R&D activities narrative disclosure: A new measure

Panayiotis C. Andreou[a,b], Neophytos Lambertides[a], Anna E. Maruska[a]

Department of Finance, Accounting and Management Science

January 15, 2023

Preliminary Working Draft

Please do not cite or circulate without authors' permission

Contact: panayiotis.andreou@cut.ac.cy, n.lambertides@cut.ac.cy, az.maruska@edu.cut.ac.cy

a. Cyprus University of Technology, Department of Commerce, Finance and Shipping, Spyrou Araouzou 115, Limassol 3036, Cyprus.

b. Durham University, Durham University Business School, Mill Hill Lane, Durham, DH1 3LB, United Kingdom

R&D activities narrative disclosure: A new measure

Abstract

We construct a new measure of R&D activities using the text of 10-K filings. We validate our measure by showing that it correctly identifies narratives devoted to research and development activities, that it varies intuitively over time and across sectors and that it predicts future patents and citations in a manner that indicates firms' R&D efforts. We predict and find that the current R&D textual measure has a positive and significant association with market valuation measured by Tobin's Q for up to four years. This suggests that firms that communicate more about R&D related activities have an increased likelihood of generating positive market evaluation. The results are robust to a battery of tests, including the effects of non- patented and R&D expensed firms.

Keywords: R&D; narrative disclosure; textual analysis.

1 INTRODUCTION

Recent evidence raised the discussion on whether firms and policymakers should concern themselves with R&D activities or the much-wider process of innovation (Schot and Steinmueller, 2018). On a firm-level, there has been criticism that R&D is not followed by more innovative output and better productivity. Particularly, Koh and Reeb (2015) examine whether missing R&D expenditures in financial statements indicates a lack of innovation activity. This is especially important as finance and accounting research interprets the blank R&D fields as firms with zero R&D, mainly proxying for absence of innovation. Koh and Reeb (2015) show that the non-reporting R&D firms file over 14 times as many patents as firms that report zero R&D expenditure and that these non-reporting firms obtained patents with broader contributions and greater citation breadth than zero R&D firms. Their findings concur that over 10% of the Compustat universe of missing R&D cases display substantial evidence they engage in innovation and R&D activities.

Koh and Reeb (2015) in order to provide further tests but mainly to provide a methodological approach of treating missing R&D observations in empirical research they generate the missing R&D observations using an empirical model based on the associations between missing R&D and firm (financial) characteristics (such as ROA, PPE, Leverage, and others). Their approach aimed at attempting to use as much of the Compustat information as possible to allow assignment of firms into the blank R&D category using an empirical model. Their results motivate researchers to treat missing R&D with care, however they are limited to relative cross sectional analysis and accounting information. Motivated by the need of developing an efficient processing of treating missing or even zero R&D firms, in this study we employ textual analysis on 10-K filings to construct an R&D measure which is able to measure more accurately R&D-related activities and innovation of US public firms. Narratives are an important information source and a critical component that helps external stakeholders to complement their understanding of firm's financial performance and objectives. Despite the recent arguments and criticisms, R&D investment and related disclosures continue to grow both within the firm and at a wider industry level.

To construct our measure, we first develop bag of words consisting of core and contextual words defining and describing R&D activities. Core words are nouns that encapsulate the main definition of research and development, and contextual words are adjectives or verbs that are identified to be able to describe various types of research and development activities or the stage of these activities. The advantage of this method is that it uses a "multiword" phrase-level analysis that

retains better the intended meaning than single-word-level analysis (Loughran and McDonald, 2011) that strips a word of its linguistic context. Our measure of R&D activities is based on the proportion of times that certain core-contextual word pairs occur in the 10-K filings. As far as we know, this is the first paper that develops a bag-of-words based textual measure by utilizing contextual information from official definition documents such as the OECD Fascati Manual 2015 (that was revised several times) and national survey instruments such as those from National Center for Science and Engineering Statistics (NCSES) and National Science Foundation (NSF).

To validate our measure, we implement several validation techniques to verify that it correctly captures R&D activities. First, we show that the R&D narrative measure achieves an acceptable level of reliability and yields consistent results. Specifically, we show that our textual measure is highly persistent over time, consistent with the general assertion that changes in R&D are quietly slow. Second, we assess the methodological quality of our measure by examining the degree of R&D activities in certain U.S. states and/or industries. We show that our textual R&D measure is more (less) pronounce in U.S. states and industries traditionally spending large (small) amounts in R&D (Ortega-Argilés and Brandsma, 2008; Bellstam et al., 2020). Third, we test the predictive ability of our textual R&D measure, and we find that it foresees future patents and citations. The findings demonstrate the predictive validity of this new measure over and beyond previous measures such as the one developed by Merkley (2011).

Finally, assuming priori that more R&D narratives or real R&D expense must lead to more patents/citations or future cash flows, it must also lead to value creation. R&D investments are considered as investments in intangible assets that contribute to the long-term growth of the firm (Chan et al., 2001). If such investment results in an innovative product, service or process that enables the firm to enhance its intangible assets, then the firm will differentiate itself from other firms. Especially if the firm effectively communicates its innovative activities, it has an increased likelihood of generating positive market evaluation which in turn can increase the firm's value (Ehie and Olibe, 2010; Majumdar et al., 2019). Previous empirical evidence demonstrates that a firm expecting successful patenting outcomes, will also expect a positive influence on its equity stock. This is because a patented technology undoubtedly signifies a potentially valuable resource and is likely to form the basis of a sustainable competitive advantage if it is valuable, rare, inimitable, and imperfectly substitutable (according to the United States Patent and Trademark Office definition for granting patents). Furthermore, subsequent citations of a given patent, which are a meaningful quality weight for older patents, signifies that the patent granted had indeed proven to be valuable. Hall et al. (2005) use patent and citations and find a significant effect on market valuation (measured by Tobin's Q).

Taken together with the above validations, we use alternative models of Tobin's Q to validate our R&D narrative measure and indicate that the R&D topic contains useful information.

Our study contributes to the literature in several ways. First, we contribute to the extant literature on research and development. Prior studies have showed that firm value depends on R&D expenditures (Eberhart et al., 2004, 2008), innovative efficiency (Hirshleifer et al., 2013), patent citations (Gu ,2005; Belenzon, 2012) and industry-level technological progress (Matolcsy and Wyatt, 2008). In this context, our paper demonstrates that narrative R&D disclosures can be insightful given the high complexity and uncertainty of R&D activities. This is important because there is scarce research on R&D using qualitative R&D disclosures as a proxy. Second, our study adds to the limited literature on R&D narrative disclosure. Nekhili et al. (2012) investigate the impact of R&D narrative disclosure on the market value of equity for a sample of French companies during the period 2000–2004. Also, La Rosa and Liberatore (2014) examine the effect of regulatory regime (mandatory versus voluntary) of R&D narrative disclosures on the cost of equity capital for biopharmaceutical and chemical listed companies from eight Western European countries across the period 2005–2009. However, these studies use a small sample period and focus on specific regions/countries and industries/sectors (Gu and Li, 2003; Jones, 2007), and hence lack the necessary power to generalize their inferences across different firms. We are the first to use a larger sample of firms and longer period (1995-2020) and thus we are able to eliminate measurement error and selection bias.

Furthermore, while our study is closely associated with the work of Merkley (2014), it is, however, very distinct. Specifically, Merkley (2014) examines the determinants of R&D disclosures while we study the impact of such disclosures. Merkley (2014), along with many other studies, use disclosure as a dependent variable and financial performance as an independent variable, while we test the opposite relation (Lev et al., 1996; Chan et al., 2001; Rutherford, 2018). We argue that by focusing on qualitative R&D disclosures, we are able to expound whether these are informative enough to predict future patents and affect the firm's market value. We provide supporting evidence that disclosure policy is an effective mechanism through which firm-specific information can be conveyed to outside investors. In that way investors' dependence on common information signals is reduced, firm opacity is improved, and firm valuation is ameliorated.

2 LITERATURE REVIEW

2.1 R&D current measurement issues

An important issue facing the U.S. and international accounting standard setters is the financial reporting of corporate R&D expenditures. Generally accepted accounting principles in the United

States (US GAAP) require public corporations to expense all internal R&D outlays (i.e to subtract from revenues (sales) in the process of calculating net income (earnings)). The major characteristic of an expense that differentiates it from an asset (capital) is that it is not supposed to generate future benefits. Although there is no doubt that R&D activities are expected to produce future benefits, such as from sales of drugs or software products, accounting conservatism calls for expensing R&D activities because individual R&D projects are highly uncertain (Kothari et al., 2002). By contrast, the capitalization regime (that is, the recognition that R&D expenditures constitute an asset that is expected to provide future benefits) has been implemented only in certain circumstances. For example, SFAS No. 86 (FASB 1985)[1] requires software companies to capitalize their development costs after reaching technical feasibility. IAS No. 38 (IASB 1998)[2] requires companies to expense all research costs, but to capitalize their development costs after establishing technical and commercial feasibility. A recent study by Mazzi et al. (2022), contrasts the thinking of the standard setters in the historical development of the standard (especially of the capitalization of development costs) with buy-side and sell-side equity investors through interviews. Results show that investors find R&D accounting information useful for decision making and are supportive of the principle of the mandatory capitalisation of development costs, but highly critical of the conditions specified in the standard. This is due to the vagueness and subjectivity currently in the standard as well as the possible manipulation facilitating earnings management. As a result, the signalling of future value creation to them, and hence decision-usefulness to them, of capitalised development costs is undermined with consequential demands for increased wider voluntary disclosure. The on-going debate among academics and practitioners on capitalization versus expensing of R&D activities lacks direct evidence on the uncertainty of future earnings and cash flows attributable to current R&D expenditures and current accounting system makes it hard to compare the R&D activities of different firms.

Due to these complexities, practitioners, policy makers and academics have mostly relied on input measures of R&D expenditures, or output metrics based on patents to understand the nature and effectiveness of firms' R&D (Leuz and Wysocki 2016; Roychowdhury et al. 2019). The problem with these measures is that they do not capture the entire extent of a firm's R&D activities. Not all companies that engage in R&D activities choose to disclose R&D expenditures in their financial statements. Similarly, not all companies that engage in R&D activities pursue the filing of patents. Since neither the reported R&D expenditure in financial statements (Koh and Reeb, 2015) nor the patent filings capture the entire scope of R&D activities within the firm (Manso et al., 2017), narrative R&D disclosures may reveal valuable additional insights that would be of interest to explore. FASB

[1] Financial Accounting Standards Board (FASB) (1985) Statement of financial accounting standards no. 86, accounting for the costs of computer software to be sold, leased, or otherwise marketed. FASB, Norwalk.
[2] Lev B (2004). Sharpening the intangibles edge. Harv Bus Rev 82(6):109–116, 138.

proposed in August 2001 to report quantitative or qualitative information about intangibles in the notes to the financial statements, to improve the quality of information provided to investors and creditors, but this was never officially applied (Ciftci and Zhou, 2016).

Because of the difficulty in measuring qualitative information, early studies that assess the qualitative aspects of the disclosure mostly employed hand-collected data and analyzed small samples. For example, Enache and Hussainey (2019) use hand-collected data from 10-K to build the disclosure index for the biotech industry. Some other studies use data provided by experts such as financial analysts in order to code the quality of disclosure (e.g., AIMR scores[3]) (Jiao, 2011; Bhattacharya et al., 2013). Recognizing these constraints, Core (2001) recommends the use of text-based measures in corporate finance from other disciplines such as computer science, computational linguistics, and artificial intelligence.

2.2 10-K Disclosure Development and Importance

The 10-K report is one, among many other, firm-issued disclosures. Other types of disclosure is earnings announcements and press releases. However, a significant stream of research corroborates that 10-K filings contain important price-relevant information that investors trade on (e.g., Huddart, Ke, and Shi 2007; You and Zhang, 2009). A recent body of research finds that investors' ability to deal with relevant information in the 10-K is impaired by disclosure complexity (e.g., Lehavy et al., 2011; Rennekamp, 2012) and disclosure volume is a vital element of that complexity (e.g., You and Zhang, 2009; Miller, 2010). 10-K filings have grown in length over the past recent decades (Cazier and Pfeiffer, 2015). As they have grown in length, content has been added in response to a number of (supposed) information demands from different users and some additional requirements under recent companies acts and listing rules. However, the narrative reporting is expected to increase in the future in relation to the assumed need to justify numerous aspects of activities not amenable to numerical conveyance. Previous studies have examined some of the trends and changes in different narrative reporting such as risk reporting (e.g., Cabedo and Tirado, 2004; Linsley and Shrives, 2006; Woods, 2004) chairman's statement (e.g., Rippington and Taffler, 1995; Smith and Taffler, 2000) and CSR reporting (e.g., Miles et al., 2002; Solomon and Solomon, 2006).

However few studies have focused on 10-K filings due to their length and complexity. 10-K filings can be informative for different type of buyers/sellers of patents or simply investors interested in an investment opportunity. For example, if an investor is looking for a new technological

[3] Disclosure quality is measured by analysts' evaluations of firms' various disclosure activities compiled by the Association for Investment Management and Research (AIMR scores). It is mainly large firms that are evaluated by AIMR between 1986 and 1996. These scores are not available after 1996.

development that she thinks might be a good investment opportunity, she can look into 10-K disclosures (through a text-based search) as a starting point to identify businesses that potentially use the technology in question. Then the patent buyer/investor has a better-defined set of firms to further check whether the technical details match the intended patented technology or the technology of interest. In the same line, we utilize firms' 10-K filings as they provide a rich volume of firm-specific narrative disclosures towards the context of R&D activities that might be of interest to certain types of investors.

