I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 13029

The Value of Publicly Available, Textual and Non-textual Information for Startup Performance Prediction

Ulrich Kaiser
Johan M. Kuhn

MARCH 2020

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 13029

The Value of Publicly Available, Textual and Non-textual Information for Startup Performance Prediction

Ulrich Kaiser *School*
University of Zurich, Copenhagen Business School, ZEW Leibniz Center for European Economic Research and IZA

Johan M. Kuhn
EPAC and Copenhagen Business

MARCH 2020

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

The Value of Publicly Available, Textual and Non-textual Information for Startup Performance Prediction*

Can publicly available, web-scraped data be used to identify promising business startups at an early stage? To answer this question, we use such textual and non-textual information about the names of Danish firms and their addresses as well as their business purpose statements (BPSs) supplemented by core accounting information along with founder and initial startup characteristics to forecast the performance of newly started enterprises over a five years' time horizon. The performance outcomes we consider are involuntary exit, above–average employment growth, a return on assets of above 20 percent, new patent applications and participation in an innovation subsidy program. Our first key finding is that our models predict startup performance with either high or very high accuracy with the exception of high returns on assets where predictive power remains poor. Our second key finding is that the data requirements for predicting performance outcomes with such accuracy are low. To forecast the two innovation-related performance outcomes well, we only need to include a set of variables derived from the BPS texts while an accurate prediction of startup survival and high employment growth needs the combination of (i) information derived from the names of the startups, (ii) data on elementary founder-related characteristics and (iii) either variables describing the initial characteristics of the startup (to predict startup survival) or business purpose statement information (to predict high employment growth). These sets of variables are easily obtainable since the underlying information is mandatory to report upon business registration. The substantial accuracy of our predictions for survival, employment growth, new patents and participation in innovation subsidy programs indicates ample scope for algorithmic scoring models as an additional pillar of funding and innovation support decisions.

JEL Classification: L26, C53

Keywords: startup, performance, prediction, text as data

Corresponding author:
Ulrich Kaiser
University of Zürich
Department of Business Administration
Affolternstrasse 56
8050 Zürich
Switzerland
E-mail: ulrich.kaiser@business.uzh.ch

* We thank Henrik Barslund Fosse of the Novo Nordisk Foundation for directing our attention to the existence of the Danish business purpose statement data and gratefully acknowledge financial support from the social science research program "The Socioeconomic Impact of Research in Denmark" of the Novo Nordisk Foundation. This research beneted from helpful comments received at the first University of Zurich Winter Workshop held in Saas Fee January 8-10, 2020, and the 2nd Digital Economy workshop held in Tel Aviv March 1-2, 2020. We gratefully acknowledge useful feedback from Jörg Claussen, Martin Murmann, Christian Peukert and, in particular, Reinhold Kesler.

INTRODUCTION

Identifying promising startups is a formidable task for investors, creditors and policy makers alike. Even though each group often has quite a wealth of information available when deciding about a possible involvement in a particular startup, this information must be processed quickly which in turn implies that simple heuristics become highly valuable (Baum and Wally 2001; Eisenhardt 1989; Kirsch et al. 2009). In addition, investors and creditors aim at identifying promisinging startups early and therefore increasingly often use algorithmic scoring models (Corea 2018; Diffey 2019; Palmer 2017).

That publicly available information can be used to effectively measure entrepreneurial success is demonstrated in seminal work by Guzman and Stern (2015; G/S hereafter). G/S use information on firm and founder names, geographical location as well as an indicator for a startup holding a patent at the time of foundation to measure entrepreneurial success defined as either an IPO or an acquisition at the ZIP-code level. In this paper, we study how well the initial G/S variables in combination with other publicly available and similarly conveniently obtainable data can predict a broad range of performance outcomes: involuntary exit, high employment growth, a return on assets of above 20 percent, new patent applications and, as a more inclusive indicator of innovative activity, participation in an innovation subsidy program.

Our set of performance predictors comprises of (i) the initial G/S firm name variables, (ii) an extended set of variables derived from firm names, (iii) basic founder characteristics such as gender and previous founding experience as well as business success, (iv) initial startup characteristics like industry affiliation, initial assets and profits as well as address information and (v) variables generated from firms' business purpose statements (BPSs). BPSs are required by most US states and most European countries as an integral part of the business formation documents. They are, e.g., mandatory for corporations worldwide where they are also referred to as "articles of organization", "articles of incorporation" or "certificate of incorporation".

We base our analysis on the population of Danish firms started as incorporated companies between 2012 and 2014, 55914 firms in total, whose data we web-scrape from government websites. To assess the changes in forecasting accuracy that our extended lists of potential predictors cause, we employ simple logit models for our five performance models and calculate the respective areas under the receiver-operator curve (AUC) as our main measure of prediction accuracy. The AUC is a frequently applied forecast performance statistic for binary firm performance models (Agarwal and Taffler 2008, Åstebro and Winter 2012; Chava and Jarrow 2004). We assess the contribution of each set of explanatory variables since not all data may

be publicly available in all countries and since there are differences in their ease of use.

Our key findings are that (i) our models predict all performance outcomes with high accuracy with the exception of high return on assets, (ii) the data needed to generate our precise forecasts are both easily obtainable and straightforward to apply in simple empirical models and (iii) prediction accuracy can be substantially improved by including variables beyond the ones initially suggested by G/S.

Predicting our two innovation-related performance indicators with high accuracy only requires the set of variables we derive from the BPSs. Combining the BPS variables with the initial G/S variables even leads to predictions of very high accuracy for new patent applications. Accurately predicting involuntary exit and high employment growth is more data demanding as both involve the combination of three different sets of variables. A satisfactory prediction of involuntary exit and high employment growth needs the basic G/S variables in combination with the set of founder characteristics. On top of these two sets of variables, predicting involuntary exit involves the set of initial startup characteristics while predicting high employment growth entails the additional inclusion of the BPS-derived set of variables. Importantly, the basic G/S variables, the founder characteristics and the data derived from the BPSs are likely to be easily accessible since they are mandatory to report to the authorities upon business registration. We hence not only demonstrate that it is possible to accurately predict startup success, we also show that the data required to generate such accurate predictions may in fact be readily available from public sources. This is of particular interest given a global trend towards the opening of business register data to the public. Initiatives like the "Open Government Partnership" with its explicit goal to ease the access to public data are getting more and more traction with now including 79 countries worldwide. Data sets similar to ours hence are or will soon be available in many other countries (`https://www.opengovpartnership.org/`).

Our paper unfolds as follows: we first present our data, then introduce our empirical methods, subsequently discuss our empirical results and finally conclude.

DATA

Our core data is generated and collected by the Danish Business Authority (DBA), an administrative unit under the authority of the Danish Ministry of Business. We track all firms started between 2012 and 2014 over a period of five years. The data comprises of the universe of 55914 firms registered as limited liability companies (LLCs), joint stock corporations or a new form of a LLC called "iværksætterselskab" (IVS) whose main difference to a standard LLC is that

it does not come with capital requirements and hence in effect without liabilities on part of the owners. The DBA data also provide us with the company names and addresses, NACE Rev. 2 industry codes, starting dates, total assets, profits, the number of employees as well as the names and person identifiers of their founders. In addition, the DBA data contain the BPSs since firms are obliged to report their business purpose as part of their general charters. Business purpose statements are mandatory by the Danish Law of Corporated Firms which provides firms with substantial leeway in their eventual formulation as there is no wordcount limit and the BPSs only need to loosely describe a startups' activity. As a consequence, many BPSs are very generic ("The purpose of this firm is to do trading.") while others are very specific.[1] We shall make use of this heterogeneity in our empirical analysis.

Dependent variables

We consider five alternative performance variables: (i) involuntary exit, (ii) high employment growth, (iii) a return on assets of above 20 percent, (iv) at least one patent after foundation and (v) participation in an innovation subsidy program; variables that, except for the last one, are commonly used in management and economics. New business survival is very widely studied (Audretsch and Mahmood 1995; Cassar 2014; Chava and Jarrow 2004; Gimmon and Levie 2010). Visitin and Pittino (2014) as well as Wennberg et al. (2011) consider employment growth as a main performance outcome. Return on asset is considered by Morgan et al. (2009) as well as Cornett and Tehranian (1992) while patents are standard indicators for innovative activity (Blundell et al. 1995; Griliches 1990; Kaiser et al. 2015, 2019). However, not all inventions are patented and not all inventions can be patented (Arundel and Kabla 1998). We therefore consider participation in an innovation subsidy program as an additional and broader indicator of innovative activity. All Danish innovation subsidy programs are competitive and reviewed, which in turn implies that the program sponsors assessed that the applicant firm exceeds the quality threshold for the respective subsidization program.

