents is the average industry expenditure for the observations the year with non-USPTO patents only, MI Patent M1 is the multiply imputed non-USPTO patents using ln(total assets), ROA, PPE, capital expenditure, sales growth, and leverage by industry (two-digit), MI Patent M2 is the multiply imputed non-USPTO patents using the same model as M1 without sales growth and leverage as conditioning information, MI Patent M3 is the multiply imputed non-USPTO patents using the same model as M1 with the addition of R&D expenditure as conditioning information. Diff. is the difference between non-USPTO patents and imputed patents. t-stats represent the t-statistic for the difference between non-USPTO patents and imputed patents. *, **, and *** represent the statistical significance of the mean being different from zeros at the 10%, 5%, and 1% levels, respectively.

*Panel A. USPTO vs. non-USPTO Patents*

| | US Firm | | Foreign Firm | |
|---|---|---|---|---|
| | USPTO | Non-USPTO | USPTO | Non-USPTO |
| Total citation | 236.5*** | 1.2*** | 620.1*** | 17.3*** |
| Number of patents | 28.2*** | 22.0*** | 124.9*** | 147.3*** |
| Citation/patent | 11.9*** | 1.1*** | 6.8*** | 0.1*** |
| Obs. | 14,608 | 9,518 | 1,676 | 1,758 |

*Panel B. Comparison with Imputation*

| Variable | N | Mean | Std. Dev. | Non-USPTO | Diff. | t-stat |
|---|---|---|---|---|---|---|
| Non-USPTO | 6,478 | 11.79 | 93.38 | | | 10.16 |
| Industry Patents | 6,350 | 32.84 | 47.84 | 11.79 | -21.05 | -15.58 |
| MI Patent M1 | 6,448 | 12.88 | 83.19 | 11.79 | -1.09 | -0.69 |
| MI Patent M2 | 6,448 | 12.01 | 79.36 | 11.79 | -0.22 | -0.12 |
| MI Patent M3 | 6,448 | 12.89 | 82.54 | 11.79 | -1.10 | -0.70 |

**Figure 1**
**Reporting Innovation**

The figure presents data in four quadrants related to the reporting of innovation. The proportions are based on the Compustat Global sample and Compustat North America of 330,790 firm-year observations for the period 1999-2012. If R&D is missing in Compustat, we classify that as unreported. If a firm does not have patent applications in a particular year with PATSTAT, we classify that as no patents. Q1 presents the proportion of observations that both report R&D and apply for patents, Q2 presents the proportion of observations that report R&D but do apply for patents, Q3 presents the proportion of observations that do not report R&D and apply for patents, and Q4 presents the proportion of firms that do not report R&D and do not apply for patents.

| | |
|---|---|
| *Q.1 Report R&D and Patents*<br>(10%) Global Compustat/Patents<br>(18%) North America Compustat/Patents | *Q.2 Report R&D and No Patents*<br>(25%) Global Compustat/Patents<br>(28%) North America Compustat/Patents |
| *Q.3 Don't Report R&D and Patents*<br>(4%) Global Compustat/Patents<br>(5%) North America Compustat/Patents | *Q.4 Don't Report R&D and No Patents*<br>(61%) Global Compustat/Patents<br>(49%) North America Compustat/Patents |

**Figure 2**
**Distribution of Imputed R&D**

This figure presents the distribution of imputed R&D using multiple imputation methods. The data is based on Compustat North America from 1992 to 2016. Frequency represents the frequency with which Recovered R&D takes certain values (bars, left axis) and Cumulative % represents the cumulative distribution of Recovered R&D (line, right axis).