Following Antweiler and Frank (2004) and Tetlock (2007), a large literature in finance uses textual analysis from financial news, social media, and company filings to forecast stock prices and studies the causal effect of new information. Also, Balahur et al. (2010) analyse the content of news articles to identify positive and negative tone according to the number of positive and negative words used. For instance, consider the sentence "The market currently loves Amazon and hates Apple.", which expresses a positive sentiment towards Amazon and a negative sentiment towards Apple. Sentiment can be defined as the explicitly stated feeling depending on the use of words in a specific context. In the context of finance, investor sentiment is the sense about the expected favourable or unfavourable direction of events that is mostly based on rumors rather than market driven information.[4]

3 Measuring Narrative R&D Disclosure

We define narrative R&D disclosure as the proportion of R&D related words in the 10-K.[5] To identify R&D-related words, we use the National Science Foundation(NSF) (2014) that compiles all the official

[4] The preference of certain words usage and particular combinations to represent meanings is underused in management studies. The paper of Illia et al. (2014) analyses the co-occurrence of words for two companies in the biometric industry. There is ambivalence towards the biometric industry due to the use of technologies such as facial recognition and finger-printing by governments and private companies. Proponents argue that biometric devices aid in the functioning of modern societies, whereas opponents criticize the devices as being unstable and unreliable and for violating an individual's privacy (Ackleson, 2003). Illia et al. (2014) analyse press releases of two companies in this sector, one with minimal criticism and one with high media criticism although both companies offer exactly the same products and services. They adopt ALCESTE, a computerized text analysis software package developed by Max Reinert (1993) that allows researchers to study co-occurrence based on positioning text analysis. After having scanned the text and lemmatized words, ALCESTE splits the text into extracts called Units of Elementary Context. These extracts have the same number of words (around 16 to 19 words long) and include keywords that are analysed. ALCESTE classifies words in a descending hierarchical way and compares how words co-occur or do not co-occur in each extract. It has been previously used to analyse and compare speeches and texts from politicians, consumers, individuals or academics. This study investigates the way the words are positioned with regard to other words in a text which is critical in understanding the effects of language and corporate discourses (Lund and Burgess, 1996).

[5] Textual_R&D = Number of R&D related words from the BoW divided by the total number of words in the 10-K filing.

definitions of research and development in U.S. unedited and as they appear in their original sources.[6] We have identified verbs, adjectives and nouns used in the NSF document and employed Princeton University's WorldNet Lexical Database and Harvard IV-4 Psychosocial Dictionary to identify synonyms as it is likely that managers use variants of words in their disclosure. To cross-check our dictionary, we refer to the dictionary of commonly used R&D keywords and phrases developed by Merkley (2014).[7] In comparison to Merkley (2014), we include both singular and plural forms to capture all possible occurrences. For example, Merkley (2014) includes the singular form of "new technology", and "study" but not their plural form; "new technologies" and "studies". We also include more modern words that appear in recent 10-fillings such as "internet" and "software". To narrow down the words that are mostly used in financial context, we use words that appear more than 100 times in Loughran and McDonald (2011) word list. Once we have an exhaustive dictionary, we notice by eye that there are six (6) distinct pillars of R&D activities: "research", "development", "technology", "patents", "clinical", "collaboration". Each pillar includes a combination of words (or word pairings) that could be split into "core" and "contextual". Core words are nouns that stem from the element topic and contextual words are adjectives or verbs that describe research and development activities and their stage. Therefore, the final bag-of-words is based on actual usage frequency of the core-contextual word pairs that is most likely associated with the target construct. The complete bag-of-words can be found in Appendix 1. To quantify R&D activities, per annual 10-K filing, we count the occurrences whereby a contextual word appears within a span of ten lexical words (-10 and +10) of one of the core words, and then normalize the count based on the firm's 10-K filing length. To enhance precision, we ignore stop words such as "the", "are", "no" that precede the core word. To illustrate our method, below is an extract of Microsoft Corp. (2014) that signals its R&D activities through the use of R&D related words:

> *"**Developing new technologies** is complex and time-consuming. It can require long **development and testing** periods. Significant delays in new releases or significant problems in creating **new products** or services could adversely affect our revenue. We expect to continue making acquisitions or entering into **joint ventures** and strategic alliances as part of our long-term business strategy. We see significant opportunities for growth by investing **research and***

[6] NSF(2014) has compiled the definitions of R&D by extracting them from the Organisation for Economic Co-operation and Development (OECD) Frascati Manual 2015, from different sectors of the U.S. economy that perform or fund R&D (i.e businesses (I), the federal government and state governments (II), and academic and nonprofit organizations (III). Sources for definitions of R&D also include the Office of Management and Budget (OMB), federal procurement, tax and accounting guidance, and surveys from the National Center for Science and Engineering Statistics (NCSES), National Science Foundation (NSF). The NSF (2014) document is available here: https://www.nsf.gov/statistics/randdef/rd-definitions.pdf

[7] Merkley's (2014) dictionary of R&D related words is available in the Appendix to his paper (http://dx.doi.org/10.2308/accr-50649.s1).

development *resources in different areas…We will continue to make significant investments in* ***research, development****, and marketing for existing* ***products****, services, and* ***technologies****, including the Windows operating system, the Microsoft Office system..."*

Microsoft Corp has a dedicated Research and Development section in 10-K filings that outlines not only its current activity in product and service development but also its future long-term commitment to research and development with the aim of finding a unique perspective on future technology trend and contributing to innovation. More examples and extracts from 10-K filings can be found in Appendix 2.

The textual analysis method used in this paper has several advantages. First, it allows to sample from a broad range of publicly listed U.S. companies that make R&D investments. This is an improvement compared to prior empirical work that used a hand-collected sample and/or focused on general voluntary disclosures and restricted industries or countries. Second, this measure allows focusing on qualitative disclosures that firms provide alongside with accounting performance measures and reduces selection (bias) concerns as firms with material innovative investments are mandatorily required to provide such information through their 10-K filings. Firms' 10-K filings are provided concurrently with the audited financial statements and typically include more detailed information than the annual report to shareholders, which often appears as colourful and glossy publication.

To conduct textual analysis, we use Natural Language Processing (NLP) capabilities for re-developing and validating a measure for R&D disclosure because it is typically used with the "bag-of-words" model. NLP allows implementing "multiword" phrase-level analysis which retains better the intended meaning than single-word-level analysis that strips a word of its linguistic context. NLP technique has been applied to diverse topics such as online product reviews (Ullah et al., 2016), political speeches (Klebanov et al., 2008), press releases (Illia et al., 2012), among others. Even though textual analysis is widely implemented in finance research, only few textual implementations describe procedures of how the bag-of-words was built and how to make sure that the words selected accurately represent the studies constructs (Neuendorf, 2002). Following the suggestion by Short et al. (2010) for conducting validity testing, we provide reliability, content and predictive validity as described in section 4.

To lend credence to our method, Figure 1 Panel A shows the frequencies of the core words along with those of the core–contextual word pairs, respectively, for the 6 pillars related to R&D activities, as detailed in Appendix 1 (i.e., "research", "development", "technology", "patents", "clinical", "collaboration"). The first two pillars are general references to research and/or

development while the remaining four pillars are R&D-specific activities. The figure shows that core words alone appear many more times than core-contextual word pairs, indicating that in a vast majority of instances, core words are used in a context probably irrelevant to R&D related activities. This is particularly evident from element "technology" where technology-oriented words appeared alone 26,457,354 times, but when combined with the contextual words the number is reduced to 8,357,614 times. The only exception is the first element because the core word "research" usually appears before or after the contextual words and the same core word can be counted twice depending on the position of the contextual word. For example, in the sentence "*we have conducted research for our new product line*" there is a contextual word "conduct" before the core word "research" and another contextual word "new" after the word "research". In this instance, the core word is counted as one, but there are two contextual words that are combined with the core. This finding portrays that the core word "research" appears in a vast majority of instances in the correct context of R&D activities. In addition, the results in Panel A of Figure 1 show that firms involved in R&D activities do not necessarily utilize the first two pillars "research" and "development", indicating the importance of our R&D dictionary.

[Insert Figure 1]

Furthermore, in Panel B of Figure 1, we also checked whether the core-contextual word pairs that we use to quantify R&D activities overlap with the positive and negative sentiment dictionaries of Loughran and McDonald (2011), which has been extensively used in the literature. We do not find any noticeable overlap since only 1 stem word out of 164 core-contextual R&D related words was negative (claim*) and 5 out of 164 core-contextual R&D related words were positive (collabor*, improv*, advance*, innovat*, breakthough). This reassures that a widely used dictionary, such as that of Loughran and McDonald (2011), cannot serve our purpose and thus a tailor-made dictionary is necessary for capturing specifically R&D activities.

We then analyze R&D narratives over time. Figure 2, Panel A plots the time series of R&D narratives for our sample of 12,564 firms between 1995-2020. As seen in Panel A, there is little variation in R&D talks over time, though there is a modest upward trend in more recent years. This is consistent with prior research that an increase of narrative reporting is thought to be associated with the increased public scrutiny of enterprise activities and the assumed necessity to give explanations for numerous aspects of operations not amenable to numerical conveyance, especially after the effect

of the financial crisis (Samkin and Schneider, 2010). Such upward trend is also evident in other type of narrative disclosures such as in corporate social responsibility (CSR) disclosures which also experience substantial increases as firms' are aiming to influence society's perceptions toward corporate activities (Chu et al., 2013).

To provide further validation tests, in Panel B of Figure 2 we also (conceptually) replicate the main results of Merkley (2011) using his R&D bag-of-words methodology (found in the Appendix of Merkley (2011)). Our findings show a similar trend of R&D narrative disclosures through the years, however, the occurrence of such narratives seems more in line with our core-contextual methodology.

[Insert Figure 2 here]

4 DATA AND TEXT MEASURE OF R&D

We construct our sample based on the intersection of firm-years available on the EDGAR filings database, where we get the 10-K filings to measure R&D activities, and the Compustat annual file for the 1995-2020 period.[8] Our main sample includes 107,917 firm-year observations. Figure 3 illustrates the word cloud of R&D words; that is, clustering the words based on the frequency they are appeared in the pool of all 10-K filings of our sample. As seen (by eye), the most frequent words (i.e., biggest clusters) related to R&D are 'research', 'development', 'product', 'activities' and 'clinical'.

[Insert Figure 3 here]

Furthermore, Figure 4 shows that R&D narrative disclosure depends on the industry sector of the firms in which they operate. It is evident that the majority of the firms that refer to R&D activities are within the Pharmaceutical Product, Medical Equipment, Electronic Equipment and Business Services which is in line with prior studies (Merkley, 2010; Bellstam et al., 2020). Specifically, firms in the Pharmaceutical Products industry intensely refer to element "clinical", which is expected as it includes core terms words such as candidate(s), stud(y)(ies), trial(s), program(s)) and contextual terms such as product(s), drug(s), laboratory(ies), feasibility. Technology related element ("technology") are the most frequently occurring across all industries, more evidently in Computers and Equipment-related industries (i.e Electronic Equipment, Electrical Equipment, Measuring and Controlling Equipment etc.)

[8] Most EDGAR filings are not available prior to 1994.

[Insert Figure 4 here]

5 VALIDATION SECTION

Textual analysis is extensively employed in finance research (Loughran and McDonald, 2009; Hoberg and Phillips, 2010, 2016; Li et al., 2013; Hoberg et al., 2014). The development, refinement, and implementations of the coding scheme are key for the quality of textual analysis (Carley, 1993; Gephart, 1993). However, according to Short et al. (2010) only few papers that use textual/content analysis describe procedures of how the word dictionary is selected and many of these studies lack supporting validity of the textual measures (Neuendorf, 2002; Short el al., 2010; Weber, 1990). Considering these concerns, Short et al. (2010) recommend researchers to conduct validation tests mainly through the following five perspectives: content, reliability, external, dimensionality and predictive validity. Hence, in this section we conduct various validity tests to determine how well our text-based measure captures R&D related activities.

5.1 Reliability Validity

We begin with reliability assessment as it is essential to demonstrate that our new developed R&D narrative measure is capable to achieve an acceptable level of reliability and consistency (Peter, 1979). Prior studies suggest that firms' R&D activities are stable over the years, as accumulation of R&D capital and self-sustained engagement in R&D activities are long-lasting (Manez et al.,2015; Esteve-Pérez and Rodríguez, 2013). Thus, in principle, if our textual R&D measure is properly constructed to capture R&D activities then it should exhibit persistency over the years. In other words, if the measure is non-random, then it is expected to remain in the same portfolio decile in two subsequent periods with greater than 10% probability. Table 1 presents the mean annual transition probabilities by deciles of the textual R&D measure. Consistent with our expectations the results show that firms in the lowest (1_{st}) decile of the measure in a year have 69% chance to remain in the lowest decile in the following year, while firms in the highest (10_{th}) decile remain in the same decile the following year with 78% probability. In general, the transition probabilities of all diagonal elements are notably much higher than 10%, while the elements representing decile changes (off-diagonal elements in the matrix) exhibit rapid probability declines. These results indicate that our textual R&D measure is indeed persistent over time and lends support to the notion that it resembles the behaviour of an R&D-related variable.