We measure all performance variables within the first five years after establishment, except for return on assets which we measure within the first three years after foundation due to a substantial increase in missing information over a five year time horizon — many firms that started in 2014 have not yet submitted in their fifth year financial report. We define involuntary exits as closures due to bankruptcy and compulsory dissolution enforced by the regulatory authorities due to non-compliance to administrative requirements. It does not include dissolution

[1] E.g., "The company's purpose is to design and develop, manufacture and assemble switchboards, steering and control boards, PLC/PC/SRO solutions, automation and pre-finished projects for use by fitters, OEM/system manufacturers and the industry in general at a quality and at a price that entails that customers, suppliers and other stakeholders regard the company as an attractive and professional partner."

after a merger or an acquisition which would count as business success (Bates 2005; Detienne and Wennberg 2014; G/S) or voluntary exits. Employment figures are provided in categories of 0, 1, 2-4, 5-9, 10-199 and more than 199 employees. We term startups that increase employment by at least two categories as "high employment growth" businesses since each category implies a doubling of the number of employees. Our final two performance measures refer to innovative performance: patents and participation in an innovation subsidy program. Our patent application data originates in the "PatStat" database provided by the European Patent Office to which researchers at Copenhagen Business School have attached the unique Danish identifiers which allow us to combine our data sets (Kaiser et al. 2015; 2019). It includes all patents filed at the European Patent Office or the World Intellectual Property Organization that involve at least one Danish applicant or inventor. We have data on the universe of Danish innovation support schemes collected by Danish Ministry of Higher Education and Science at our disposal.[2]

Explanatory variables

We relate our five performance variables to our five sets of explanatory variables, e.g. the basic G/S variables, the extended G/S firm name variables, founder characteristics, startup characteristics and BPS information, as well as combinations thereof.

(i) The G/S variables: Our first set of explanatory variables follows Guzman and Stern (2015). Their model to predict startup performance contains dummy variables for (i) the firm name being eponymous (i.e. it reflects one of the founder's names), (ii) the firm name being short or long, (iii) the geographical location appearing in the firm name (specified as a dummy for any geographical location like a city, village or region appearing in the firm name and another dummy for the terms "Denmark", "Danish" or "Dan" in the firm name) (iv) the legal form (dummies for corporations and IVSs with LLCs as base category), (v) the geographic regions the firm is residing in and (vi) the startup commanding over at least one patent at the time of foundation.

(ii) The extended G/S variables: We extend this basic set of variables derived from the firm names by dummy variables for the firm name containing (i) a "proper" word which we define based on the dictionary of Danish words as a proxy for the firm name containing information on what the firm actually does (like "baking", "consulting" or "plumbing"), (ii) the terms "holding", "capital", "invest" or "share" in the firm name to identify holding companies as well as (iii) a female name and (iv) a male name. We in addition include a (v) founder name index since social psychology and economics suggest that person names constitute strong indicators

[2]This data was made available to us via the project "Investments, Incentives and the Impact of Danish Research" sponsored by the Novo Nordisk Foundation.

of a persons' background (Fryer et al. 2004; Gerhards and Hans 2009; Goldstein and Stecklov 2016; Mehrabian 1997). To account for potential information contained in founder names, we build a "name index" by calculating the name-specific average performance of firms started by founders with a focal given name. We e.g. find that 15 percent of the founders named "Ulrich" face a forced exit within the first five years while this is the case for ten percent of the founders with the given name "Johan". For solo founders named Ulrich this generates an index of 0.15, for founders named Johan it is 0.1. For team foundations we take the averages across the set of founder names.

(iii) The human capital variables: As a third set of variables we employ information on the startups' founders at the time of business foundation. These include dummy variables for (i) at least one founder being a legal entity, (ii) at least one founder having a female first name, (iii) at least one founder having a male first name, (iv) the startup being founded by a team (e.g. more than one person founder), (v) the five number of employees categories described above with this information being missing as the comparison group, (vi) one of the founders having previously founded between one and three other firms and (vii) one of the founders having previously founded between more than three other firms and (viii) one of the founders having previously experienced an involuntary exit.

Firm size at startup has been shown to be highly correlated with post-entry performance (Arora and Nandkumar 2011; Bonardo et al. 2011; Brüderl et al. 1992; Clarysse et al. 2011; Delmar and Shane 2004; Ensley and Hmieleski 2005; Visintin and Pittini 2014; Zahra et al. 2007) and the same is true for gender (Delmar and Shane 2004; Davidsson and Honig 2003; Wennberg et al. 2011). We control for team foundations because they are said to have an edge over solo founders since teams pool human and financial resources instead of being dependent on the solo entrepreneur (Eesley et al. 2014; Eisenhardt and Schoonhoven 1990), a view recently challenged by Greenberg and Mollick (2018). The importance of previous founder experience is widely demonstrated (e.g. Baron and Ensley 2006; Dencker and Gruber 2015; Eesley and Roberts 2006a,b; Gompers et al. 2006; Westhead et al. 2005) which motivates our inclusion of the previous founding experience dummies. We additionally control for previous involuntary exit (Cope 2010; Hayward et al. 2010; Nielsen and Sarasvathy 2016; Wagner 2002).

(iv) The firm characteristics: Our fourth set of explanatory variables concerns itself with the characteristics of the startup. Financial information has long been used as a predictor for business performance (Altman 1968, 1984; Brüderl et al. 1992; Dambolena and Khoury 1980; Huyghebaert et al. 2000; Laitinen 1992). We account for total assets and total profits in the first year. Both variables are missing for half of our observations, a "sparsity of data" problem that is very common in big datasets. Following Gelman and Hill (2007), we set the missing

corresponding explanatory variables to zero and in order to distinguish genuine 0s from the artificially created 0s introduce an additional dummy for such replacements having taken place. Since information on total assets is missing in all cases where information on profits is missing as well, we only need to include a single indicator for such a replacement having taken place. We operationalize total profits by using quantiles dummies while we take the natural logarithm of profits.

Our data contains detailed address information and we use this text as data to create indicators for the business history of each address and for the address being shared with other firms. Specifically, we include a dummy variable for at least one involuntary exit at the respective address as well as another dummy variable for nine or more involuntary exits at the address. These two dummy variables may serve as proxies for the overall attractiveness of the location and other characteristics associated with the given address. We control for how many other firms reside under the same address since many corporations often co-reside with their associated holding companies by including dummies for the address being shared by 2-5, 6-10 and more than 10 other firms with the address being unshared being the base category. In addition, we account for the present address having previously been used by 1-5, 6-10, 11-100 and more than 100 firms. To account for differences across sectors (Brüderl et al. 1992; Clarysse et al. 2011), we include a set of NACE Rev. 2 one digit sector dummy variables. A missing sector classification constitutes our base category. To more precisely account for sectoral heterogeneity without being forced to include a large set of sectoral dummy variables, we include mean industry performance for all our five performance indicators which we calculate on the basis of the Danish Industry Classification that is slightly more detailed than NACE Rev. 2 four digit level.[3]

(v) The BPS data: Our fifth set of explanatory variables uses the BPS data. Before turning the BSP text data into explanatory variables we remove words and phrases which do not contain information relevant to our analysis, an approach called "stopping" in computer linguistics. Examples for stopwords are "the", "because", "between" or "against". In addition, we "stemmed" all words in the BPSs. Stemming reduces words to their roots, e.g. the words "automation" and "automated" would both be reduced to their root "autom". We use the dictionary of the Danish Language Authority as our source for stemming. After stopping and stemming we define three subsets of BPS-related variables that either relate to BPS complexity, to its specificity or to its very content. As measures of BPS complexity we consider (i) the "LIX" due to Björnsson (1968) which has found widespread application in text analysis.

[3]For example, if 30 percent of *other* firms in the regression sample in a focal firm's industry experience high employment growth, the associated mean performance for in this industry is 0.3.

It is calculated as the sum of the percentage of words of more than six letters and the average number of words per BPS in our context. The higher the LIX, the higher is the complexity of the text. We in addition use complexity-related variables measuring (ii) mean word length, (iii) BPS length and (iv) dummy variables for the quintiles of the BPS length distribution to put BPS length into perspective. To measure BPS specificity we use counts of how many times a "proper" word in a focal BPS appears in the universe of BPSs. We operationalize these counts as (v) the frequency with which the least common word in a focal BPS appears in the universe of BPSs and (vi) the frequency with which the most common word in a focal BPS appears in the universe of BPSs. We also control for the ratio of these two variables. Similar to our treatment of our firm name information we finally create the following content-related variables: (vii) a dummy for a geographic term appearing in the BPS, (viii) a dummy variable for a male name appearing in the BPS and (ix) a dummy variable for a female name appearing in the BPS. As a final subset of the BPS variables we generate (x) "wordscore indices" that measure the mean "performance" of firms' BPSs for each of our five performance indicators. The wordscore approach has been developed in political sciences where it has found widespread application in inferring political positions in text documents on the basis of scores for words derived from documents (Laver et al. 2003). It is perhaps best illustrated by providing an example. A share of 47.4 percent of the startups with the word "discotheque" in their BPSs face an involuntary exit while this is true for 36.4 percent of the startups with the word "delivery" in their BPSs. The wordscore associated with the word "discotheque" is defined as the word's average "performance" and hence is 0.474 while the other wordscore is 0.364. To aggregate the individual wordscores at the firm name level, we take the average of the individual wordscores.