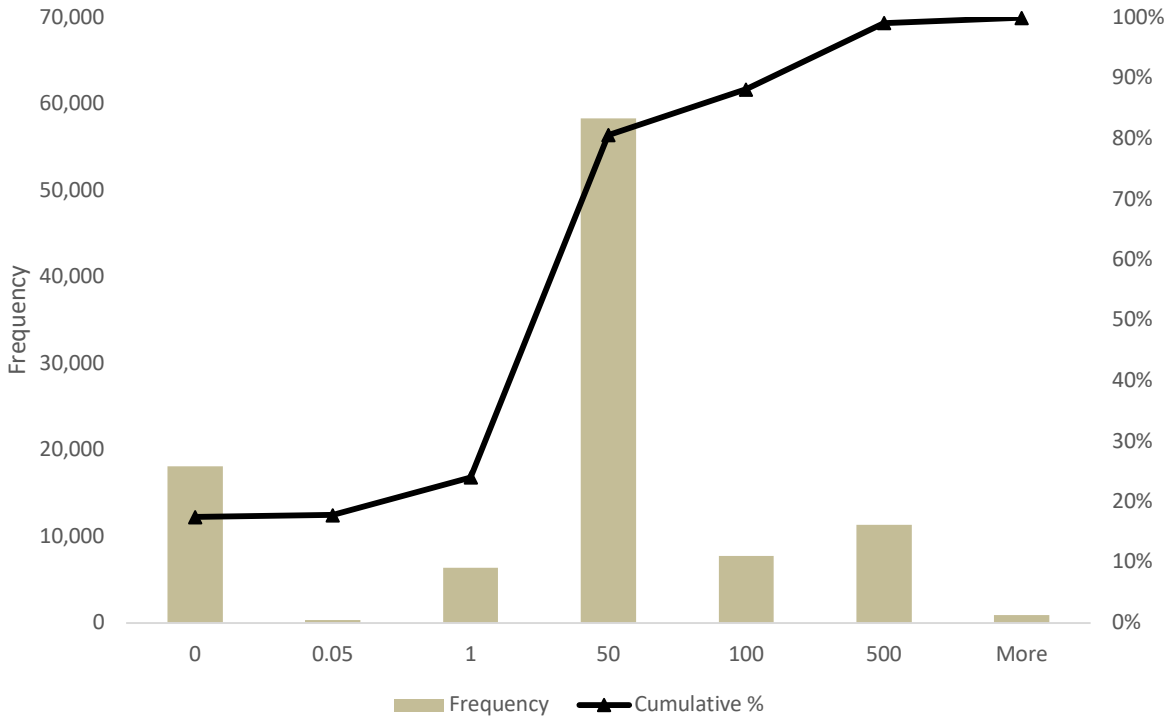
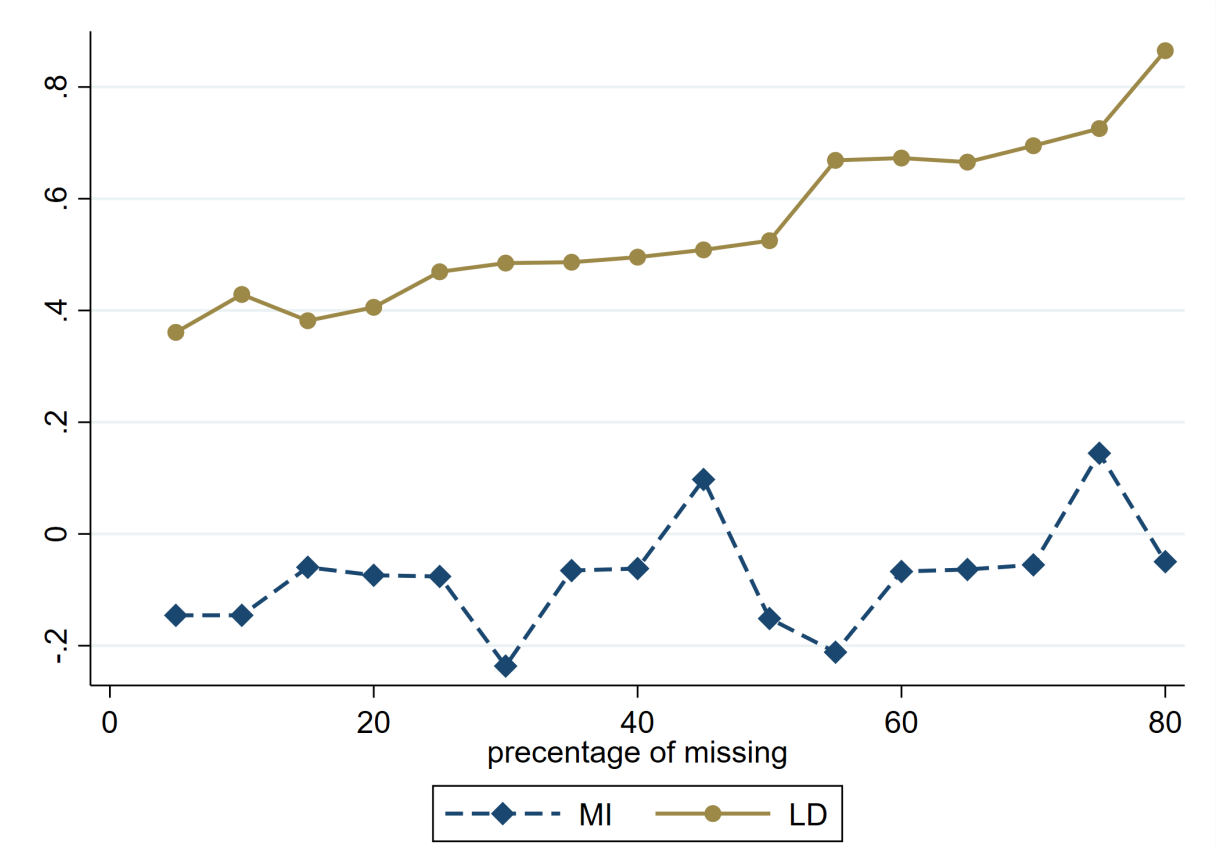


**Figure 3**
**Bias across Missingness Level**

This figure presents the bias of the R&D coefficient for listwise deletion (LD) and multiple imputation (MI) across different missingness levels. The simulation is based on the empirical distribution of the panel of 783 firms with non-missing information for all variables except R&D. MI uses all the variables in the regression in Section 5.1 and is estimated using MCMC with 200 iterations for convergence. We present results for data missing at random (MAR). We conduct 500 simulations.