[Insert Table 1 here]

5.2 Content Validity

In this section we assess the methodological quality of our measure using certain characteristics of various U.S. states. If our measure is properly constructed, it is expected to be more (less) pronounce in U.S. states traditionally spending large (small) amounts in R&D.[9] As seen in Figure 5, our R&D textual measure is commonly highlighted in U.S. states with similar levels of R&D intensity. Notably, our measure distinguishes the six U.S. states with the highest R&D investments: California, Massachusetts, Washington, New Jersey, Maryland and Utah. For example, California has the highest R&D intensity because it has the largest private and public support for research and development in the U.S (Von Zedtwitz and Gassmann, 2002).[10] The consistency in the findings indicates that our textual measure correctly measures the level of R&D intensity of U.S. states.

[Insert Figure 5 here]

Next, we look at the industry level to validate whether our R&D textual measure correctly captures the traditional level of R&D intensity of certain industries using the 48 Fama and French industry classification. As expected, Figure 6 indicates that the distribution (box plot) of our R&D textual measure is consistent with the (out of sample) industry distribution of R&D intensity as indicated by Ortega-Argilés and Brandsma (2008). Specifically, it is evident that the majority of firms talking intensively about R&D activities in their 10-Ks belong to traditional R&D industries such as Pharmaceutical Product, Electronic Equipment and Business Services. This finding is in line with prior studies (Merkley, 2010; Bellstam et al., 2020).

[Insert Figure 6 here]

[9] Reasons for why some U.S. states being more R&D intensive than others include cash grants, rebates, and R&D tax credits that attract a lot of businesses to relocate, expand, or stay in a specific locality (Wu, 2005). Most state governments offer R&D tax credits, many of which are tied to national tax credit levels. The only state that doesn't offer state R&D tax credit but is R&D active is Washington. This is because the federal government is in Washington and most of the R&D spending is used for the development of its own state. Around $15 billion a year is spent locally on R&D and being near to resources means greater chance of getting an R&D incentive.

[10] California has high R&D intensity mainly due to the fact that California is the state with some of the world's best universities such as the Stanford University, UC Berkeley and Caltech, and therefore these research institutions do basic science and share it with private partners to stimulate technological breakthroughs. Universities and research institutions are a big part of what drives innovation in Silicon Valley. Universities offer the educational background and foundation of many of the scientists and specialists placed in the tech giants' R&D departments, who work towards revolutionary discoveries while university alumni often take the leap of starting their own business. Thus, it's not surprising that many well-known and valuable technology corporations, like Apple and Facebook are located in California (Mahou et al., 2022).

5.3 Predictive Validity

In this section we test the predictive ability of our textual R&D measure to foresee future patents and citations. Patents incentivize investors to invest in R&D and stimulate the interest for further discovery and production of new things (Lincoln, 1953; Antonipillai and Lee, 2016). Citations on the other hand are typically observed only years after the grant of the cited patent, establishing that the original R&D that led into innovation was feasible and commercially worth it. To do this test, we use two alternative proxies of future patents and citations as dependent variables. The first is the natural logarithm of the number of patents (*log(1+Patents$_{t+1 \rightarrow t+3}$)*) and second the ratio of the number of citations to the number of patents (*log(1+(Citations $_{t+1 \rightarrow t+3}$ /Patents$_{t+1 \rightarrow t+3}$)*)) summed over the next three years, following Kogan et al. (2017). Our model specification includes the following control variables; POSTONE (*Positive words) and NEGTONE (negative words)*, RDINTNS (R&D expense/ Total Assets), ADINTNS (Advertisement expense/ Total Assets), TANGB (Property, plant and equipment (net) scaled by total assets), *CASHTA*(*Cash/Total Assets)*, *LEV* (long-term debt over total assets), *AGE* (the natural logarithm of the number of years since the firm first appears in CRSP), log (*assets*) and SOFTWARE (capitalized software and development costs over total assets). All control variables have been winsorized at 1%. The definition of all variables is tabulated in Appendix 3.

Table 2 presents the descriptive statistics of our sample. Consistent with prior research, the firms that invest in R&D have a mean (median) of $5,829 ($368) million total assets and $3,020 ($162) million market value. The average (median) firm in the sample has been public for 17 (12) years and the average (median) Tobin's Q is 0.34 (1.18). The mean ratio of R&D expenses to total assets is 0.06, the mean ratio of advertising expenses to total assets is 0.01 and the mean ratio of debt-to-asset (leverage) is 0.56. The average firm in the sample has 10 patents and 141 citations. Finally, the mean of our textual R&D measure is 0.0061, implying that for the average firm in our sample 0.61% of the disclosure text is related to R&D.

[Insert Table 2 here]

Table 3 shows the results of our regression analysis based on the following model:

$$\begin{aligned} PATENT_{t+1,t+3} = {} & b_0 + b_1 TextRD_t + b_2 POSTONE_t + b_3 NEGTONE_t + b_4 TANGB_t + b_5 CASHTA_t \\ & + b_6 LEV_t + b_7 RDINTNS_t + b_8 ADVINTNS_t + b_9 AGE_t + b_{10} LnAT_t \\ & + b_{11} SOFTWARE_t + FirmFE \\ & + YearFE \end{aligned} \qquad (1)$$

where $PATENT_{t+1,t+3}$ is the natural logarithm of the number of patents (*log(1+Patents$_{t+1 \rightarrow t+3}$)*) or the ratio of the number of citations to the number of patents (*log(1+(Citations $_{t+1 \rightarrow t+3}$ /Patents$_{t+1 \rightarrow t+3}$)*))

summed over the next three years. All model specifications include standard errors clustered at the firm-level.

The results are consistent with our expectations. Specifically, our findings show that our textual R&D measure has a robust positive and statistically significant impact on the number of future patents and on the citations to patents ratio. All columns show that a standard-deviation increase in the textual R&D measure is associated with around 12% - 20% more patents and citation impact, an effect that is statistically significant at the 1% level. The relation remains statistically significant when controlling for the past number of patents and citations as well as firm and/or year fixed effects.

[Insert Table 3 here]

5.4 Are narrative efforts by firms valued by the market?

Next, we proceed to investigate the valuation content of our proposed textual R&D measure. To this end, in this section we examine whether firm's effort to disclose narrative R&D activities is valued by the market. Prior research is scarce on the value relevance of textual measures. For example, Majumdar et al. (2019) adopt a text mining-based approach and examine the relationship between Twitter related activities of manufacturing firms and the market reaction towards these firms. Their results indicate that firms that communicate more about new product developments using Twitter posts have an increased likelihood of generating positive market evaluation. Following the same logic, we are interested in testing whether a firm with more R&D narrative discussion in its 10-K filings receives higher market valuation. To test this assertion, we examine the impact of our textual R&D measure on several alternative measures of Tobin's Q (see Appendix) that are widely used in economics and finance (e.g., Kaplan and Zingales, 1997; Gompers et al.,2003; Bebchuk and Cohen, 2004; among others)[11]. Our first Tobin's Q measure is the market value of equity plus total assets minus common equity and deferred taxes divided by total assets. This way of calculating Tobin's Q has

[11] Extensive research recently has established that Tobin's Q is a valid measure of firm valuation (El Ghoul et al., 2017; Aboud and Diab, 2018; Li et al., 2018). Tobin's Q has traditionally reflected firm's future development expectations as it is derived from stock market prices (Li et al., 2019; Shan & McIver, 2011; Bozec, Dia & Bozec, 2010; Demsetz & Villalonga, 2001). Tobin's Q reveals the value investors allocate to a firm's tangible and intangible assets based on predicted future revenue and cost streams. It is a forward-looking measure of valuation (Anderson, Fornell, and Mazvancheryl 2004) that is affected by investor's perception (acumen, optimism or pessimism) and psychology about future events (herd behaviour, mistakes, manipulations, etc.) (Wei, 2007). According to Demsetz and Villalonga (2001), Tobin's Q is also favoured by most economists, who have a better understanding of market constraints. In addition, Tobin's Q is more appropriate than any accounting profit ratio where the latter is affected by accounting practices and different taxation systems that depend on the different ownership structure. Since Tobin's Q is not affected by accounting conventions, it can be used as comparison tool across industries (Chakravarthy 1986). Furthermore, a high Tobin's Q ratio shows that a firm has successfully leveraged its investments in a way that is more valued in respect to its market-value compared to its book-value (Kapopoulos and Lazaretou, 2007).

been widely used in recent papers (i.e., Bellstam et al., 2021; Jiao, 2011). Our second Tobin's Q is calculated by summing the market value of common equity, liquidating value of preferred stock and book value of debt scaled by total assets. Book value of debt is computed as the difference between current liabilities and current assets plus inventory plus long-term debt (Vomberg et al., 2015; Lee and Grewal, 2004; Titman and Wessels, 1988). The third alternative measure of Tobin's Q is calculated using the book value of assets minus book value of common equity plus the market value of common equity divided by total assets (Florackis, 2005; Belkhir, 2005; Barontini and Capri, 2006).

The results reported in Table 4 show an overall positive and statistically significant association between our textual R&D measure and following years Tobin's Q after controlling for firm, year and industry fixed effects. Except of the results of using the TobinQ1 (that show a weaker and less statistical significance than the results based on the other two definitions of Tobin's Q), we find that a standard-deviation increase in textual R&D variable is associated with around 14% increase in Tobin's Q. This suggests that firms that communicate more their R&D related activities have an increased likelihood of generating positive market value. This is because R&D narrative disclosures influence the perception of investors regarding firms' performance (Salvi et al., 2020).

[Insert Table 4 here]

A notable advantage of our narrative R&D measure is that it can be computed for firms that either have no patents (and thus no citations) or they have no R&D expenses. Thus, our measure is suitable to evaluate R&D activities for a broader set of firms. We highlight this feature by including the interaction term between our textual R&D measure and an indicator for no citations (i.e., a dummy variable that takes one if the firm has zero number of citations and zero otherwise). These results are reported in Table 5. Similarly we also include the interaction term between the textual R&D measure and an indicator for NonRDexpense (i.e., a dummy variable that takes one if the firm has zero R&D expenses in a given year, and zero otherwise). These results are presented in Table 6. The interaction terms provide a test for significant differences in the relation between the textual R&D variable and Tobin's Q between firms with and without citations (Table 5) and between R&D and non-R&D firms (Table 6). The results in both tables show that the interaction term is insignificant in all model specifications, indicating that the impact of the textual R&D measure on Tobin's Q is similar for firms with and without citations as well as for R&D and non-R&D firms.

[Insert Table 5 and Table 6 here]

6 CONCLUSION

In this research we propose a new measure to understand the R&D activities of firms through their 10-K reporting. First, we conduct a series of validation tests to ensure that our measure behaves in the correct intended way and then we utilize the new textual R&D measure to examine whether the market values firms' efforts in reporting narrative R&D disclosures. Our findings suggest that there is a positive association between divulging R&D related information and Tobin's Q.

Our research contributes to the growing literature on understanding the impact of R&D narrative disclosures. First, from our review of existing literature we find that R&D informativeness is impaired due to the perceived vagueness and subjectivity of the criteria currently in the standards and lack of mandated disclosure. Our study contributes to this issue by analyzing narrative R&D disclosures with consistent and validated methods. Second, this research shows that the use of the content of the 10-K fillings is associated with the firm value. With the growth of textual-analysis algorithms and several social media technologies, existing research has analyzed narrative disclosure but not with consistent and validated methods. Today's researchers use advanced text mining methods, such as LDA, which are more suitable for shorter texts rather than longer reports. Additionally, prior research that examined the role of narrative disclosures and firm value has mostly focused on the tone and volume aspect (Luo et al., 2013; Hitt et al., 2015; Ren et al., 2017) and not the content aspect. Fourth, we examine the largest dataset of firms reporting to SEC and spanning multiple time periods, which to the best of our knowledge, is the largest dataset analyzed in the area of narrative disclosures.

7 REFERENCES

Aalst, J.V., 2010. Using Google Scholar to Estimate the Impact of Journal Articles in Education. Educational Researcher 39(5), 387-400.

Abraham, Lincoln. 1953. In R.P. Pasler, M.D. Pratt, & L.A. Dunlap (Eds.), Complete Works of Abraham Lincoln, Vol. 5. Rutgers University Press.

Ackleson, J., 2003. Securing through technology? "Smart borders" after September 11. Knowledge, Technology and Policy 16, 56–74.

Andreou, P.C., Harris, T., Philip, D., 2020. Measuring Firms' Market Orientation Using Textual Analysis of 10-K Filings. British Journal of Management. 31, 872–895.

Antonipillai, J. and Lee, M.L., 2016. Intellectual Property and the US Economy: 2016 Update. Jointly produced by the Economics and Statistics Administration and the US Patent and Trademark Office.

Antweiler, Werner, and Murray Z. Frank. 2004. Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards. Journal of Finance 59 (3): 1259–94.

Arora, A., S. Athreye. 2012. *Patent Incentives: Returns to Patenting and the Inducement for Research & Development*. UK Intellectual Property Office

Arora, Ashish, W.M. Cohen, J.P. Walsh. 2014. *The Acquisition and Commercialization of Invention in American Manufacturing: Incidence and Impact*. NBER Working Paper No. 20264

Audretsch, D.B. and Feldman, M.P., 2004. Knowledge spillovers and the geography of innovation. In *Handbook of regional and urban economics* (Vol. 4, pp. 2713-2739). Elsevier.

Balvers, R.J., Gaski, J.F, McDonald, B., 2016. Financial Disclosure and Customer Satisfaction: Do Companies Talking the Talk Actually Walk the Walk? J Bus Ethics. 139, 29-45

Balvers, R.J., Gaski, J.F. and McDonald, B., 2016. Financial disclosure and customer satisfaction: do companies talking the talk actually walk the walk?. *Journal of business ethics*, *139*(1), pp.29-45.