Table 1 presents descriptive statistics of our dependent and explanatory variables. It shows that involuntary exits are comparatively rare events with 17.2 percent of the firms in our data involuntarily exiting within five years of operation, a figure that is substantially lower than the 50 percent overall exits reported by e.g. Headd (2003) or Mata and Portugal (1994). Different to those studies we focus, however, on involuntary exits as well as firms with a legal form that requires registration and consider the universe of startups instead of merely technology-driven ones. A tenth of the firms in our data generate substantial employment growth while 17 percent achieve a high return on assets. By contrast, participation in an innovation subsidy program and taking out a new patent are both rare events. A mere 1.6 percent of our firms participate in innovation subsidy programs while only 0.3 percent apply for a new patent within their first five years of existence.

More than 40 percent of all startups are founded in the capital greater Copenhagen region, only 0.4 percent of all firms has applied for a patent at the time of foundation, about a third of

the startups involve another firm as a founder, 87 percent of the startups are founded by men, more than 89 percent are solo foundations and 46 percent are founded by serial entrepreneurs which compares to a European average of 30 percent and a US average of 13 percent (Plehn-Dujowich 2010). Turning to the information contained in the BPSs, the average LIX is 54 which is considered as "difficult" by Björnsson (1968). The mean word length is at 9.4 characters while average BSP lengths is 41.3 characters.

The correlations between our explanatory variables are modest with our largest variance inflation factor being 2.56 which is well below the critical value of 10 (Belsley et al. 1980).

EMPIRICAL ANALYSIS

Empirical strategy

Our empirical aim is twofold: we want to analyze (i) the degree of accuracy to which publicly available data can be used to forecast business startup performance and (ii) what sets of variables — and combinations thereof — are best at predicting performance since not all variables may be equally easy to get a handle on. We seek to achieve our goals by subsequently introducing our five different sets of explanatory variables as well as their combinations in logit performance regressions and by subsequently assessing the out-of-sample prediction accuracy of our specifications. We estimate our models on a 70 percent random sample and retain the remaining 30 percent for prediction, following G/S. We calculate our firm name indices and our BPS wordscores as well as the average industry performance index on the regression sample and extrapolate them to our holdout sample.

Our focus is on the prediction of outcomes and we therefore present the forecasting accuracy statistics only and relegate logit coefficient estimation results for our full models to Appendix A. We apply three different prediction accuracy measures: (i) the AUC, (ii) the log-likelihodd value and (iii) the Bayesian Information Criterion (BIC). Our focus is on the AUC as a standard measure of forecast performance of binary models (Hand 2001). It illustrates the performance of a classification model like ours by plotting the observed rate of outcomes against the rate of false positive outcomes at pre-specified threshold levels (the receiver-operator curve, ROC), deciles in our case as in Cooper et al. (1993). The area under thes curve is a measure of predictive accuracy where an AUC of 0.5 suggests no predictive power at all while a value of 1 corresponds to perfect prediction. Bradley (1997) defines a model that corresponds to an AUC of between 0.5 and 0.6 as a "fail', values between 0.6 and 0.7 as "poor", between 0.7 and 0.8 as "fair", between 0.8 and 0.9 as "good" and values above 0.9 as "excellent". In

addition, we calculate the percentage changes in the AUC compared to the specification that uses the Guzman/Stern set of variables only. Almost all our models include the basic G/S set of variables which allows us to compare the log-likelihood values of the basic G/S model to the richer models as suggested by standard textbooks (Greene 2017; Wooldridge 2016) as a second prediction accuracy statistic. Our results table displays the percentage change in the log-likelihood statistics compared to the G/S benchmark model which is equal to the relative change in the associated likelihood-ratio test statistics. These test statistics cannot reject that all models that include variables beyond the basic G/S ones are jointly statistically highly significant; i.e., the fuller models have statistically significantly larger explanatory power than the base G/S specification. This is why we do not provide the p-values in our results table. Our third alternative useful textbook statistic to study differences between both non-nested and nested models is the Bayesian Information Criterion (BIC), a statistic frequently used for model selection (Kass and Raftery 1995). Adding additional parameters may lead to overfitting, a problem which the BIC attempts to solve by penalizing extra parameters added to the empirical model. The preferred model is the one with the lowest BIC. Our results table displays the changes in the BIC relative to the basic G/S model along with a categorization of these percentage changes into "not worth more than a mention" (abbreviated in the table by "none"), "positive", "strong" and "very strong" which correspond to changes between 0 and 2, 2 to 6, 6 to 10 and above 10 respectively (Jeffreys 1935; Kass and Raftery 1995).

Results

Table 2 presents our prediction outcomes. A first striking finding is that the information contained in our BPS-related variables is rich enough to *alone* predict the two innovation–related outcomes with "good" accuracy. An even "excellent" accuracy is achieved once the BPS data is combined with both the human capital variables and the basic G/S variables. An "excellent" predictive performance is not obtained for any other performance outcome. A second striking result is that all our specifications poorly predict a high return on assets. Even though combining the initial G/S variables with the firm characteristics and the BPS information leads to a massive improvement in predictive accuracy by 22.2 percent as measured by the AUC, it still remains"poor" with an AUC of 0.686.

Involuntary exit is predicted with "good" accuracy with an AUC of 0.801 when the basic G/S variables are combined with the set of human capital variables and the BPS data. Predictive power can be increased by 2.7 percent if the BPS variables are added as well, leading to an AUC of 0.823. Adding even more variable sets does, however, not increase predictive power. Similarly, it also needs the combination of at least three sets of variable, the basic G/S variables, the human capital and the BPS variables, to attain "good" predictive accuracy for high employment

growth. Adding variables sets beyond these three actually decreases AUC since AUC penalizes the number of explanatory variables.

Turning to the changes in the log-likelihood function as alternative prediction accuracy measures, we naturally find that the more sets of variables we include, the larger the log-likelihood values become because logit models maximize the log-likelihood functions without penalizing the number of explanatory variables. These improvements are largest for adding the set of startup characteristics and the set of BPS-related variables to the initial G/S specification, which is a finding that reinforces our initial AUC-based results. Likewise, the changes in the BIC echo these initial results as well since the addition of the set of initial startup characteristics and the BPS-related variables lead to "strong" or "very strong" reductions in the BIC.

To sum up, our models predict startup survival, high employment growth and participation in an innovation subsidy program well. They predict new patents very well but fail to predict high returns on assets with acceptable accuracy. Our results show that it is sufficient to include the BPS-related variables to generate a "good" predictive accuracy for new patents and participation in an innovation subsidy program. To get the "excellent" predictive accuracy for new patents the BPS variables need to be combined with the basic G/S variables and the founder characteristics. An accurate prediction of involuntary exit and high employment growth requires the combination of the three sets of variables where both predictions need the basic G/S variables and the human capital characteristics. Predicting involuntary exit in addition involves the inclusion of the startup characteristics while the high employment growth forecast additionally entails the BPS-generated variables. We hence find that the initial G/S variables, the human capital variables and the BPS-related variables are key contributors to startup success prediction. At the same time, these variables are particularly easy to obtain since they are primarily based on textual information on the names of the startups and their founders and therefore data that is mandatory to report upon business registration.

Robustness checks

Even though all data we use in our analysis is publicly available, not all variables may be easily obtainable in all countries. In addition, not all variable are equally simple to process. The initial G/S variables include information on whether or not a startup has applied for a patent at the time of incorporation while the set of human capital variables includes initial firm size. Even though information on previous patenting activity is easily gathered via online searches for individual firms, it is very cumbersome to match startups to their corresponding patenting history on a broader scale. Likewise, publicly available information on startups often does not contain information on initial firm size which limits the direct applicability of our startup characteristics variables. Finally, the wordscores constitute important elements of the set of

BPS-related variables. Again, while individual BPSs indeed are easily obtainable, processing the universe of BPSs may be more demanding. In our robustness checks we therefore test the extend to which leaving out the information on inital patents, initial firm size and wordscores affects prediction accuracy.

Omitting initial firm size reduces the predictive accuracy for involuntary exit and participation in an innovation subsidy program by 0.28 and 0.03 percentage points, respectively. Not surprisingly given that there is likely to be state dependence in firm size (Audretsch et al. 1999), it is more relevant for high employment growth where the average reduction is 4.65 percentage points. It also matters for applying for at least one new patent where the average reduction is 1.92 percentage points. We nevertheless obtain an "excellent" prediction accuracy for new patents and a "good" prediction accuracy for high employment growth.