**Appendix**

**Table A1**
**Variable Definitions**

This table shows the variable definitions.

| Variable Names | Variable Definitions | Code |
|---|---|---|
| R&D Expenditure | R&D expenditure divided by total assets | XRD/AT |
| Report R&D | Indicator variable: 1 if a firm reported zero or positive R&D expenditure; 0 otherwise | |
| PPE | Net property, plant, and equipment divided by total assets | PPENT/AT |
| Tobin's Q | Tobin's Q, measured as market value of equity divided by total assets | MKTVAL/AT |
| Leverage | Total liabilities divided by total assets | LT/AT |
| Ln (Total Assets) | Natural log of total assets | Ln(AT) |
| Capital Expenditure | Capital expenditure divided by total assets | CAPX/AT |
| ROA | EBIT divided by total assets | EBIT/AT |
| Sales Growth | Annual sales growth | $(Sale_t\text{-}Sale_{t-1})/Sale_{t-1}$ |
| HH Index | Herfindahl industry concentration index | |
| No. of Patent Applications | Total number of patent applications | |
| No. of Patents Granted | Total number of patents granted | |
| Citations | Total number of citations per patent | |

**Table A2**
**Patent Simulation Based on the Empirical Distribution of Data**

This table 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 USPTO data. The empirical distribution comes from the panel of 783 firms with non-missing information for all variables except USPTO patents. The methods evaluated are listwise deletion (LD), imputation with zero (ImpZero), imputation with industry mean, two-digit SIC code (ImpMean), inverse probability weighting (IPW), Heckman procedure (Heckman), and multiple imputation (MI). MI uses all the variables in sample and is estimated using MCMC with 200 iterations for convergence. The regressions for imputation with zero and industry mean include a dummy variable for the imputed observations. Absolute average represents the average of the absolute bias across all variables. Variable definitions are presented in Table A1. We present results for two missingness mechanisms: missing at random (MAR) in Panel A and missing not at random (MNAR) in Panel B. We generate missingness in patents for 70% of the sample. We conduct 500 simulations.

| | | LD | Imp Zero | Imp Mean | IPW | Heckman | MI |
|---|---|---|---|---|---|---|---|
| | | *Panel A. MAR* | | | | | |
| Bias | Patent | -0.07 | -0.38 | -0.37 | 0.32 | -0.21 | 0.04 |
| | Ln(Total Assets) | 0.29 | -0.29 | -0.30 | 53.54 | 2.98 | 0.41 |
| | Tobin's Q | -0.09 | 0.02 | 0.03 | -14.61 | -6.58 | -0.08 |
| | Leverage | -0.06 | -0.01 | -0.01 | 0.74 | -0.61 | 0.06 |
| | ROA | 0.01 | 0.00 | 0.00 | 0.84 | -0.57 | 0.00 |
| | ***Avg. Abs. Bias*** | ***0.10*** | ***0.14*** | ***0.14*** | ***14.01*** | ***2.19*** | ***0.12*** |
| RMSE | Patent | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | Ln(Total Assets) | 0.00 | 0.00 | 0.00 | 0.10 | 0.09 | 0.00 |
| | Tobin's Q | 0.00 | 0.00 | 0.00 | 0.03 | 0.10 | 0.00 |
| | Leverage | 0.03 | 0.01 | 0.01 | 0.08 | 0.49 | 0.01 |
| | ROA | 0.05 | 0.02 | 0.02 | 0.63 | 1.01 | 0.02 |
| | | *Panel B. MNAR* | | | | | |
| Bias | Patent | 0.43 | -0.30 | -0.35 | 1.04 | 8.47 | 0.16 |
| | Ln(Total Assets) | 1.97 | 1.12 | 1.13 | 47.68 | 44.42 | 1.82 |
| | Tobin's Q | 0.17 | 0.04 | 0.05 | -17.68 | -16.97 | -0.09 |
| | Leverage | 0.01 | 0.01 | 0.01 | 0.15 | 0.34 | 0.04 |
| | ROA | 0.00 | -0.01 | -0.01 | 0.75 | -0.08 | -0.01 |
| | ***Avg. Abs. Bias*** | ***0.52*** | ***0.30*** | ***0.31*** | ***13.46*** | ***14.06*** | ***0.42*** |
| RMSE | Patent | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | Ln(Total Assets) | 0.01 | 0.00 | 0.00 | 0.09 | 0.08 | 0.00 |
| | Tobin's Q | 0.00 | 0.00 | 0.00 | 0.04 | 0.04 | 0.00 |
| | Leverage | 0.02 | 0.01 | 0.01 | 0.04 | 0.06 | 0.01 |
| | ROA | 0.05 | 0.03 | 0.03 | 0.56 | 0.30 | 0.03 |