Bauin, S., Michelet, B.; Schweighoffer, M. G.; & Vermeulin, P.,1991. Using bibliometrics in strategic analysis: "Understanding chemical reactions" at the CNRS. Scientometrics 22(1), 113-137.

Belkin, N.J. and Croft, W.B., 1992. Information filtering and information retrieval: Two sides of the same coin?. *C*ommunications of the ACM, 35(12), 29-38.

Berchicci, L., 2013. Towards an open R&D system: Internal R&D investment, external knowledge acquisition and innovative performance. *Research policy*, *42*(1), pp.117-127.

Bhattacharya, S. and Guriev, S., 2006. Patents vs. trade secrets: Knowledge licensing and spillover. Journal of the European Economic Association, 4(6), pp.1112-1147.

Biber, D., 1993. Co-occurrence patterns among collocations: a tool for corpus-based lexical knowledge acquisition. Computational linguistics, 19(3), 531-538.

Bindi, R., Calzolari, N., Monachini, M., Pirrelli, V., 1991. Lexical knowledge acquisition from textual corpora: A multivariate statistic approach as an integration to traditional methodologies. In Proceedings, Seventh Annual Conference of the UW Centre for the New OED and Text Research. Oxford, U.K.

Bourke, J. and Roper, S., 2017. Innovation, quality management and learning: Short-term and longer-term effects. *Research Policy*, *46*(8), pp.1505-1518.

Brown, B. et al. 2010. Music Sharing as a Computer Supported Collaborative Application, 179-198

Cabedo, J. and Tirado, J. M. (2004), 'The Disclosure of Risk in Financial Statements', Accounting Forum, 28/2: 181–200.

Cazier, R.A. and Pfeiffer, R.J., 2016. Why are 10-K filings so long?. Accounting Horizons, 30(1), pp.1-21.

Chu, C.L., Chatterjee, B. and Brown, A.,2013. "The current status of greenhouse gas reporting by Chinese companies: a test of legitimacy theory", Managerial Auditing Journal, Vol. 28 No. 2, pp. 114-139.

Church, K., Gale, W., Hanks, P. and Hindle, D., 1991. Using statistics in lexical analysis. Lexical acquisition: exploiting on-line resources to build a lexicon, 115, p.164.

Ciftci, M. and Zhou, N., 2016. Capitalizing R&D expenses versus disclosing intangible information. Review of Quantitative Finance and Accounting, 46(3), pp.661-689.

Cohen, W.M., Nelson, R. and Walsh, J.P., 2000. Protecting their intellectual assets: Appropriability conditions and why US manufacturing firms patent (or not).

Cook, D., Kieschnick, R., Van Ness, R., 2006. On the marketing of IPOs. Journal of Financial Economics 82, 35-61.

Cooper, J.R., 1998. A multidimensional approach to the adoption of innovation. *Management decision*.

Core, J.E., 2001. A review of the empirical disclosure literature: discussion. *Journal of accounting and economics*, *31*(1-3), pp.441-456.

DellaVigna, S., Kaplan, E., 2007. The Fox News effect: media bias and voting. Quarterly Journal of Economics 122, 1187 – 1234

Dumais, S., Platt, J., Heckerman, D., & Sahami, M. 1998. Inductive learning algorithms and representations for text categorization. Proceedings of the 7th International Conference on Information and Knowledge Management, 148–155.

Evans, M., Wayne M., Cates, C. L., & Lin, J. (2005, April). Recounting the court? Toward a text-centered computational approach to understanding they dynamics of the judicial system. Paper presented at the annual meeting of the Midwest Political Science Association, Chicago.

Fisher IE, Garnsey MR. 2010. Improving information retrieval from accounting documents: a prototype digital thesaurus for employee benefits. In 2010 AAA Midyear Meeting of the Information Systems and the Strategic and Emerging Technologies Sections, Clearwater, FL.

Foray, D., Mowery, D.C. and Nelson, R.R., 2012. Public R&D; and social challenges: What lessons from mission R&D; programs?. Research policy, 41(ARTICLE), pp.1697-1702.

Freimane, R. and Bāliņa, S., 2016. Research and development expenditures and economic growth in the EU: A panel data analysis. Economics and Business, 29(1), pp.5-11.

Gans, J.S., Hsu, D.H. and Stern, S., 2008. The impact of uncertain intellectual property rights on the market for ideas: Evidence from patent grant delays. Management science, 54(5), pp.982-997.

Gentzkow, M., 2006. Television and voter turnout. Quarterly Journal of Economics 121, 931 – 972.

Gerybadze, A., Hommel, U., Reiners, H.W. and Thomaschewski, D. eds., 2010. Innovation and international corporate growth. Heidelberg: Springer.

Graham, S.J., Grim, C., Islam, T., Marco, A.C. and Miranda, J., 2018. Business dynamics of innovating firms: Linking US patents with administrative data on workers and firms. *Journal of Economics & Management Strategy*, *27*(3), pp.372-402.

Hall, B.H., Jaffe, A.B. and Trajtenberg, M., 2000. Market value and patent citations: A first look.

Huddart, S., B. Ke, and C. Shi. 2007. Jeopardy, non-public information, and insider trading around SEC 10-K and 10-Q filings. Journal of Accounting and Economics 43 (1): 3–36.

Illia, L., Sonpar, K. and Bauer, M.W., 2014. Applying co-occurrence text analysis with ALCESTE to studies of impression management. British Journal of Management, 25(2), 352-372.

Jaffe, A.B. and De Rassenfosse, G., 2019. Patent citation data in social science research: Overview and best practices. *Research handbook on the economics of intellectual property law*.

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. Lecture Notes in Computer Science (ECML'98), 1398, 137–142.

Kearney C, Liu S. 2014. Textual sentiment in finance: a survey of methods and models. International Review of Financial Analysis 33: 171–185.

Kim, J. and Valentine, K., 2021. Financial Statement Information and the Market for Innovation.

Kim, S.B., Seo, H.C. and Rim, H.C., 2004. Information retrieval using word senses: root sense tagging approach. In Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, 258-265.

Klebanov,B.B, Diermeier, D., and Beigman, E., 2008. Lexical Cohesion Analysis of Political Speech. Political Analysis 16, 447-463.

Kleinknecht, A., Van Montfort, K. and Brouwer, E., 2002. The non-trivial choice between innovation indicators. Economics of Innovation and new technology, 11(2), pp.109-121.

Kogan, L., Papanikolaou, D., Seru, A. and Stoffman, N., 2017. Technological innovation, resource allocation, and growth. *The Quarterly Journal of Economics*, *132*(2), pp.665-712.

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

Kortum, S., 1993. Equilibrium r&d and the patent--r&d ratio: Us evidence. The American Economic Review, 83(2), pp.450-457.

Kothari, S.P., Laguerre, T.E. and Leone, A.J., 2002. Capitalization versus expensing: Evidence on the uncertainty of future earnings from capital expenditures versus R&D outlays. Review of accounting Studies, 7(4), pp.355-382.

Kwon, N., Zhou, L., Hovy, E., & Shulman, S. W. (2006). Identifying and classifying subjective claims. Proceedings of the 8th Annual International Digital Government Research Conference, 76–81.

Lehavy, R., F. Li, and K. Merkley. 2011. The effect of annual report readability on analyst following and the properties of their earnings forecasts. The Accounting Review 86 (3): 1087–1115.

Leuz, C. and Wysocki, P.D., 2016. The economics of disclosure and financial reporting regulation: Evidence and suggestions for future research. Journal of accounting research, 54(2), pp.525-622.

Levin, R.C., Klevorick, A.K., Nelson, R.R., Winter, S.G., Gilbert, R. and Griliches, Z., 1987. Appropriating the returns from industrial research and development. Brookings papers on economic activity, 1987(3), pp.783-831.

Li et al., 2010. Contextual Bag-of-Words for Visual Categorization. IEEE 21 (4).

Li F. 2010. Textual analysis of corporate disclosures: a survey of the literature. Journal of Accounting Literature 29: 143–165.

Linden G. et al. 2003. Amazon.com Recommendations. IEEE, 76-80

Linsley, P. M. and Shrives, P. J. (2006), 'Risk Reporting: A Study of Risk Disclosures in the Annual Reports of UK Companies', British Accounting Review, 38/4: 387–404.

Loughran, T., and B. McDonald.2011. "When is a Liability not a Liability? Textual Analysis, Dictionaries, and 10-Ks." Journal of Finance, 66, 35–65.

Lund, K., C. Burgess,1996. Producing high dimensional semantic spaces from lexical co-occurrence. Behavior Research Methods, Instrumentation, and Computers 28, 203–208.

Mahou, Y., Ben Youssef, S. and Ben Jebli, M., 2022. Inspecting the influence of renewable energy and R&D in defending environmental quality: evidence for California. *Environmental Science and Pollution Research*, pp.1-12.

Manual, F., 2015. Guidelines for collecting and reporting data on research and experimental development. URL: http://www. oecd. org/sti/frascati-manual-2015-9789264239012-en. htm.

Mazzi, F., Slack, R., Tsalavoutas, I. and Tsoligkas, F., 2022. Exploring investor views on accounting for R&D costs under IAS 38. *Journal of Accounting and Public Policy*, *41*(2), p.106944.

McGahan, A.M. and Silverman, B.S., 2006. Profiting from technological innovation by others: The effect of competitor patenting on firm value. *Research Policy*, *35*(8), pp.1222-1242.

McTavish, D., Pirpo, E., 1990. Contextual Content Analysis. Quality & Quantity 24, 245-265.

Megna, P. and Klock, M., 1993. The impact of intangible capital on Tobin's q in the semiconductor industry. *The American Economic Review*, *83*(2), pp.265-269.

Miles, S., Hammond, K. and Friedman, A. L. (2002), Social and Environmental Reporting and Ethical Investment, ACCA Research Report No. 77, Certified Accountants Educational Trust (London).

Mille, B.N et al., 2003. MovieLens Unplugged: Experiences with an Occasionally Connected Recommender Systems.

Miller, B. 2010. The effects of reporting complexity on small and large investor trading. The Accounting Review 85 (6): 2107–2143.

Nassirtoussi AK, Aghabozorgi S, Wah TY, Ngo DCL. 2014. Text mining for market prediction: a systematic review. Expert Systems with Applications 41(16): 7653–7670.

NSF, 2014. Definitions of Research and Development: An Annotated Compilation of Official Sources.

O'Leary DE. 2013. 'Big data', the 'Internet of Things' and the 'Internet of Signs'. Intelligent Systems in Accounting, Finance and Management 20(1), 53–65.

Pathan, Kamble,V., 2019. A Review Various Techniques for Content Based Spam Filtering.

Qin, H., 1999. Knowledge discovery through co-word analysis. Graduate School of Library and Information Science. University of Illinois at Urbana-Champaign. Library Trends 48 (1) Knowledge Discovery in Bibliographic Databases: 133-159.

Rennekamp, K. 2012. Processing fluency and investors' reactions to disclosure readability. Journal of Accounting Research 50: 1319–1354.

Rippington, F. and Taffler, R. (1995), 'The Information Content of Firm Financial Disclosures', Journal of Business Finance and Accounting, 22: 345–362.

Roychowdhury, S., Shroff, N. and Verdi, R.S., 2019. The effects of financial reporting and disclosure on corporate investment: A review. *Journal of Accounting and Economics*, *68*(2-3), p.101246.

Salvi, A., Vitolla, F., Giakoumelou, A., Raimo, N. and Rubino, M., 2020. Intellectual capital disclosure in integrated reports: The effect on firm value. Technological Forecasting and Social Change, 160, p.120228.

Samkin, G. and Schneider, A., 2010. Accountability, narrative reporting and legitimation: the case of a New Zealand public benefit entity. Accounting, Auditing & Accountability Journal.

Sarnoff, J. D. 2011. Derivation and Prior Art Problems with the New Patent Act. *PATENTLY-OPat. LJ*, *2011*, 12.

Schot, J. and Steinmueller, W.E., 2018. Three frames for innovation policy: R&D, systems of innovation and transformative change. *Research policy*, *47*(9), pp.1554-1567.

Scott, Steven L., and Hal R. Varian. 2014. Predicting the Present with Bayesian Structural Time Series. International Journal of Mathematical Modeling and Numerical Optimisation 5 (1–2): 4–23.

Sebastiani, F. (2002). Machine learning in automated text categorization. ACM Computing Surveys, 34, 1–47.

Shen, X., Tan, B. and Zhai, C., 2005. Context-sensitive information retrieval using implicit feedback. In Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval, 43-50.

Sinclair, J., 1991. Corpus, concordance, collocation. Oxford University Press.

Smadja, F., 1991. Macrocoding the lexicon with co-occurrence knowledge. Lexical acquisition: Exploiting on-line resources to build a lexicon, 165-189.

Smith, M. and Taffler, R. J. (2000), 'The Chairman's Statement: A Content Analysis of Discretionary Narrative Disclosures', Accounting, Auditing & Accountability Journal, 13/5: 624–46.

Solomon, J. F. and Solomon, A. (2006), 'Private Social, Ethical and Environmental Disclosure', Accounting, Auditing & Accountability Journal, 19/4: 564–91.

Stromberg, D., 2004. Radio's impact on public spending. Quarterly Journal of Economics 119, 189 – 221.