Omitting initial patents from the specifications leaves the prediction accuracy for involuntary exit and high employment growth essentially unchanged. The predictive power for new patents only drops by 0.89 percentage points, despite state dependence in patenting activity being well documented (Blundell et al. 1995; Kaiser et al. 2015, 2018). It does matter, however, for predicting participation in an innovation subsidy program where the decrease is 4.83 percentage points on average and where we no longer achieve "good" predictive accuracy with a maximum AUC of 0.799, or 0.13 percentage points short of "good" category.

Leaving out the three variables that are either likely to be harder to gather or to process has hence very little effect on prediction accuracy overall.

CONCLUSIONS

Easily accessible and publicly available data, both textual and non-textual, are starting to become easily accessible in most modern economies. We show how such data can be used to predict the expected performance of newly started enterprises with substantial accuracy. Such performance predictions are of great importance to investors, creditors and policy makers alike. Investors may not only want to assess the prospects of a business that asks for funding, they may also be interested in identifying promising startups before they even apply. Some investors have already embraced "algorithmic scoring" models (Corea 2018; Diffey 2019; Palmer 2017) and our paper indicates that it is well possible to successfully use such methods. Even though banks are unlikely to be equally proactive, they may as well want to more firmly base their debt financing decisions on objective data-driven grounds. Finally, policy makers may gain from the improved identification of promising startups in order to be better gear innovation support

programs towards such firms and to improve the tailoring of startup promotion programs more generally.

For our predictions, we use data on the universe of Danish firms started between 2012 and 2014 to run simple logit regressions to show that key performance outcomes such as survival, employment growth, patenting activity as well as participation in competitive and audited innovation support programs can be predicted with high accuracy using publicly available data alone. Our models essentially only require "text as data" information that startups have to report when they register: startup names, founder identities, addresses and business purpose statements. Even though including hard-to-get or hard-to-process additional information on initial firm size, initial patents and an index of the relatedness of words used in the buiness purpose statements to aggregate startup performance improves prediction accuracy, such more intricate data is not necessary to forecast startup performance with substantial precision. However, even our most complex model was unable to predict returns on asset of above 20 percent with even modest accuracy.

Our finding that we are — apart for our outcome variable high return on assets — are able to forecast startup performance with substantial accuracy using publicly available data alone suggests that there are ample opportunities for the early identification of promising startups. The fact that we use simple logit models, a standard workhorse in the anaylsis of binary outcomes, makes our approach applicable to a wide range of users.

The data we use have recently become publicly available through the open data policy adopted by the Danish government in 2017 as part of the "Open Government Partnership" initiative that started in 2011 and now includes 79 countries worldwide. This paper shows that such open data policies indeed are effective in improving economic decision making at very low cost.

Table 1: **Descriptive statistics**

	Dummy	Mean	Std.dev.		Dummy	Mean	Std.dev.
Dependent variables				*Startup characteristics*			
Involuntary exit	1	0.172		Total assets year 1	0	12280	186841
High employment growth	1	0.095		Total assets year 1 is missing	1	0.503	
High return on assets	1	0.170		1st quintile profits year 1	1	0.099	
Innovation subsidy program	1	0.016		2nd quintile profits year 1	1	0.102	
New patent	1	0.003		3rd quintile profits year 1	1	0.092	
Explanatory variables				4th quintile profits year 1	1	0.102	
Guzman/Stern firm name variables				5th quintile profits year 1	1	0.102	
Eponymous firm name	1	0.146		Previous exit at same address	1	0.265	
Short firm name	1	0.190		>9 previous exits at same address	1	0.047	
Medium long firm name	1	0.646		Adress unshared	1	0.151	
Long firm name	1	0.164		Adress shared with 2-5 other firms	1	0.500	
Firm name: w/ geogr. location	1	0.056		Adress shared with 6-10 other firms	1	0.134	
Firm name w/ Danmark, Danish, Dan	1	0.021		Adress shared with >10 other firms	1	0.214	
Legal form: corporation	1	0.050		Address previously unused	1	0.132	
Legal form: IVS	1	0.099		Address previously used by 1-5 others	1	0.423	
Legal form: LLC	1	0.851		Address previously used by 6-10 others	1	0.138	
Geogr. region: Midtjylland	1	0.217		Address previously used by 11-100 others	1	0.230	
Geogr. region: Nordjylland	1	0.081		Address previously used by > 100 others	1	0.077	
Geogr. region: Sjælland	1	0.111		Sector 1	1	0.032	
Geogr. region: Syddanmark	1	0.165		Sector 2	1	0.145	
Geogr. region: Greater Copenhagen	1	0.425		Sector 3	1	0.039	
Patents at foundation	1	0.004		Sector 4	1	0.161	
Extended Guzman/Stern firm name variables				Sector 5	1	0.062	
In firm name: a proper danish word	1	0.475		Sector 6	1	0.029	
In firm name: Holding, capital, shares	1	0.282		Sector 7	1	0.012	
In firm name: female name	1	0.081		Sector is missing	1	0.520	
In firm name: male name	1	0.158		Mean ind. perf. invol. exit	0	0.173	0.084
Founder name index invol. exit	0	0.171	0.070	Mean ind. perf. high empl. growth	0	0.079	0.057
Founder name high empl. growth	0	0.082	0.035	Mean ind. perf. high ret. on assets	0	0.142	0.057
Founder name high ret. on assets	0	0.135	0.038	Mean ind. perf. new patents	0	0.016	0.038
Founder name new patents	0	0.018	0.012	Mean ind. perf. innov. subsidy program	0	0.005	0.024
Founder name innov. subsidy program	0	0.006	0.006	*BPS information*			
Human capital variables				LIX	0	54.0	9.6
At least one founders is firm	1	0.369		Mean word length	0	9.4	2.2
At least one of founder has female first name	1	0.180		BPS lengths	0	41.3	33.0
At least one of founder has male first name	1	0.868		BPS 1st quintile	1	0.192	
Team	1	0.111		BPS 2nd quintile	1	0.195	
# employees year 1: 0	1	0.042		BPS 3rd quintile	1	0.201	
# employees year 1: 1	1	0.068		BPS 4th quintile	1	0.203	
# employees year 1: (2,4)	1	0.069		BPS 5th quintile	1	0.208	
# employees year 1: (5,9)	1	0.030		Frequency of least common word	0	1247	2160
# employees year 1: (6,49)	1	0.025		Frequency of most common word	0	5673	3818
# employees year 1: missing	1	0.767		Freq. least/freq. most common word	0	0.291	0.370
No previous founding experience	1	0.198		A geogr. term name is in BPS	1	0.020	
Previously founded 1-3 firms	1	0.463		A male name is in BPS	1	0.022	
Previously founded more than 3 firms	1	0.212		A female name is in BPS	1	0.011	
Earlier invol. exit by one founder	1	0.084		BPS wordscore invol. exit	0	0.163	0.071
				BPS wordscore high empl. growth	0	0.080	0.045
				BPS wordscore high ret. on assets	0	0.144	0.045
				BPS wordscore new patents	0	0.019	0.020
				BPS wordscore innov. subsidy program	0	0.006	0.009

Standard deviations are displayed for continuous variables only.

Table 2: **Prediction accuracy**

Sets of variables included											
G/S basic (i)	G/S ext. (ii)	Hum. cap. (iii)	Firm char. (iv)	BPS (v)	ROC-AUC Value	ROC-AUC Cat.	ROC-AUC Δ	Log likeli-hood Δ	BIC Δ	BIC Cat.	dof
Involuntary exit											
x					0.623	poor	0.0	0.0	0.0	none	14
	x				0.636	poor	2.0	—	-3.0	positive	5
		x			0.699	poor	12.1	—	-4.5	positive	12
			x		0.719	fair	15.3	—	-7.7	strong	23
				x	0.720	fair	15.5	—	35.2	very strong	14
x	x				0.683	poor	9.6	5.4	-5.2	positive	19
x		x			0.740	fair	18.7	8.3	-7.8	strong	26
x			x		0.745	fair	19.5	10.4	-9.6	strong	37
x				x	0.743	fair	19.2	8.2	-7.8	strong	28
x	x	x			0.763	fair	22.5	12.4	-11.8	very strong	31
x	x		x		0.755	fair	21.2	13.4	-12.5	very strong	42
x	x			x	0.751	fair	20.5	11.0	-10.3	very strong	33
x		x	x		0.801	good	28.5	17.7	-16.5	very strong	49
x		x		x	0.796	fair	27.7	15.4	-14.5	very strong	40
x			x	x	0.780	fair	25.2	14.3	-13.1	very strong	51
x		x	x	x	0.823	good	32.0	21.0	-19.4	very strong	63
x	x		x	x	0.784	fair	25.8	16.4	-15.0	very strong	56
x	x	x		x	0.801	good	28.5	17.5	-16.4	very strong	45
x	x	x		x	0.825	good	32.4	21.4	-20.1	very strong	54
x	x	x	x	x	0.824	good	32.2	22.5	-20.7	very strong	68
High employment growth											
x					0.591	fail	0.0	0.0	0.0	none	14
	x				0.668	poor	13.0	—	-8.0	strong	5
		x			0.673	poor	13.8	—	-4.5	positive	12
			x		0.679	poor	14.8	—	-6.5	strong	23
				x	0.790	fair	33.6	—	25.2	very strong	14
x	x				0.684	poor	15.8	8.1	-7.7	strong	19
x		x			0.694	poor	17.5	5.6	-5.0	positive	26
x			x		0.699	poor	18.2	8.1	-6.9	strong	37
x				x	0.794	fair	34.3	12.5	-11.7	very strong	28
x	x	x			0.750	fair	26.9	13.0	-12.0	very strong	31
x	x		x		0.729	fair	23.3	13.4	-11.8	very strong	42
x	x			x	0.787	fair	33.2	14.7	-13.6	very strong	33
x		x	x		0.743	fair	25.6	12.8	-10.9	very strong	49
x		x		x	0.829	good	40.2	17.9	-16.4	very strong	40
x			x	x	0.790	fair	33.6	16.2	-14.1	very strong	51
x		x	x	x	0.824	good	39.4	21.2	-18.5	very strong	63
x	x		x	x	0.785	fair	32.9	18.0	-15.7	very strong	56
x	x	x		x	0.825	good	39.6	20.4	-18.6	very strong	45
x	x	x		x	0.828	good	40.1	21.3	-19.1	very strong	54
x	x	x	x	x	0.822	good	39.0	23.4	-20.5	very strong	68
Return on assets of above 20 percent											
x					0.561	fail	0.0	0.0	0.0	none	14
	x				0.588	fail	4.8	—	-2.1	positive	5
		x			0.596	fail	6.3	—	-1.5	none	12
			x		0.629	poor	12.0	—	-4.2	positive	23
				x	0.660	poor	17.5	—	37.4	very strong	14
x	x				0.596	fail	6.2	2.1	-1.9	none	19
x		x			0.608	poor	8.3	1.8	-1.4	none	26
x			x		0.637	poor	13.4	5.0	-4.2	positive	37
x				x	0.661	poor	17.8	3.9	-3.4	positive	28
x	x	x			0.620	poor	10.5	3.7	-3.1	positive	31
x	x		x		0.640	poor	13.9	6.4	-5.3	positive	42
x	x			x	0.652	poor	16.2	4.9	-4.2	positive	33
x		x	x		0.654	poor	16.5	6.2	-4.9	positive	49
x		x		x	0.674	poor	20.1	5.5	-4.5	positive	40
x			x	x	0.677	poor	20.6	7.2	-5.9	positive	51
x		x	x	x	0.686	poor	22.2	8.3	-6.5	strong	63
x	x		x	x	0.670	poor	19.4	8.1	-6.5	strong	56
x	x	x		x	0.666	poor	18.6	6.4	-5.3	positive	45
x	x	x		x	0.681	poor	21.4	7.8	-6.3	strong	54
x	x	x	x	x	0.679	poor	20.9	9.2	-7.3	strong	68
At least one new patent											
x					0.721	fair	0.0	0.0	0.0	none	14
	x				0.747	fair	3.7	—	-3.3	positive	5
		x			0.733	fair	1.7	—	-0.3	none	12
			x		0.752	fair	4.3	—	-6.9	strong	23
				x	0.885	good	22.7	—	14.7	very strong	14
x	x				0.795	fair	10.3	7.0	-6.0	positive	19
x		x			0.787	fair	9.2	5.5	-3.3	positive	26
x			x		0.794	fair	10.2	12.1	-7.9	strong	37
x				x	0.894	good	24.1	21.6	-18.7	very strong	28
x	x	x			0.814	good	13.0	9.8	-6.7	strong	31
x	x		x		0.831	good	15.3	16.9	-11.8	very strong	42
x	x			x	0.898	good	24.6	24.9	-21.1	very strong	33
x		x	x		0.825	good	14.4	15.4	-9.1	strong	49
x		x		x	0.903	excellent	25.3	25.0	-20.0	very strong	40
x			x	x	0.899	good	24.7	25.8	-18.9	very strong	51
x		x	x	x	0.903	excellent	25.3	28.4	-19.4	very strong	63
x	x		x	x	0.901	excellent	24.9	28.8	-20.9	very strong	56
x	x	x		x	0.906	excellent	25.6	27.5	-21.6	very strong	45
x	x	x		x	0.911	excellent	26.4	27.4	-19.9	very strong	54
x	x	x	x	x	0.904	excellent	25.4	30.7	-20.8	very strong	68

Sets of variables included					ROC-AUC			Log likeli-hood	BIC		
G/S basic (i)	G/S ext. (ii)	Hum. cap. (iii)	Firm char. (iv)	BPS (v)	Val.	Cat.	Δ	Δ	Δ	Cat.	dof
Participation in an innovation subsidy program											
x					0.745	fair	0.0	0.0	0.0	none	14
	x				0.655	poor	-12.1	—	-12.8	very strong	5
		x			0.635	poor	-14.8	—	-19.7	very strong	12
			x		0.745	fair	0.0	—	-15.8	very strong	22
				x	0.883	good	18.5	—	-49.2	very strong	14
x	x				0.758	fair	1.8	6.5	1.3	none	19
x		x			0.763	fair	2.5	5.4	-5.7	positive	26
x			x		0.769	fair	3.3	8.1	-12.1	very strong	36
x				x	0.872	good	17.1	11.2	-0.6	none	26
x	x	x			0.774	fair	3.9	10.5	-5.6	positive	31
x	x		x		0.775	fair	4.0	13.3	-11.8	very strong	41
x	x			x	0.862	good	15.7	15.7	-1.1	none	31
x		x	x		0.784	fair	5.2	12.3	-18.8	very strong	48
x		x		x	0.863	good	15.8	16.1	-6.8	strong	38
x			x	x	0.835	good	12.1	17.3	-14.3	very strong	48
x		x	x	x	0.836	good	12.2	21.2	-21.3	very strong	60
x	x		x	x	0.831	good	11.5	21.2	-15.3	very strong	53
x	x	x		x	0.856	good	14.9	19.6	-8.0	strong	43
x	x	x		x	0.836	good	12.2	19.3	-17.0	very strong	53
x	x	x	x	x	0.834	good	12.0	24.3	-22.9	very strong	65

The five sets of explanatory variables are (i) the basic G/S variables, (ii) the extended set of G/S variables, (iii) the human capital variables, (iv) the firm characteristics and (v) the BPS variables. "Val.' refers to the value of the respective test statistic while Δ refers to its percentage change relative to the basic G/S model. Changes in the log-likelihood statistic cannot be calculated for the models not including the G/S variables. "Dof" denotes the degrees of freedom of the respective estimation model.

References

Åstebro, T. and Winter, J.K. (2012). More than a dummy: The probability of failure, survival and acquisition of firms in financial distress, European Management Review 9(1), 1-17.

Agrawal, V. and Taffler, R. (2008). Comparing the performance of market-based and accounting-based bankruptcy prediction models. Journal of Banking & Finance 32: 1541–1551.

Altman, E.I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. Journal of Finance 23: 589–609.

Altman, E.I. (1968). The success of business failure prediction models: an international survey. Journal of Banking & Finance 8: 171-184.

Arora, A. and Nandkumar, A. (2011). Cash-Out or flameout! Opportunity cost and entrepreneurial strategy: theory, and evidence from the information security industry. Management Science, 57(10): 1844-1860.

Arundel, A. and Kabla, I. (1998). What percentage of innovations are patented? Empirical estimates for European firms. Research Policy, 27: 127-141.

Audretsch, D.B. and Mahmood, T. (1995). New firm survival: New results using a hazard function. The Review of Economics and Statistics, 77(1): 97-103.

Audretsch, D., Santarelli, E. and Vivarelli, M. (1999). Start-up size and industrial dynamics: some evidence from Italian manufacturing. International Journal of Industrial Organization 17: 965-983.

Baron , R.A. and Ensley, M.D. (2006). Opportunity recognition as the detection of meaningful patterns: evidence from comparisons of novice and experienced entrepreneurs. Management Science, 52(9): 1331-1344.

Bates, T. (2005). Analysis of young, small firms that have closed: delineating successful from unsuccessful closures. Journal of Business Venturing, 20(3): 343-358.

Baum, J.R. and Wally, S. (2003). Strategic decision speed and firm performance. Strategic Management Journal 24(11): 1107–129.