Tetlock, Paul C. 2007. Giving Content to Investor Sentiment: The Role of Media in the Stock Market. Journal of Finance 62 (3), 1139–68.

Thomas, M., Pang, B., & Lee, L. 2006. Get out the vote: Determining support or opposition from Congressional floor-debate transcripts. Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP'06), 327–335.

Twiss, B. (1992), Managing Technological Innovation, Pitman, London.

Von Zedtwitz, M. and Gassmann, O., 2002. Market versus technology drive in R&D internationalization: Four different patterns of managing research and development. *Research policy*, *31*(4), pp.569-588.

Wiegand, M., Balahur, A., Roth, B., Klakow, D. and Montoyo, A., 2010. A survey on the role of negation in sentiment analysis. In Proceedings of the workshop on negation and speculation in natural language processing, 60-68.

Woods, M. (2004), 'Accounting for Derivatives: An Evaluation of Reporting Practices by UK banks', European Accounting Review, 13/2: 373–91.

Yang, Y., & Liu, X. (1999). A re-evaluation of text categorization methods. Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'99), 42–49.

You, H., and X. Zhang. 2009. Financial reporting complexity and investor underreaction to 10-K information. Review of Accounting Studies 14 (4): 559–586.

Yu, B., Kaufmann, S., Diermeier, D., 2008. Classifying Party Affiliation from Political Speech. Journal of Information Technology & Politics 5(1).

Zairi, M., 1994. Innovation or innovativeness? Results of a benchmarking study. *Total Quality Management*, *5*(3), pp.27-44.

Zernik, U. ed., 1991. Lexical acquisition: exploiting on-line resources to build a lexicon. Psychology Press.

Appendix 1: R&D Wordlist

This table shows our R&D core-contextual wordlist. Following Loughran and McDonald (2011), we add the plural form of nouns, the simple past tense, the past participle, gerund and the third person present tense for verbs. Additionally, we include appropriate synonyms and words with similar meaning.

Element 1	Core	**research, rd&e, r&d**
	Contextual	product, products, project, projects, program, programs, facility, facilities, initiative, initiatives, center, centers, activity, activities, operation, operations, pipeline, pipelines, laboratory, laboratories, process, processes, processing, engineering, advance, advanced, advances, advancing, advancement, advancements, conduct, conducted, conducts, conducting undergo, underwent, undergoes, undergoing test, tested, tests, testing, develop, developed develops, developing, development, developments, basic, applied experimental, clinical, preclinical, joint, new, innovative, conceptual, service, services, method, methods, technique, techniques, experiment, experiments
Element 2	Core	**development**
	Contextual	Project, projects, program, programs, facility, facilities, initiative, initiatives, center, centers, activity, activities, operation, operations, laboratory, laboratories, process, processes, processing, design, designs, engineering, advance, advanced, advances, advancing, advancement, advancements, test, tested, tests, testing, continue, continued, continues, continuing, experimental, clinical, preclinical, new, innovative, service, services, method, methods, technique, techniques, experiment, experiments,

		component, component, device, devices, model, models, equipment, equipments, tool , tools
Element 3	Core	**technology, technologies, algorithm, algorithms, system, systems, product, products, software, internet, technological, scientific**
	Contextual	develop, developed, develops, developing, development, developments, advance, advanced, advances, advancing, advancement, advancements, improve, improved, improves, improving, improvement, improvements, expand, expanded, expands, expanding, test, tested, tests, testing, research, application, applications, innovation, innovations, knowledge, milestone, breakthrough, research, new
Element 4	Core	**Patent, patents, trademark, trademarks**
	Contextual	apply, applied, applies, applying, application, applications, claim, claimed ,claims, claiming, file, filed, files, filling, grant, granted , grants, granting, issue, issued, issues, issuing, issuance, issuances, receive, received, receives, receiving, award, awarded, awards, awarding, pending, product, products
Element 5	Core	**candidate, candidates, trial, trials, study, studies, program, programs, data, testing, testings**
	Contextual	Clinical, preclinical, pilot, safety, experimental, drug, drugs, product, products, laboratory, laboratories, feasibility, feasibilities, research
Element 6	Core	**Collaboration, collaborations, collaborative, venture**
	Contextual	Establish, established, establishes, establishing, establishment, establishments, initiate, initiated, initiates, initiating, initiative, initiatives, announce, announced, announces, announcing, announcement, announcements, research, joint

Appendix 2: Excepts from Selected 10-Ks

Name	Filing Date	Description
CELLDEX THERAPEUTICS, INC.	02/03/2009	The Company's deliverables under this collaboration primarily include an exclusive license to its CDX-110 product candidate and its EGFRvIII technologies, research and development services as required under the collaboration and participation in the joint clinical development committee.
FAIR ISAAC CORP	22/12/1999	Technological innovation and excellence have been goals of the Company since its founding. The Company devotes, and intends to continue to devote, significant funds to research and development to develop both new products and enhancements to its existing products. In addition, the Company has ongoing projects for improving its fundamental knowledge in the area of algorithm design, its capabilities to produce algorithms efficiently, and its ability to specify and code algorithm executing software.

COMMVAULT SYSTEMS INC	16/05/2008	Research and Development Our research and development organization is responsible for the design, development, testing and certification of our data management software applications. As of March 31, 2008, we had 241 employees in our research and development group, of which 61 are located at our Hyderabad, India development center. Our engineering efforts support product development across all major operating systems, databases, applications and network storage devices. A substantial amount of our development effort goes into certification, integration and support of our applications to ensure interoperability with our strategic partners' hardware and software products. We have also made substantial investments in the automation of our product test and quality assurance laboratories. We spent $26.9 million on research and development activities in fiscal 2008, $23.4 million in fiscal 2007 and $19.3 million in fiscal 2006.
HARRIS INTERACTIVE INC	31/08/2001	Our Internet-based and traditional market research and polling services include: - research studies conducted on specific issues for specific customers - custom research, - research studies on issues of general interest developed and sold to numerous clients - multi-client research, - research conducted for other research firms - service bureau research
SANGAMO BIOSCIENCES INC	29/03/2002	We are responsible for advancing product candidates into preclinical animal testing.
SANGAMO BIOSCIENCES INC	29/03/2002	In January 2001, we announced our first plant agriculture collaboration with Renessen LLC, a joint venture between Cargill and Monsanto Company…Registrant's divisions, subsidiaries and affiliates conduct research and development activities in laboratories and test facilities within their particular fields for the purposes of improving existing products and developing new ones to meet the needs of their customers. In addition, research and development programs are directed toward development of new products and services for diversification or expansion.
MEASUREX CORP /DE/	25/02/1994	The Company is actively engaged in basic technology and applied research and development programs which are designed to develop new or improved products and process applications.

ADVANCED TECHNOLOGY LABORATORIES INC	04/03/1994	The Company has obtained patents on certain of its products and has applied for patents which are presently pending.
CELL PATHWAYS INC	22/03/2002	The agreement also provides for future potential payments to Sinclair of up to $3 million depending on achievements related to sales, patent and clinical trial milestones.
AUTOBYTEL INC	22/03/2002	We have been issued a patent directed toward an innovative method and system for forming and submitting purchase requests over the Internet and other computer networks from consumers to suppliers of goods and services… We have applied for additional service marks and patents. We regard our trademarks, service marks, brand names and patent as important to our business.
BAUSCH & LOMB INC	22/03/2002	The company is currently involved in several pending patent proceedings relating to its <PureVision> contact lens product line.
IDACORP INC	22/03/2002	Currently, six-20 year US patents have been issued to IdaTech. More than 50 pending domestic and foreign patent applications addressing various aspects of fuel processor design, operation, materials, and integration with fuel cell stacks.
PAIN THERAPEUTICS INC	22/03/2002	We seek to protect our technology by, among other methods, filing and prosecuting U.S. and foreign patents and patent applications with respect to our technology and products and their uses. The issued patents are scheduled to expire no earlier than September 2012.
PHARMANETICS INC	22/03/2002	The Company pursues patent applications to provide protection from competitors. A number of U.S. and corresponding international patents have been issued to CVDI covering various aspects of the TAS technology. These patents expire between 2004 and 2013. The Company has filed, and is pursuing, a number of additional U.S. and international patent applications.

VALUECLICK INC	22/03/2002	We do not know if our current patent applications or any future patent application will result in a patent being issued within the scope of the claims we seek, if at all, or whether any patents we may receive will be challenged or invalidated. Although patents are only one component of the protection of intellectual property rights, if our patent applications are denied, it may result in increased competition and the development of products substantially similar to our own.
TELLABS INC	22/03/2002	Important factors that could cause our actual results to differ materially from those in forward-looking statements include, but are not limited to: economic changes impacting the telecommunications industry; new product acceptance; product demand and industry capacity; competitive products and pricing; manufacturing efficiencies; research and new product development; protection and access to intellectual property, patents and technology; ability to attract and retain highly qualified personnel; availability of components and critical manufacturing equipment; facility construction and start-ups; the regulatory and trade environment; availability and terms of business partnering arrangements and future acquisitions; uncertainties relating to synergies, charges, and expenses associated with business combinations and other transactions; and oth er risks and future factors that may be detailed from time to time in the Company's filings with the SEC.
TELLABS INC	22/03/2012	All of such patents, patent applications, registered trademarks, trademark applications and registrations and registered copyrights, if any, have been duly registered in, filed in or issued by the United States Patent and Trademark Office, the United States Register of Copyrights, or the corresponding offices of other jurisdictions as identified on Schedule 2.15(b), and have been properly maintained and renewed in accordance with all applicable provisions of law and administrative regulations in the United States and each such jurisdiction except as set forth on Schedule 2.15(b).
GEORGIA PACIFIC CORP	22/03/2002	The Corporation is the owner of numerous patents, copyrights, trademarks, licenses and trade secrets, as well as substantial know-how and technology (herein collectively referred to as technology, relating to its products and the processes for their production, the packages used for its products, the design and operation of various processes and equipment used in its business and certain quality assurance and financial software.

HON INDUSTRIES INC	22/03/2002	As of December 29, 2001, the Company owned 217 U.S. and 119 foreign patents and had applications pending for 58 U.S. and 73 foreign patents. In addition, the Company holds registrations for 136 U.S. and 184 foreign trademarks and has applications pending for 55 U.S. and 68 foreign trademarks.
HON INDUSTRIES INC	22/03/2009	The Company accomplishes this through improving existing products, extending product lines, applying ergonomic research, improving manufacturing processes, applying alternative materials and providing engineering support and training to its operating units.
MAUI LAND & PINEAPPLE CO INC	22/03/2002	In 1999, the Company was granted a U.S. patent on its fresh cut pineapple technology, which enhances the quality of the product while extending the shelf life.
MINERALS TECHNOLOGIES INC	22/03/2002	The Company believes that its rights under its existing patents, patent applications and trademarks are of value to its operations, but no one patent, application or trademark is material to the conduct of the Company's business as a whole.
ADC TELECOMMUNICATIONS INC	12/01/1994	The Company is committed to an ongoing program of new product development which combines internal development efforts with acquisition, joint venture, licensing or marketing
Delphi Automotive PLC	06/02/2017	We believe these markets are likely to experience substantial long term growth, and accordingly have made and expect to continue to make substantial investments, both directly and through participation in various partnerships and joint ventures, in numerous manufacturing operations, technical centers, research and development activities and other infrastructure to support anticipated growth in these areas.
UNITED TECHNOLOGIES CORP /DE	09/02/2017	The Introduction of New Products and Technologies Involves Risks and We May Not Realize the Degree or Timing of Benefits Initially Anticipated. We seek to achieve growth through the design, development, production, sale and support of innovative products that incorporate advanced technologies. The product, program and service needs of our customers change and evolve regularly, and we invest substantial amounts in research and development efforts to pursue advancements in a wide range of technologies, products and services.

NeuroMetrix	09/02/2017	We believe that we have research and development (R&D) capability that is unique to the industry with nearly two decades of experience in developing diagnostic and therapeutic devices involving the stimulation and measurement of nerve signals for clinical purposes.
ARKANOVA ENERGY CORP	10/02/2017	Management believes that future growth of our company will primarily occur through the exploration and development of our existing properties. However, we may elect to proceed through collaborative agreements and joint ventures in order to share expertise and reduce operating costs with other experts in the oil and gas industry.
ADC TELECOMMUNICATIONS INC	29/12/2021	Under the Original Agreement, we have engaged in exclusive research and development efforts with EMRE to evaluate and develop new and/or improved carbonate fuel cells to reduce carbon dioxide emissions from industrial and power sources in exchange for;(i) payment by EMRE of certain fees and costs as well as certain milestone-based payments to be paid only if certain technological milestones are met, two of which had not been satisfied as of the execution of Amendment No. 1, and (ii) certain licenses, in each case as described in the Original Agreement.
FUELCELL ENERGY INC	29/12/2021	Advanced Technologies contract revenues recognized under the EMRE Joint Development Agreement were approximately $1.3 million higher during the year ended October 31, 2021, reflecting continued performance under the EMRE Joint Development Agreement during the year ended October 31, 2021... FuelCell Energy has leveraged five decades of research and development to become a global leader in delivering environmentally responsible distributed baseload power platform solutions through our proprietary fuel cell technology.
MULLEN AUTOMOTIVE INC.	29/12/2021	We will strive to undertake significant testing and validation of our products in order to ensure that we meet the demands of our customers. We attempt to protect our intellectual property rights, both in the United States and abroad, through a combination of patent, trademark, copyright and trade secret laws, as well as nondisclosure and invention assignment agreements with our consultants and employees.
JANEL CORP	27/12/2021	Life Sciences faces an inherent business risk of exposure to product and other liability claims if its products, services or product candidates are alleged or found to have caused injury, damage or loss.