Belsley D., Kuh, E. and Welsch, R. (1980). Regression diagnostics: identifying influential data and sources of collinearity. Wiley & Sons.

Laver, M., Benoit, K. and Garry, J. (2003). Extracting policy positions from political texts using words as data. The American Political Science Review, 97(2): 311-331.

Björnsson, C. H. (1968). Läsbarhet. Stockholm: Liber.

Blundell, R., Griffith, R. and van Reenen, J. (1995). Dynamic count data models of technological innovation. Economic Journal, 105: 333-344.

Bonardo, D., Paleari, S. and Vismara, S. (2011). Valuing university-based firms: the effects of academic affiliation on IPO performance, Entrepreneurship Theory & Practice, 35(4): 755-776.

Bradley, A.P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern recognition, 30(7): 1145-1159.

Brüderl, J., Preisendörfer, P. and Ziegler, R. (1992). Survival chances of newly founded business organizations. American Sociological Review, 4(1): 227-242.

Cassar, G. (2014). Industry and startup experience on entrepreneur forecast performance in new firms. Journal of Business Venturing 29: 137-151.

Chava, S. and Jarrow, R.A. (2004). Bankruptcy prediction with industry effects. Review of Finance 8: 537-569.

Clarysse, B., Tartari, V. and Salter, A. (2011b). The impact of entrepreneurial capacity, experience and organizational support on academic entrepreneurship. Research Policy, 40(8): 1084–1093.

Cooper, A. (1993). Challenges in predicting new firm performance. Journal of Business Venturing 8: 241-253.

Cope, J. (2010). Entrepreneurial learning from failure: an interpretive phenomenological analysis. Journal of Business Venturing.

Corea, F. (2018). *An Introduction to Data*. Springer.

Cornett, M.M. and Tehranian, H. (1992). Changes in corporate performance associated with bank acquisitions. Journal of Financial Economics, 31(2): 211-234.

Dambolena, I.G. and Khoury, S.J. (1980). Ratio stability and corporate failure. The Journal of Finance, 35(4): 1017-1026.

Davidsson, P. and Honig, B. (2003). The role of social and human capital among nascent entrepreneurs. Journal of Business Venturing, 18(3): 301-331.

Delmar, F. and Shane, S. (2004). Legitimating first: organizing activities and the survival of new ventures. Journal of Business Venturing, 19: 385-410

Detienne, D.R. and Wennberg, K. (2014). What do we really mean when we talk about "exit"? — a critical review of research on entrepreneurial exit. International Small Business Journal, 32(1): 4-16.

Dencker, J.C. and Gruber, M. (2015). The effects of opportunities and founder experience on new firm performance. Strategic Management Journal 36: 1035-1052.

Diffey, C. (2019). Motherbrain: How AI is helping this VC firm to pick the next big startup. TechRound, May 6, 2019; `https://techround.co.uk/news/motherbrain-using-ai-to-pick-start-up/`

Eesley, C.E., Hsu, D.H. and Roberts, E.B. (2014). The contingent effects of top management teams on venture performance: aligning founding team composition with innovation strategy and commercialization environment. Strategic Management Journal, 35(12): 1798-1817.

Eisenhardt, K.M. (1989). Making fast strategic decisions in high-velocity environments. Academy of Management Journal 32(3): 543–576.

Eisenhardt, K.M. and Schoonhoven, C.B. (1990). Organizational Growth: Linking Founding Team, Strategy, Environment, and GrowthAmong U.S. Semiconductor Ventures, 1978-1988. Administratice Science Quarterly, 35(3): 504-529.

Ensley, M.D. and Hmieleski, K.M. (2005). A comparative study of new venture top management team composition, dynamics and performance between university-based and independent startups. Research Policy, 34: 1091–1105.

Fryer, R.G., Levitt, S.D. (2004). The causes and consequences of distinctively black names. Quarterly Journal of Economics 119(3): 767–805.

Gelman, A. and Hill, J. (2007). Data analysis using regression and multilevel/hierarchical models. Cambridge University Press.

George, G., Zahra, S.A. and Wood, D.R. (2002). The effects of business-university alliances on innovative output and financial performance: a study of publicly traded biotechnology companies. Journal of Business Venturing, 17 (6): 577–609.

Gerhards, J. and Hans, S. (2009). From Hasan to Herbert: Name-Giving Patterns of Immigrant Parents between Acculturation and Ethnic Maintenance. American Journal of Sociology, 114: 4, 1102-1128.

Gimmon, E. and Levie, J. (2010). Founder's human capital, external investment, and the survival of new high-technology ventures. Research Policy 39: 1214-1226.

Goldstein, J.R. and Stecklov, G. (2016). From Patrick to John F.: ethnic names and occupational success in the last era of mass migration. American Sociological Review, 81(1): 85–106.

Gompers, P., Kovner, A. R., Lerner, J. and Scharfstein, D. (2006). Skill vs. luck in entrepreneurship and venture capital: Evidence from serial entrepreneurs. Journal of Financial Economics, 96:18–32.

Greenberg, J. and Mollick, E.R. (2018). Sole survivors: solo ventures versus founding teams. NYU working paper http://dx.doi.org/10.2139/ssrn.3107898.

Greene, W. (2017). Econometric analysis, Pearson.

Griliches, Z. (1990). Patent statistics as economic indicators: a survey. Journal of Economic Literature, 28(4): 1661-1990.

Guzman, J. and Stern, S. (2015). Where is Silicon Valley? Science, 347(6222): 606-609.

Hand, D. J. (2001). Measuring diagnostic accuracy of statistical prediction model. Statistica Neerlandica, 55: 3-16.

Hayward, M., Forster, W. and Fredrickson, B. (2010). Beyond hubris: how highly confident entrepreneurs rebound to venture again. Journal of Business Venturing, 25(6): 569-578.

Headd, B. (2003). Redefining business success: Distinguishing between closure and failure. Small Business Economics, 21(1): 51–61.

Huyghebaert, N., Gaeremynck, A., Roodhooft, F. and van de Gucht, L.M. (2000). New firm survival: the effect of start-up characteristics, Journal of Business Finance & Accounting 27(5): 627-651.

Jeffreys, H. (1935). Some tests of significance, treated by the theory of probability. Proceedings of the Cambridge Philosophy Society 31, 203-222.

Kaiser, U., Kongsted, H.C. and Rønde, T. (2015). Does the mobility of R&D labor increase innovation? Journal of Economic Behavior and Organization, 110, 91-105.

Kaiser, U., Kongsted, H.C., Laursen, K. and Ejsing, A.-K. (2018). Experience matters: the role of academic scientist mobility for industrial innovation. Strategic Management Journal, 39(7): 1935-1958.

Kass, R.E. and Raftery, A.E. (1995). Bayes Factors. Journal of the American Statistical Association, 90(430): 773–795.

Kirsch, D., Goldfarb, B. and Gera, A. (2009). Form or substance: the role of business plans in venture capital decision making. Strategic Management Journal 30: 487–515.

Laitinen, E. (1992). Prediction of failure of a newly founded firm. Journal of Business Venturing 7: 323-340.

Mata, J. and Portugal, P. (1994). Life duration of new firms. The Journal of Industrial Economics, 42(3): 227–245.

Mehrabian, A. (1997). Impressions created by given names. Names, 45(1) : 19-33.

Morgan, N.A., Vorhies, D.W. and Mason, C.H. (2009). Market orientation, marketing capabilities, and firm performance. Strategic Management Journal 30(8), 909-920.

Nielsen, K. and Sarasvathy, S.D. (2016). A market for lemons in serial entrepreneurship? Exploring type I and type II errors in the restart decision. Academy of Management Discoveries 2(3): 247-271.

Palmer, M. (2017). Artificial intelligence is guiding venture capital to startups. Financial Times: Dec. 11, 2017; https://www.ft.com/content/dd7fa798-bfcd-11e7-823b-ed31693349d3.

Plehn-Dujowich, J. (2010). A theory of serial entrepreneurship. Small Business Economics 35(4): 377-398.

Visintin, F. and Pittino, D. (2014). Founding team composition and early performance of university-based spin-off companies. Technovation, 34: 31-43.

Wagner, J. (2002). Taking a second chance: entrepreneurial restarters in Germany. The Institute for the Study of Labor (IZA) Discussion Paper Series.

Wennberg, K., Wiklund, J. and Wright, M. (2011). The effectiveness of university knowledge spillovers: performance differences between university spinoffs and corporate spinoffs. Research Policy, 40: 1128– 1143.

Westhead, P., Ucbasaran, D., Wright, M. and Binks, M. (2005). Novice, serial and portfolio entrepreneur behaviour and contributions. Small Business Economics 25: 109-132.

Wooldridge, J.M. (2010). Econometric analysis of cross section and panel data. MIT press.

Zahra, S.A., van de Velde, E. and Larraneta, B. (2007). Knowledge conversion capability and the performance of corporate and university spin-offs. Industrial and Corporate Change, 16(4): 569-608.

Appendix A: Logit estimates

	Involuntary exit exit	High empl. growth	High ret. On assets	Inno. subsidy program	New patent
Eponymous firm name	-0.222***	-0.098	0.077*	-0.219	0.736
	(0.053)	(0.089)	(0.045)	(0.267)	(0.593)
Short firm name	0.005	0.05	0.048	0.151	0.368
	(0.041)	(0.054)	(0.043)	(0.112)	(0.273)
Long firm name	-0.007	0.04	-0.027	-0.229	-0.761
	(0.046)	(0.063)	(0.044)	(0.160)	(0.495)
Firm name: w/ geogr. location	-0.225***	0.056	-0.092	-0.884***	-0.194
	(0.075)	(0.091)	(0.078)	(0.277)	(0.745)
Firm name w/ Danmark, Danish, Dan	0.284***	0.263**	0.151	0.167	-0.921
	(0.105)	(0.128)	(0.120)	(0.273)	(1.003)
Legal form: corporation	-0.880***	0.718***	-0.257***	0.468***	0.587
	(0.107)	(0.081)	(0.090)	(0.157)	(0.365)
Legal form: IVS	0.989***	-0.435***	-0.382***	0.281	-0.284
	(0.052)	(0.096)	(0.071)	(0.236)	(0.781)
Geogr. region: Midtjylland	0.187***	0.104*	0.007	0.201	-0.106
	(0.044)	(0.061)	(0.043)	(0.125)	(0.354)
Geogr. region: Nordjylland	0.285***	0.062	-0.065	-0.264	-0.532
	(0.062)	(0.085)	(0.061)	(0.215)	(0.691)
Geogr. region: Sjælland	0.186***	0.07	0.011	-0.097	0.113
	(0.053)	(0.075)	(0.055)	(0.181)	(0.431)
Geogr. region: Syddanmark	0.261***	0.133**	0.036	-0.023	0.361
	(0.048)	(0.064)	(0.047)	(0.147)	(0.337)
Patents at foundation	-0.351	0.659***	-1.232**	0.992***	3.220***
	(0.338)	(0.239)	(0.498)	(0.364)	(0.433)
In firm name: a proper danish word	0.088**	0.134***	-0.037	-0.167	-0.484*
	(0.035)	(0.047)	(0.035)	(0.102)	(0.285)
In firm name: Holding, capital, shares	0.069	-1.272***	0.072	-1.172***	-0.637
	(0.050)	(0.114)	(0.046)	(0.277)	(0.492)
In firm name: female name	-0.004	-0.226**	-0.081	-0.774***	0.064
	(0.058)	(0.088)	(0.060)	(0.273)	(0.570)
In firm name: male name	-0.079	0.092	-0.015	-0.207	-0.031
	(0.048)	(0.066)	(0.044)	(0.173)	(0.438)
At least one founders is firm	-0.803***	-0.355***	0.408***	0.231*	1.126***
	(0.049)	(0.067)	(0.048)	(0.129)	(0.422)
At least one of founder has female first name	0.072*	-0.123*	-0.01	0.195	0.764**
	(0.043)	(0.066)	(0.046)	(0.155)	(0.383)
Team	-0.051	0.242***	-0.166***	0.555***	-0.694
	(0.052)	(0.072)	(0.056)	(0.161)	(0.568)
# employees year 1: 0	0.142**	1.055***	0.198**	0.357*	0.491
	(0.072)	(0.074)	(0.086)	(0.207)	(0.510)
# employees year 1: 1	-0.276***	-0.671***	0.397***	0.563***	0.39
	(0.073)	(0.087)	(0.064)	(0.160)	(0.448)
# employees year 1: (2,4)	0.286***	-0.922***	0.430***	0.780***	0.581
	(0.062)	(0.092)	(0.068)	(0.143)	(0.431)
# employees year 1: (5,9)	0.263***	-1.276***	0.288***	1.248***	0.523
	(0.095)	(0.141)	(0.104)	(0.181)	(0.676)
# employees year 1: (6,49)	0.322***	-3.177***	0.540***	1.428***	-0.943
	(0.104)	(0.334)	(0.116)	(0.200)	(1.405)
No previous founding experience	-1.902***	0.878***	1.180***	0.298	0.342
	(0.062)	(0.105)	(0.092)	(0.194)	(0.593)
Previously founded 1-3 firms	-1.985***	1.071***	1.027***	0.354**	0.561
	(0.054)	(0.092)	(0.087)	(0.148)	(0.475)
Previously founded more than 3 firms	-2.563***	1.242***	0.929***	0.582***	1.246***
	(0.065)	(0.092)	(0.090)	(0.154)	(0.424)
Earlier invol. exit by one founder	0.574***	0.08	0.048	0.103	-0.32
	(0.061)	(0.075)	(0.062)	(0.168)	(0.366)
ln(total assets year 1)	-0.141***	0.079***	-0.192***	0.046	0.119
	(0.016)	(0.020)	(0.014)	(0.040)	(0.125)
Total assets year 1 is missing	-0.361***	0.916***	-0.929***	-0.108	1.366
	(0.115)	(0.153)	(0.102)	(0.307)	(0.913)
2nd quintile profits year 1	-0.497***	-0.347***	-0.158**	-0.278	0.523
	(0.083)	(0.110)	(0.079)	(0.203)	(0.492)
3rd quintile profits year 1	-0.725***	-0.022	-0.362***	-0.590**	0.489
	(0.089)	(0.112)	(0.084)	(0.244)	(0.527)
4th quintile profits year 1	-0.811***	0.077	0.505***	-0.121	0.631
	(0.086)	(0.099)	(0.072)	(0.179)	(0.508)
5th quintile profits year 1	-1.272***	0.236**	0.939***	-0.457**	-1.5
	(0.119)	(0.106)	(0.076)	(0.200)	(1.115)
Previous exit at same address	0.240***	-0.004	0.072	-0.035	-0.162
	(0.046)	(0.062)	(0.048)	(0.144)	(0.336)
>9 previous exits at same address	0.283***	-0.063	0.139	-0.262	-0.029
	(0.110)	(0.134)	(0.114)	(0.247)	(1.052)
Adress unshared	-0.015	0.057	0.173**	-0.399	0.263
	(0.088)	(0.121)	(0.088)	(0.292)	(0.756)
Adress shared with 2-5 other firms	0.136*	-0.091	0.054	-0.137	0.007
	(0.072)	(0.093)	(0.075)	(0.215)	(0.488)
Adress shared with 6-10 other firms	0.272***	-0.063	0.019	0.197	0.167
	(0.068)	(0.088)	(0.068)	(0.190)	(0.381)
Address previously used by 1-5 others	0.067	-0.065	0.008	-0.307*	0.011
	(0.055)	(0.082)	(0.053)	(0.178)	(0.633)
Address previously used by 6-10 others	0.079	0.056	-0.072	-0.614***	0.304
	(0.070)	(0.101)	(0.072)	(0.222)	(0.677)
Address previously used by 11-100 others	-0.127	-0.045	0.051	-0.578**	0.408
	(0.083)	(0.115)	(0.084)	(0.258)	(0.755)
Address previously used by > 100 others	-0.430***	0.031	-0.12	-0.1	-0.042
	(0.132)	(0.165)	(0.131)	(0.340)	(1.148)
ln(LIX)	-0.355***	0.227	0.085	0.308	0.507
	(0.105)	(0.148)	(0.113)	(0.307)	(0.806)
Mean word length	0.021**	-0.036***	-0.012	0.006	-0.093
	(0.009)	(0.013)	(0.009)	(0.029)	(0.085)
BPS lengths	-0.001	0.001	0.001	0.004***	0.004
	(0.001)	(0.001)	(0.001)	(0.001)	(0.004)
BPS 2nd quintile	0.034	0.018	0.024	-0.777***	0.331
	(0.054)	(0.077)	(0.057)	(0.213)	(0.527)
BPS 3rd quintile	0.019	-0.045	0.044	-0.209	0.28
	(0.061)	(0.084)	(0.064)	(0.195)	(0.639)
BPS 4th quintile	0.036	0.011	0.077	-0.221	1.108*
	(0.070)	(0.093)	(0.071)	(0.207)	(0.609)
BPS 5th quintile	0.049	-0.043	0.175**	-0.117	1.225*
	(0.090)	(0.119)	(0.089)	(0.234)	(0.675)
ln(frequency of least common word)	-0.008	0.014	0.008	0.016	-0.181**
	(0.012)	(0.016)	(0.013)	(0.032)	(0.089)
ln(frequency of most common word)	0.02	0.028	-0.035**	0.113**	0.265**
	(0.016)	(0.022)	(0.018)	(0.049)	(0.112)
Freq. least/freq. most common word	-0.003	0	-0.146*	-0.02	2.323***
	(0.077)	(0.108)	(0.078)	(0.261)	(0.630)