Energy Services of America CORP	29/12/2021	While we are constantly monitoring our health and safety programs, our industry involves a high degree of operating risk and there can be no assurance given that we will avoid significant liability exposure and/or be precluded from working for various customers due to high incident rates... We are a biopharmaceutical company focused on acquiring, developing and commercializing clinical-stage drugs for inflammatory and immune-related diseases with clear unmet medical needs. Our two lead product candidates, EB05 and EB01, are in later stage clinical studies.
Edesa Biotech, Inc.	28/12/2021	Forward-looking statements are based upon our current expectations, speak only as of the date hereof, are subject to change and include statements about, among other things: the status, progress and results of our clinical programs; our ability to obtain regulatory approvals for or successfully commercialize any of our product candidates....
BIO VASCULAR INC	18/06/1996	These programs include surgical trade shows, support of the presentation of clinical data and new product information by key physicians
CEB Inc	31/12/2016	We do this by combining our advanced **research** and analytics with best practices from thousands of member companies with our proprietary research methodologies, benchmarking assets, and human capital analytics… Through our proprietary research, we identify key economic leverage points and isolate high return-on-investment solutions for executives to implement. We offer multiple memberships that align with functional and key industry leadership roles. We deliver our research through various channels, including web-based resources, interactive workshops, live meetings, and published studies.
DATATRAK International, Inc	31/12/2008	DATATRAK International, Inc. is a technology and services company focused on global eClinical solutions, which assist companies in the clinical pharmaceutical, biotechnology, contract research organization (“CRO”) and medical device research industries in accelerating the completion of clinical trials... Development costs incurred in the research and development of new software products and enhancements to existing software products are expensed as incurred until technological feasibility has been established.

META Group INC	31/12/2003	META Group, Inc. and its subsidiaries (collectively the "Company") is a leading provider of information technology ("IT") research, advisory services, and strategic consulting. All costs incurred in the **development** of new **products** and services are expensed as incurred. General and administrative costs related to **developing** or obtaining such **software** are expensed as incurred.
SANGAMO THERAPEUTICS, INC	30/06/2017	Some statements contained in this report are forward-looking with respect to our operations, research, development and commercialization activities, clinical trials, operating results and financial condition. Sangamo and Pfizer may also collaborate in the research and development of additional adeno-associated virus ("AAV")-based gene therapy products for hemophilia A. As of July 26, 2017, we either owned outright or have exclusively licensed the commercial rights to approximately 812 patents issued in the United States and foreign national jurisdictions, and 617 patent applications pending worldwide. We continue to license and file new patent applications that strengthen our core and accessory patent portfolio.
SYMYX TECHNOLOGIES INC	31/12/2009	Symyx Software provides a suite of scientific software, content and technology, as well as associated professional services, to support R&D information lifecycle management across the enterprise, improving scientists' ability to search, develop, manage, manipulate and store research data and to manage intellectual property. We discover and patent a range of materials in our collaborations and internal research programs. These discovered materials provide us licensing opportunities, offer a path to commercialization of new materials, and demonstrate the capabilities of our high-throughput research technologies.
AVIGEN INC	31/12/2004	As of March 1, 2005, we owned, co-owned, or held licenses to 43 issued U.S. patents and 53 pending U.S. patent applications, as well as 32 issued non-U.S. patents and 70 pending non-U.S. patent applications. Some licenses require us to exercise our best efforts to achieve research, clinical, and commercial milestones

Appendix 3: Variable Definitions

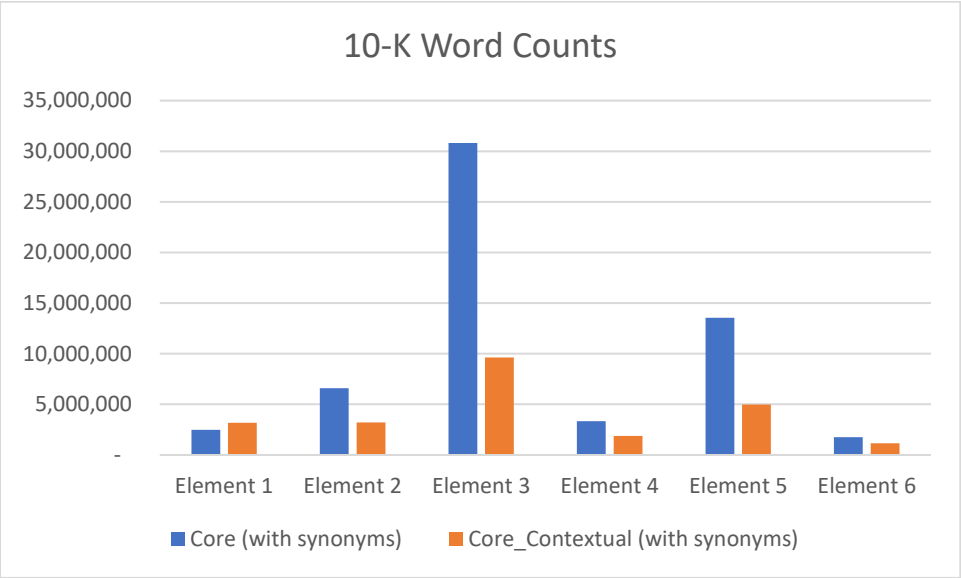
Note: This table includes variable definitions and descriptions for outcome and control variables used throughout the paper. The data source is Compustat and CRSP unless otherwise noted. As the main text includes a full discussion of the text-based r&d measure, the reader should refer to those sections for a description.

Variable	Name	Description
Textual	R&D narratives	*Firm's R&D narratives, computed as the number of words in the 10-K filings divided by the total number of words.*
Positive	Positive tone	*The percentage of words in 10-K with positive tone (following Loughran and McDonald (2011) dictionary)*
Negative	Negative tone	*The percentage of words in 10-K with negative tone (following Loughran and McDonald (2011) dictionary)*
Tangibility	Asset tangibility	*Property plant and equipment divided by total assets*
Cash/Assets	Cash to assets ratio	*The ratio of cash to assets taken from Compustat for year t*
Leverage	Leverage	*Total liabilities divided by assets, replacing book equity with market equity as of the last day of the fiscal year*
xrdintensity	R&D intensity	*R&D over Assets*
adintensity	Advertising intensity	*Advertising expenses over Assets*
Log (age)	Age	*The number of years since the first entered Compustat (earliest date 1975) expressed as a logarithm*
Log(assets)	Size	*The natural logarithm of total assets at fiscal year-end*
Log(Patents +1)	Patent count	*The number of patents issued in year t (following Kogan et al.(2017))*
Log(Cites +1)	Citation count	*Forward citations (following Kogan et al.(2017))*
Tobin's Q1		Market value of equity plus total assets minus common equity and balance sheet deferred taxes divided by total assets **(MKVALT + AT – CEQ-TXDB)/AT**
Tobin's Q2		The sum of market value of common equity (CSHO×PRCC_F from Compustat), liquidating value of preferred stock (PSTKL or PSTKRV if PSTKL is missing from Compustat) and book value of debt scaled by total assets (AT from Compustat) measured at the fiscal year end of year t. Book value of debt is computed as the difference between current liabilities (LCT from Compustat) and current assets (ACT from Compustat) plus inventory (INVT from Compustat) plus long-term debt (DLTT from Compustat). **((CSHO×PRCC_F) + PSTKL + (ACT -LCT +INVT + DLTT)) / AT**
Tobin's Q3		Book value of assets (at) minus book value of common equity (ceq) plus the market value of common equity (csho*prcc_f or mkvalt) divided by assets (at) **((PRCC_F * CSHO) + AT – CEQ) / AT**

Figure 1
Frequency and Tonal Classification of R&D Activities Words

This graph presents in Panel A the total 10-K frequencies for the occurrence of the six pillars that encompass R&D activities core words as well as core words that appear along with contextual words. Panel B highlights the percentage of R&D related words (core and contextual) classified into the different tonal classes identified in Loughran and McDonald (2011).

Panel A. Frequency of R&D words



Panel B. Tonal Classification of R&D words

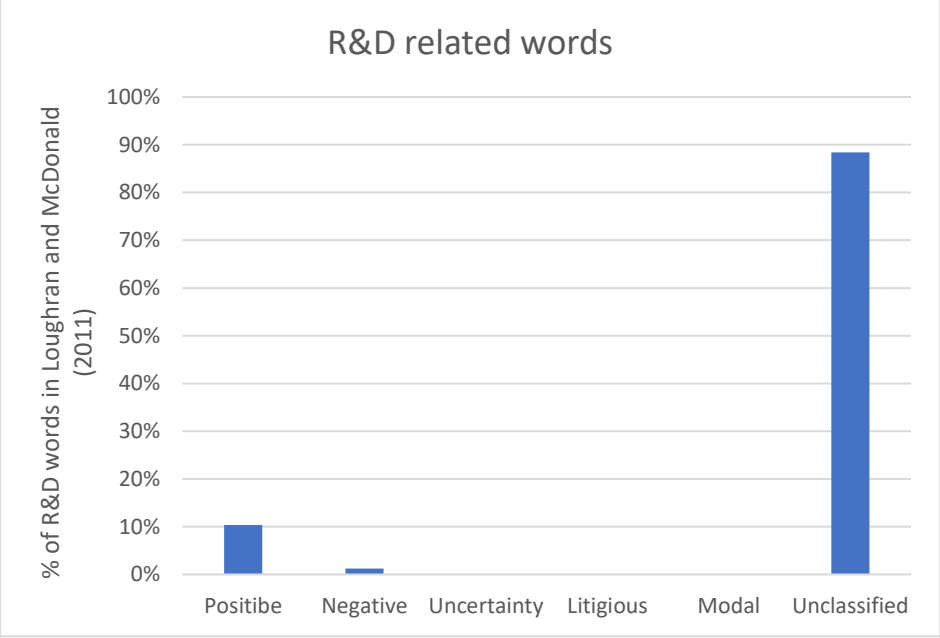
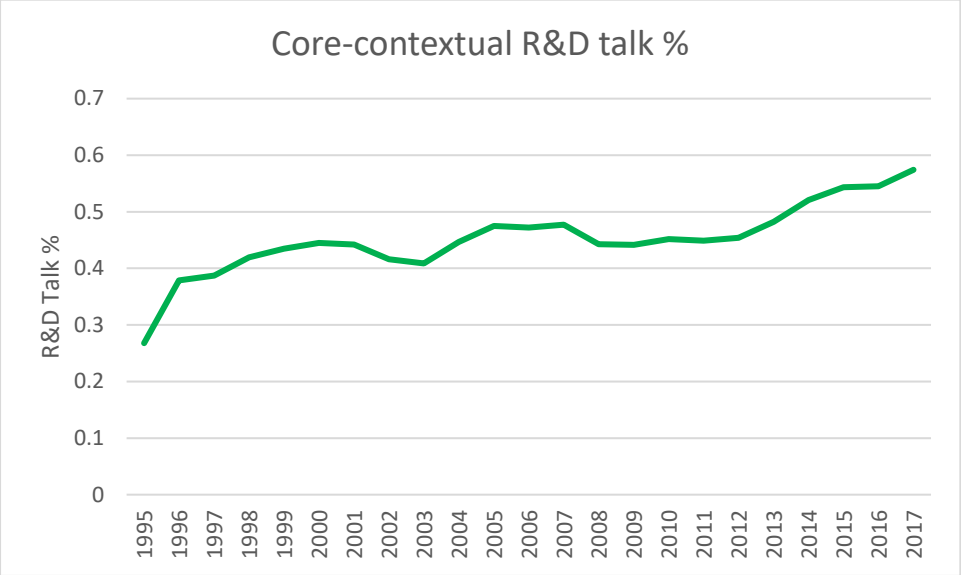


Figure 2
R&D talk over time

These figures plot the average frequency of R&D words relative to the total word count in 10K fillings over time. The frequency is calculated for the whole sample and is expressed in percentage terms. Panel A presents the frequency of R&D words using our core-contextual methodology and Panel B shows the frequency of R&D words of Merkley's (2011) Bag-of-words.