Appendix B: changes in AUC if variables are left out

Sets of variables included						Restricted models					
(i)	(ii)	(iii)	(iv)	(v)	Full model	# employees excluded	Diff. to full model (in %)	Patents excluded	Diff. to full model (in %)	Wordscores excluded	Diff. to full model (in %)
Involuntary exit											
x					0.623	0.623	0.000	0.623	-0.041	0.623	0.000
			x		0.719	0.719	0.000	0.719	0.000	0.719	0.000
				x	0.719	0.719	0.000	0.719	0.000	0.611	-0.177
x	x				0.683	0.683	0.000	0.683	0.020	0.683	0.000
x		x			0.740	0.734	-0.718	0.740	-0.010	0.740	0.000
x			x		0.745	0.745	0.000	0.745	-0.001	0.745	0.000
x				x	0.743	0.743	0.000	0.743	0.023	0.668	-0.112
x	x	x			0.763	0.762	-0.141	0.763	0.010	0.763	0.000
x	x		x		0.755	0.755	0.000	0.756	0.004	0.755	0.000
x	x			x	0.751	0.751	0.000	0.751	0.022	0.695	-0.080
x		x	x		0.801	0.798	-0.342	0.801	0.004	0.801	0.000
x		x		x	0.796	0.796	-0.030	0.796	0.006	0.755	-0.055
x			x	x	0.780	0.780	0.000	0.780	0.008	0.751	-0.038
x		x	x	x	0.823	0.823	-0.018	0.823	0.006	0.805	-0.022
x	x		x	x	0.784	0.784	0.000	0.784	0.009	0.759	-0.032
x	x	x		x	0.801	0.800	-0.020	0.801	0.011	0.769	-0.040
x	x	x		x	0.751	0.751	0.000	0.751	0.022	0.695	-0.080
x	x	x	x	x	0.824	0.824	-0.033	0.824	0.007	0.808	-0.020
Average change across affected models (in %)							-0.267		0.006		-0.066
High employment growth											
x					0.591	0.591	0.000	0.589	-0.294	0.591	0.000
			x		0.679	0.679	0.000	0.679	0.000	0.679	0.000
				x	0.789	0.789	0.000	0.789	0.000	0.662	-0.191
x	x				0.684	0.684	0.000	0.682	-0.317	0.684	0.000
x		x			0.694	0.652	-6.487	0.692	-0.397	0.694	0.000
x			x		0.699	0.699	0.000	0.697	-0.287	0.699	0.000
x				x	0.794	0.794	0.000	0.793	-0.071	0.682	-0.164
x	x	x			0.750	0.713	-5.272	0.749	-0.220	0.750	0.000
x	x		x		0.729	0.729	0.000	0.727	-0.239	0.729	0.000
x	x			x	0.787	0.787	0.000	0.787	-0.105	0.707	-0.114
x		x	x		0.743	0.715	-3.852	0.741	-0.219	0.743	0.000
x		x		x	0.829	0.801	-3.446	0.828	-0.064	0.738	-0.122
x			x	x	0.790	0.790	0.000	0.789	-0.082	0.721	-0.096
x		x	x	x	0.824	0.797	-3.418	0.823	-0.072	0.762	-0.082
x	x		x	x	0.785	0.785	0.000	0.785	-0.108	0.736	-0.067
x	x	x		x	0.825	0.795	-3.733	0.824	-0.083	0.766	-0.077
x	x	x		x	0.787	0.787	0.000	0.787	-0.105	0.707	-0.114
x	x	x	x	x	0.822	0.793	-3.604	0.821	-0.084	0.779	-0.055
Average change across affected models (in %)							-4.647		-0.172		-0.108
Return on assets of above 20 percent											
x					0.561	0.561	0.000	0.561	0.011	0.561	0.000
			x		0.629	0.629	0.000	0.629	0.000	0.629	0.000
				x	0.658	0.658	0.000	0.658	0.000	0.548	-0.201
x	x				0.596	0.596	0.000	0.596	0.053	0.596	0.000
x		x			0.608	0.610	0.262	0.608	0.027	0.608	0.000
x			x		0.637	0.637	0.000	0.637	0.018	0.637	0.000
x				x	0.661	0.661	0.000	0.663	0.171	0.575	-0.150
x	x	x			0.620	0.618	-0.347	0.621	0.060	0.620	0.000
x	x		x		0.640	0.640	0.000	0.640	0.017	0.640	0.000
x	x			x	0.652	0.652	0.000	0.653	0.132	0.598	-0.090
x		x	x		0.654	0.654	0.116	0.654	0.015	0.654	0.000
x		x		x	0.674	0.676	0.206	0.675	0.115	0.612	-0.102
x			x	x	0.677	0.677	0.000	0.677	0.060	0.638	-0.061
x		x	x	x	0.686	0.687	0.143	0.686	0.047	0.654	-0.049
x	x		x	x	0.670	0.670	0.000	0.671	0.053	0.639	-0.048
x	x	x		x	0.666	0.666	0.004	0.666	0.112	0.621	-0.072
x	x	x		x	0.652	0.652	0.000	0.653	0.132	0.598	-0.090
x	x	x	x	x	0.679	0.679	0.051	0.679	0.055	0.652	-0.041
Average change across affected models (in %)							0.123		0.068		-0.091
At least one new patent											
x					0.721	0.721	0.000	0.691	-4.141	0.721	0.000
			x		0.752	0.752	0.000	0.752	0.000	0.752	0.000
				x	0.880	0.880	0.000	0.880	0.000	0.693	-0.270
x	x				0.795	0.795	0.000	0.773	-2.755	0.795	0.000
x		x			0.787	0.761	-3.428	0.772	-1.932	0.787	0.000
x			x		0.794	0.794	0.000	0.788	-0.733	0.794	0.000
x				x	0.894	0.894	0.000	0.893	-0.171	0.762	-0.173
x	x	x			0.814	0.800	-1.808	0.798	-1.992	0.814	0.000
x	x		x		0.831	0.831	0.000	0.824	-0.866	0.831	0.000
x	x			x	0.898	0.898	0.000	0.897	-0.179	0.812	-0.106
x		x	x		0.825	0.816	-1.080	0.820	-0.578	0.825	0.000
x		x		x	0.903	0.898	-0.529	0.902	-0.145	0.805	-0.121
x			x	x	0.899	0.899	0.000	0.898	-0.100	0.805	-0.116
x		x	x	x	0.903	0.901	-0.271	0.902	-0.102	0.830	-0.088
x	x		x	x	0.901	0.901	0.000	0.900	-0.097	0.836	-0.077
x	x	x		x	0.906	0.900	-0.632	0.904	-0.177	0.828	-0.094
x	x	x		x	0.898	0.898	0.000	0.897	-0.179	0.812	-0.106
x	x	x	x	x	0.904	0.901	-0.360	0.903	-0.116	0.845	-0.071
Average change across affected models (in %)							-1.920		-0.891		-0.122
Participation in an innovation subsidy program											
x					0.745	0.745	0.000	0.590	-20.775	0.745	0.000
			x		0.745	0.745	0.000	0.745	0.000	0.745	0.000
				x	0.880	0.880	0.000	0.880	0.000	0.751	-0.172
x	x				0.758	0.758	0.000	0.634	-16.379	0.758	0.000
x		x			0.763	0.764	0.034	0.641	-15.986	0.763	0.000
x			x		0.769	0.769	0.000	0.739	-3.965	0.769	0.000
x				x	0.872	0.872	0.000	0.865	-0.893	0.789	-0.105
x	x	x			0.774	0.774	0.035	0.671	-13.318	0.774	0.000
x	x		x		0.775	0.775	0.000	0.739	-4.627	0.775	0.000
x	x			x	0.862	0.862	0.000	0.865	0.411	0.797	-0.081
x		x	x		0.784	0.781	-0.397	0.732	-6.532	0.784	0.000
x		x		x	0.863	0.865	0.308	0.854	-1.036	0.792	-0.089
x			x	x	0.835	0.835	0.000	0.853	2.175	0.789	-0.058
x		x	x	x	0.836	0.835	-0.022	0.842	0.699	0.792	-0.055
x	x		x	x	0.831	0.831	0.000	0.850	2.281	0.790	-0.052
x	x	x		x	0.856	0.856	0.068	0.852	-0.459	0.799	-0.071
x	x	x		x	0.862	0.862	0.000	0.865	0.411	0.797	-0.081
x	x	x	x	x	0.834	0.832	-0.278	0.840	0.687	0.796	-0.048
Average change across affected models (in %)							-0.031		-4.832		-0.081

(i) basic G/S variables, (ii) extended set of G/S variables, (iii) human capital variables, (iv) firm characteristics and (v) BPS variables.