Panel A. Frequency of R&D words using core-contextual methodology



Panel B. Replication of Merkley's (2011) frequency of R&D words

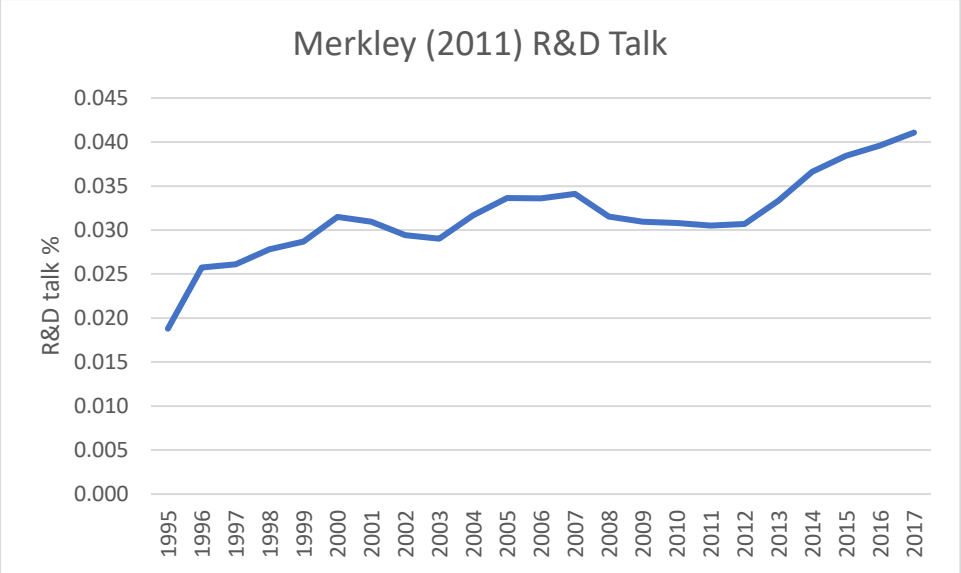
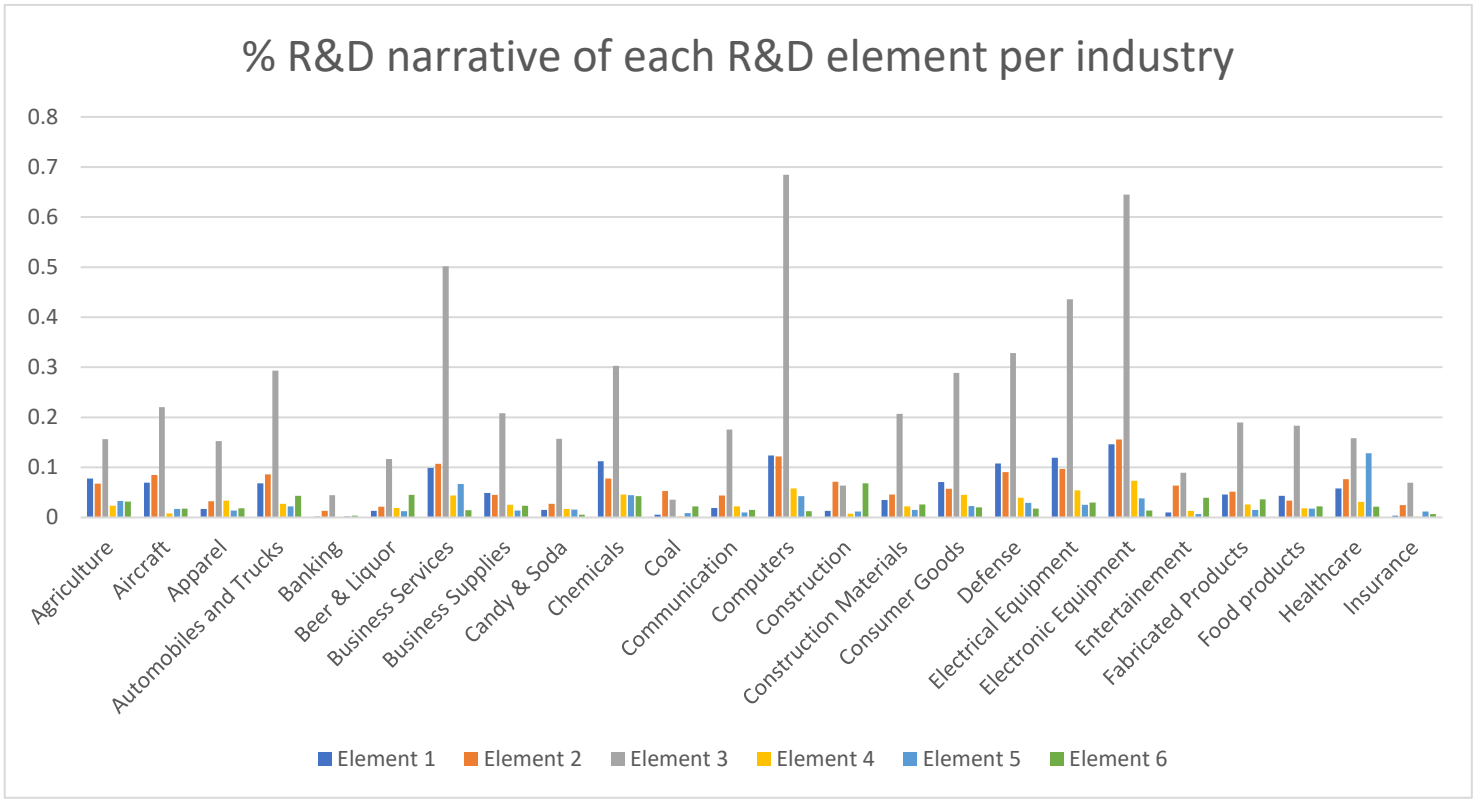


Figure 3
R&D word cloud

Figure 4
R&D narrative of each category per Fama and French 48 industry classification

Panel A: The first 24 Fama and French industry classification



Panel B: The remaining 24 Fama and French industry classification

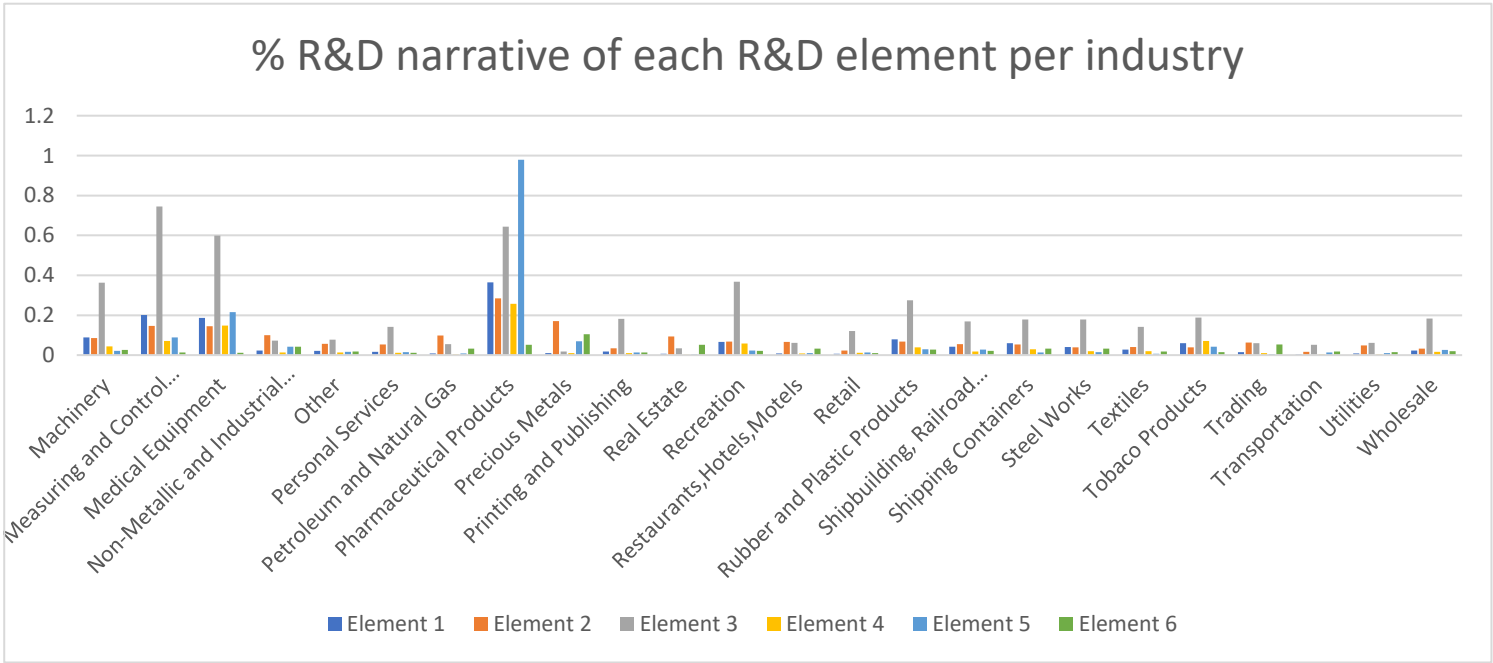


Figure 5

U.S State map with the most R&D active regions. State-level averages are calculated using the firm-year observations from 1995-2020

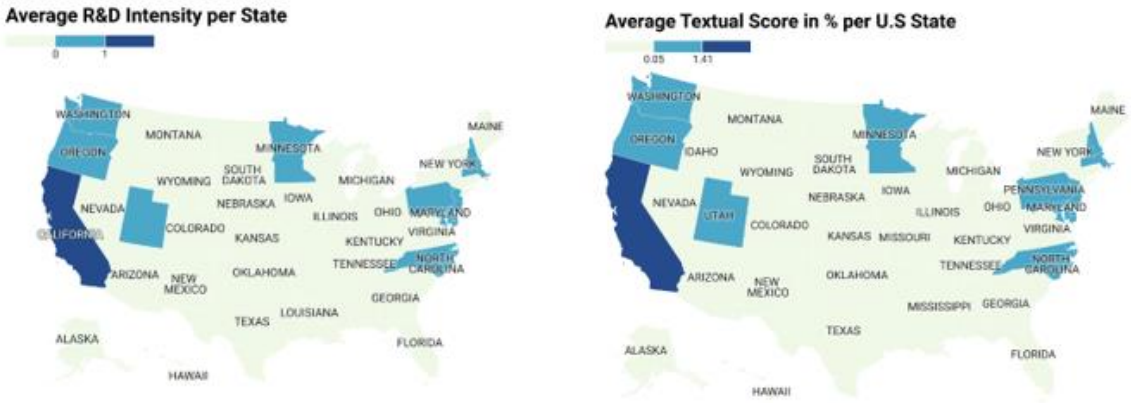


Figure 6

Average R&D talk per industry

This figure plots the distribution of R&D textual across industries in the form of box and whisker plot sorted by the mean R&D narrative disclosure. A box is drawn from the first quartile to the third quartile. A vertical line goes through the box at the median. The lines extending parallel from the boxes are known as the “whiskers”, which are used to indicate variability outside the upper and lower quartiles. R&D textual is defined as the frequency of R&D words over the total number of words in 10-K fillings, expressed as a percent for the period 1995-2020. The sample consists of 11,904 firms split into the 48 Fama and French industry categories.

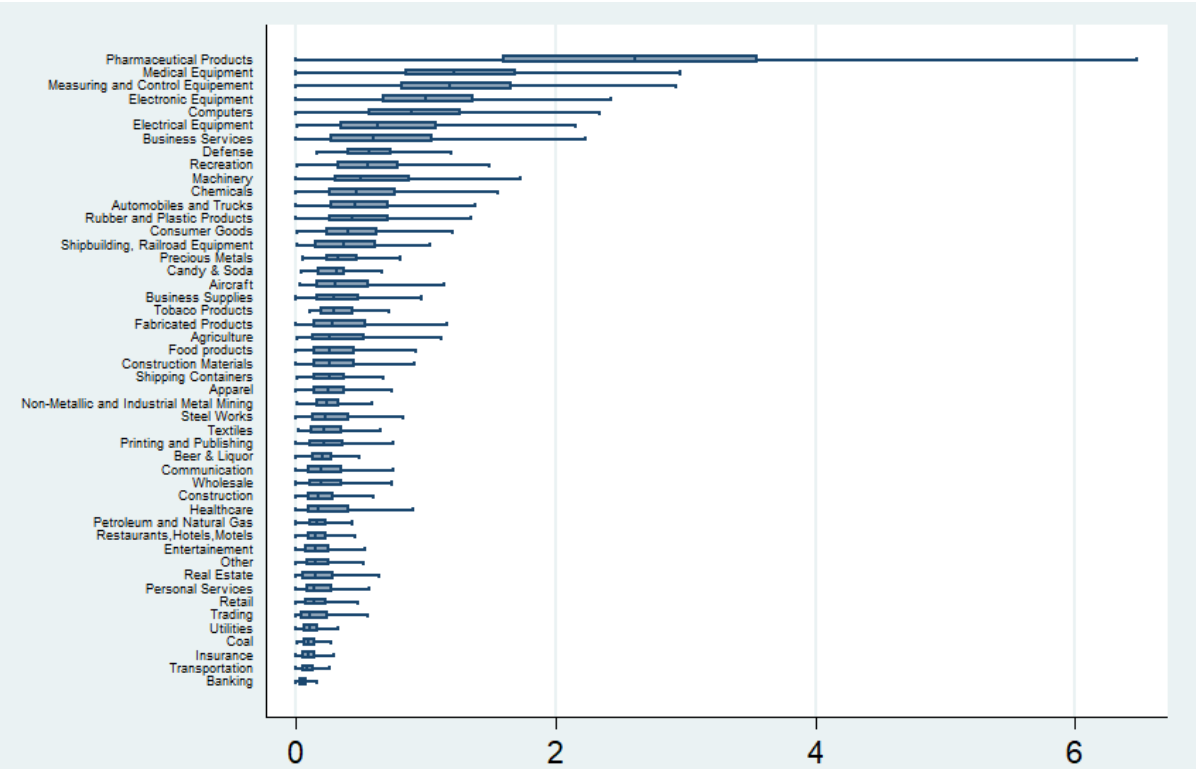


Table 1

This table reports the mean annual transition matrix between current and future period deciles of textual R&D. The diagonals are presented in bold figures.

	1	2	3	4	5	6	7	8	9	10
1	**68.76**	21.1	6.53	2.32	0.81	0.27	0.12	0.03	0.02	0.04
2	19.21	**47.95**	21.55	7.67	2.65	0.75	0.16	0.06	0	0
3	5.38	20.78	**41.49**	21.01	7.42	2.98	0.75	0.16	0.02	0.01
4	1.82	6.64	20.51	**40.37**	20.43	7.22	2.47	0.42	0.06	0.05
5	0.79	2.64	7.17	19.92	**42.36**	19.21	5.76	1.73	0.36	0.06
6	0.2	0.8	2.78	7.05	18.96	**43.77**	19.86	5.12	1.24	0.22
7	0.03	0.19	0.75	2.22	6.28	20.13	**45.75**	18.58	5.21	0.87
8	0.03	0.03	0.17	0.58	1.65	5.79	19.43	**49.82**	18.92	3.58
9	0	0.01	0.03	0.09	0.53	1.52	5.48	20.97	**57.98**	13.38
10	0.02	0.01	0.01	0.03	0.06	0.28	1.2	4.15	16.55	**77.68**

Table 2

Sample descriptive statistics for 1995-2020

Variable	Mean	Median	Std.Dev
Textuals (%)	0.61	0.26	0.83
Negative words (%)	0.02	0.02	0
Positive words (%)	0.01	0	0
Asset Total ($mil)	5,829.10	368.40	57,791.47
Common Equity ($mil)	1,141.49	121.53	6,793.65
Sales ($mil)	2,090.19	181.32	11,418.84
Market Value ($mil)	3,020.50	162.38	20,611.66
Tangibility	0.39	0.26	0.52
Cash/Assets	0.13	0.06	0.18
Leverage	0.56	0.54	0.67
Tobin Q	0.34	1.18	51.40
R&D intensity	0.06	0	0.20
Advertising intensity	0.01	0	0.05
Cites	141.26	0	1,595.97
Patent Number	10.17	0	116.97
Age	17	12	14.25

Table 3
Predictive validity with weighted citations per patent

This table presents output from ordinary least squares regressions that link our text-based R&D measure to patent counts and citation impact. In this table, the dependent variables we consider are logged patent counts over the following four years (t + 1 to t + 3), Log(1 + Patentst+1→t+3), and logged citation impact of patents over the following three years, Log(1 + (Citationst+1→t+3 /Patentst+1→t+3)). T-statistics are reported in parentheses. FE, fixed effects.
*p < 0.10; **p < 0.05; ***p < 0.01.

	log(1+Patents$_{t+1 \rightarrow t+3}$))		(log(1+(Citations $_{t+1 \rightarrow t+3}$ /Patents$_{t+1 \rightarrow t+3}$)))	
	(1)	(2)	(3)	(4)
textuals	0.121***	0.170***	0.220***	0.232***
	(11.891)	(13.002)	(15.096)	(17.976)
n_negative	0.026***	0.045***	0.039***	0.047***
	(5.923)	(6.391)	(6.330)	(7.526)
n_positive	-0.015***	0.022***	-0.013*	0.009
	(-2.945)	(2.881)	(-1.777)	(1.317)
tangibility	0.057***	0.033***	0.053***	0.033***
	(5.508)	(3.351)	(3.903)	(4.153)
cash_at	-0.057***	-0.064***	-0.080***	-0.090***
	(-9.130)	(-8.282)	(-9.475)	(-13.379)
leverage	0.045***	0.107***	0.030***	0.072***
	(6.123)	(12.293)	(2.979)	(8.035)
xrdintensity	0.007	0.038***	0.015	0.032***
	(1.152)	(4.768)	(1.605)	(4.814)
adintensity	0.055***	0.061***	0.021	0.025***
	(4.627)	(7.125)	(1.268)	(3.757)
logAge	0.239***	0.423***	0.134***	0.271***
	(13.959)	(28.589)	(5.830)	(30.680)
logAT	0.121***	0.170***	0.220***	0.232***
	(11.891)	(13.002)	(15.096)	(17.976)
software	0.012*	0.017**	-0.008	0.003
	(1.771)	(2.050)	(-1.001)	(0.584)

Year FE	YES	YES	YES	YES
Firm FE	YES	NO	YES	NO
Industry FE	NO	YES	NO	YES
Observations	106147	107373	106147	107373
Adjusted R-sq	0.719	0.342	0.555	0.277

Table 4
Regression with TobinQs

The variable definitions are tabulated in Appendix 5 and are analogous to those of Table 3. This table reports the link between narrative R&D disclosures and firm performance measured by different Tobin Q models. All specifications account for the full set of other controls, firm fixed effects and industry-year fixed effects. T-stats that are in parentheses.

	(1) $TobinQ1_{t+1}$	(2) $Tobinq2_{t+1}$	(3) $Tobinq3_{t+1}$	(4) $TobinQ1_{t+1}$	(5) $Tobinq2_{t+1}$	(6) $Tobinq3_{t+1}$
Textuals	0.064*	0.144***	0.133***	0.082***	0.149***	0.137***
	(1.863)	(3.798)	(3.503)	(2.833)	(4.586)	(4.202)
n_negative	-18.223***	-21.872***	-19.752***	-28.140***	-36.313***	-31.953***
	(-8.060)	(-9.129)	(-8.245)	(-9.281)	(-11.128)	(-9.571)
n_positive	-0.510	-9.075	-4.407	-22.159**	-33.554***	-26.956**
	(-0.077)	(-1.221)	(-0.603)	(-2.062)	(-2.861)	(-2.261)
tangibility	-0.131**	-0.274***	-0.194***	-0.168***	-0.243***	-0.257***
	(-2.158)	(-4.118)	(-2.914)	(-2.814)	(-3.869)	(-3.967)
Cash/at	0.901***	1.141***	1.134***	1.819***	2.149***	2.320***
	(10.074)	(11.867)	(11.819)	(12.165)	(12.857)	(13.803)
Leverage	0.591***	-0.011	0.500***	0.487***	-0.406***	0.301***
	(10.051)	(-0.182)	(8.085)	(6.550)	(-4.999)	(3.663)
R&D intensity	1.759***	1.749***	1.857***	3.003***	3.102***	3.380***
	(8.528)	(7.637)	(8.085)	(15.672)	(14.850)	(16.082)
Adv. intensity	0.395	0.130	0.537	3.523***	3.746***	4.141***
	(0.589)	(0.184)	(0.776)	(6.954)	(6.980)	(7.731)

logAge	-0.100***	-0.146***	-0.113***	-0.005	0.014	0.024
	(-3.956)	(-5.192)	(-3.998)	(-0.303)	(0.750)	(1.242)
logAT	-0.315***	-0.372***	-0.409***	-0.086***	-0.089***	-0.116***
	(-16.816)	(-17.806)	(-19.391)	(-6.051)	(-5.837)	(-7.316)
logPatents	-0.056**	-0.123***	-0.126***	0.085***	0.021	0.044**
	(-1.992)	(-3.981)	(-4.079)	(4.078)	(0.939)	(1.986)
logcites	0.013	0.025**	0.027**	0.049***	0.081***	0.079***
	(1.194)	(2.228)	(2.413)	(4.712)	(7.124)	(6.967)
Year FE	YES	YES	YES	YES	YES	YES
Firm FE	YES	YES	YES	NO	NO	NO
Industry FE	NO	NO	NO	NO	NO	NO
Observations	92233	92233	92233	93769	93769	93769
Adjusted R-sq	0.585	0.646	0.642	0.281	0.344	0.326

Table 5
Regression with interaction term for firms without citations

This table reports the link between narrative R&D disclosures and firm performance measured by different Tobin Q models. The variable definitions are tabulated in Appendix 5 and are analogous to those of Table 3 and Table 5. We include an interaction dummy term for firms without citations (=one if a firm has zero citations for the entire period). All specifications account for the full set of other controls, firm fixed effects and industry-year fixed effects. T-stats that are in parentheses.

	(1) **TobinQ1**$_{t+1}$	(2) **Tobinq2** $_{t+1}$	(3) **Tobinq3** $_{t+1}$	(4) **TobinQ1** $_{t+1}$	(5) **Tobinq2** $_{t+1}$	(6) **Tobinq3** $_{t+1}$
Textuals	0.087**	0.163***	0.157***	0.111***	0.201***	0.166***
	(2.295)	(3.991)	(3.781)	(3.271)	(5.361)	(4.424)
x nocitesdummy	-0.040	-0.024	-0.032	-0.041	-0.077**	-0.037
	(-1.334)	(-0.795)	(-1.027)	(-1.309)	(-2.282)	(-1.090)
n_negative	-18.189***	-21.699***	-19.576***	-27.132***	-35.103***	-30.635***
	(-8.015)	(-9.018)	(-8.134)	(-8.951)	(-10.776)	(-9.181)
n_positive	-0.321	-8.659	-3.989	-21.185**	-33.169***	-26.013**
	(-0.048)	(-1.162)	(-0.544)	(-1.964)	(-2.822)	(-2.174)
tangibility	-0.132**	-0.278***	-0.198***	-0.157***	-0.232***	-0.245***
	(-2.169)	(-4.162)	(-2.956)	(-2.633)	(-3.703)	(-3.777)
Cash/at	0.904***	1.142***	1.136***	1.844***	2.179***	2.342***
	(10.097)	(11.862)	(11.817)	(12.132)	(12.842)	(13.702)
leverage	0.587***	-0.019	0.491***	0.474***	-0.423***	0.283***
	(9.955)	(-0.311)	(7.914)	(6.318)	(-5.182)	(3.414)
R&D intensity	1.747***	1.725***	1.833***	3.099***	3.188***	3.483***
	(8.445)	(7.513)	(7.957)	(16.169)	(15.258)	(16.569)
Ad. intensity	0.404	0.142	0.551	3.678***	3.902***	4.304***
	(0.604)	(0.203)	(0.800)	(7.239)	(7.242)	(7.989)
logAge	-0.099***	-0.146***	-0.112***	0.002	0.019	0.030
	(-3.951)	(-5.159)	(-3.958)	(0.124)	(0.988)	(1.529)
logAT	-0.318***	-0.379***	-0.416***	-0.062***	-0.069***	-0.093***
	(-16.857)	(-17.920)	(-19.495)	(-4.751)	(-4.893)	(-6.312)
nocitesdummy	0.068**	0.091***	0.097***	-0.253***	-0.223***	-0.284***
	(2.522)	(3.205)	(3.376)	(-7.166)	(-5.937)	(-7.397)
Year FE	YES	YES	YES	YES	YES	YES
Firm FE	YES	YES	YES	NO	NO	NO
Industry FE	NO	NO	NO	NO	NO	NO

Observations	92233	92233	92233	93769	93769	93769
Adjusted R-sq	0.585	0.646	0.642	0.281	0.344	0.326

Table 6
Regression with a Non-R&D interaction term

This table reports the link between narrative R&D disclosures and firm performance measured by different Tobin Q models. The variable definitions are tabulated in Appendix 5 and are analogous to those of Table 3 and Table 5. We include an interaction dummy term for firms without R&D related expenses (=one if a firm has zero R&D expense for the entire period). All specifications account for the full set of other controls, firm fixed effects and industry-year fixed effects. T-stats that are in parentheses.

	(1)	(2)	(3)	(4)	(5)	(6)
	TobinQ1$_{t+1}$	**Tobinq2$_{t+1}$**	**Tobinq3$_{t+1}$**	**TobinQ1$_{t+1}$**	**Tobinq2$_{t+1}$**	**Tobinq3$_{t+1}$**
Textuals	0.106***	0.184***	0.178***	0.261***	0.337***	0.332***
	(2.911)	(4.639)	(4.453)	(8.331)	(9.864)	(9.643)
x NonRDexpense	-0.026	0.023	0.005	0.096	0.096	0.183**
	(-0.387)	(0.293)	(0.061)	(1.252)	(1.177)	(2.169)
Negative words	-17.866***	-21.337***	-19.199***	-22.331***	-30.135***	-25.091***
	(-7.805)	(-8.798)	(-7.897)	(-7.239)	(-9.092)	(-7.379)
Positive words	-0.979	-9.725	-4.976	-17.848*	-29.483**	-23.015*
	(-0.145)	(-1.283)	(-0.668)	(-1.667)	(-2.535)	(-1.944)
Tangibility	-0.068	-0.217***	-0.132*	-0.086	-0.159**	-0.164**
	(-1.101)	(-3.175)	(-1.940)	(-1.399)	(-2.472)	(-2.459)
Cash/AT	0.905***	1.146***	1.139***	2.124***	2.460***	2.658***
	(10.036)	(11.846)	(11.785)	(14.476)	(15.078)	(16.180)
Leverage	0.684***	0.076	0.593***	0.681***	-0.209***	0.517***
	(11.401)	(1.192)	(9.374)	(9.106)	(-2.588)	(6.288)
Ad.Intensity	0.441	0.175	0.590	3.389***	3.601***	3.953***
	(0.665)	(0.251)	(0.863)	(6.511)	(6.502)	(7.097)
logAge	-0.091***	-0.139***	-0.105***	0.005	0.022	0.033
	(-3.570)	(-4.876)	(-3.664)	(0.294)	(1.108)	(1.623)
logAT	-0.360***	-0.421***	-0.460***	-0.073***	-0.081***	-0.105***
	(-19.306)	(-20.019)	(-21.776)	(-5.846)	(-6.009)	(-7.515)
NonRDexpense	-0.066	-0.105*	-0.080	-0.457***	-0.456***	-0.542***
	(-1.173)	(-1.712)	(-1.329)	(-10.445)	(-9.721)	(-11.278)
Year FE	YES	YES	YES	YES	YES	YES
Firm FE	YES	YES	YES	NO	NO	NO
Industry FE	NO	NO	NO	NO	NO	NO

Observations	92233	92233	92233	93769	93769	93769
Adjusted R-sq	0.582	0.644	0.639	0.259	0.326	0.304