Essays on Contest Experiments and Supervised Learning in the Pharmaceutical Industry

Ph.D. in Economics, Management and Data Science

XXXII Cycle

By

Jan Niederreiter

2020

The dissertation of Jan Niederreiter is approved.

Program Coordinator:
Prof. Dr. Massimo Riccaboni, IMT School for Advanced Studies Lucca

Supervisor:
Prof. Dr. Massimo Riccaboni, IMT School for Advanced Studies Lucca

The dissertation of Jan Niederreiter has been reviewed by:

Prof. Dr. Andrea Mina, Sant'Anna School of Advanced Studies Pisa

Prof. Dr. Michael Hopkins, University of Sussex

Prof. Dr. Pietro Panzarasa, Queen Mary University of London

IMT School for Advanced Studies, Lucca
2020

Contents

List of Figures

List of Tables

List of Abbreviations

3L	Three player lottery contest
3S	Three player share contest
5 CF	5-fold cross validation
5L	Five player lottery contest
5NN	Five nearest neighbor
5S	Five player share contest
ACC	Accuracy
AUC	Area under the ROC curve
BART	Bayesian additive regression tree
BART_IP	BART variable inclusion proportion
BIC	Bayesian information criterion
BR	Best response
C5.0	Boosted Decision Trees using C.5.0 algorithm
CAR	Cumulated abnormal return
$\overline{\text{CAR}}$	Acerage cumulated abnormal return
CC	Complete Cases
C-Lasso	Classifier Lasso
DCF	Discounted cash flow
DI	Development index
DISCR	Backward/forward probabilistic regression
DT	Decision Tree
EMA	European Medicine Agency
EWA	Experience-weighted attraction model

FDA	Food and Drug Administration
FFT	Fast and frugal decision tree
FM	Fixed matching
HIST	Historical method
IC	information criterion
II	Internal imputation
IND	Investigational new drug
indlev3	Indication level three
LASSO	Least absolute shrinkage and selection operator
LOGIT	Linear logistic regression
ML	Machine Learning
MoA	Mechnism of Action
NDA	New drug application
NME	New molecular entity
NPV	Net present value
OLS	Ordinary least squares
PI	Phase one
PII	Phase two
PIII	Phase three
Port	Product portfolio importance
Prob	Product success probability
PROBIT	Linear probabilistic regression
QRE	Quantal response equilibrium
R&D	Research and development
RF	Random forest
RF_SE	Random Forest successive variable elimination
RM	Random matching
ROC	Receiver operating characteristic curve
SL	Supervised Learning

Acknowledgements

Chapter 2 is based on the paper *"Learning and drop out in contests: An experimental approach."*, co-authored with Francesco Fallucchi and Massimo Riccaboni. The paper has been submitted to *Theory and Decision* (Springer) and is currently under review.
Chapter 3 is based on the paper *"Identifying types in contest experiments"*, co-authored with Francesco Fallucchi and Andrea Mercatanti. The paper has been submitted to *International Journal of Game Theory* (Springer) and is currently under review.
Chapter 4 is based on the paper *"Improving The Prediction of Clinical Success Using Machine Learning"*, co-authored with Bernard Munos and Massimo Riccaboni. The paper has been submitted to *Nature Biotechnology* (Springer) and is currently under review.
Chapter 5 is based on the paper *"The Impact of Product Innovation on Firm Value: Evidence from the Bio-Pharmaceutical Industry"*, co-authored with Massimo Riccaboni. The paper has been submitted to Industrial and Corporate Change (Oxford University Press) and is currently under review.

Over the last three years, I experienced great support from my community, was constantly encouraged to broaden my horizons, and moreover got the great opportunity to work with people from many different cultural and academic backgrounds. Therefore, I am very grateful to have been part of the IMT micro universe and will always treasure the time spent in Lucca.

In particular, I would like to express my gratitude to Prof. Massimo Riccaboni for his valuable support and helpful guidance while supervising my Ph.D. project. Brainstorming together about new research ideas, discussing potential solutions to open problems, and deciding on next project steps never felt exhaustive but rather energizing. Over the last years I learned a lot from working with him, and I am grateful for his patience, his positive way of thinking, and his ability to challenge my work in a constructive way. Moreover, he introduced me to many supportive colleagues including researchers that eventually became collaborators of chapters of this Ph.D. project.

First, I would like to deeply thank Dr. Francesco Fallucchi for his hands-on support, his helpful feedback and his valuable insights that have aided me to learn more about experimental economics and about academic writing in general. Thanks to his leap of faith, we started our collaboration already in my first Ph.D. year and then extended it on a second project. Here, I had the pleasure to work with Dr. Andrea Mercatanti, who I want to especially thank for providing me guidance with his strong background in statistics.

Further, I am especially grateful for having had the great opportunity to work with Bernard Munos whose remarks greatly influenced the focus of the third chapter. From him I learned a lot about how to integrate results in order to address an audience of practitioners in the pharmaceutical industry.

The research questions in chapter three and four would not have been possible to address without the support of Evaluate Ltd. Not only did their team grant us the opportunity to use their data platform as empirical base, but also supported the project indirectly with interesting information and insightful comments during various online meetings. Thank you, Jon, Chris, Emily, Anthony, Anastasia and many others for providing such a collaborative and friendly working environment.

I take the opportunity to also thank Prof. Dr. Andrea Mina, Prof. Dr. Michael Hopkins, and Prof. Dr. Pietro Panzarasa for having dedicated their time to thoroughly evaluate my thesis, to provide valuable suggestions, and to point me to interesting considerations for future work.

Further, I am very thankful for all the colleagues and scholars that in seminars, workshops, or conferences have somehow positively influenced with their talks, feedback, and informal conversations this Ph.D. project. The thanks obviously extend to the professors at IMT whose offered coursework provided me with a great foundation for further research. At this point I would also like to thank Prof. Ennio Bilancini for granting me the opportunity to assist teaching his class "Evolutionary Game Theory".

On a more personal note, I would like to thank all my friends and family that have morally supported me during the last three years. I always felt that I can turn to you, if need be. The resulting tranquility has helped me throughout my whole Ph.D. experience. Especially, thank you Lisa for your valuable suggestions while proofreading my thesis. Obviously, my gratitude extends to my Cuban family that has always kept me in their thoughts and prayers.

Lastly, I would like to dedicate a couple of words to my fiancée Isabel who has never stopped supporting me, to motivate me, and to cheer me up. Whenever I felt blue, she was there for me, regardless of her own workload and her own struggles. In fact, I can entrust her with any kind of problem and there would be nobody who understands me like she does. I feel extremely grateful for having shared all the ups and downs of my Ph.D. life with her, for growing together, and finally for making our time at IMT Lucca the beginning of something bigger.

Vita

September 08, 1990	Born in Bad Nauheim, Germany
Education	
2017–	Ph.D. student in Economics, Management and Data Science (XXXII cycle) IMT School for Advanced Studies Lucca, Italy
2015	Double Degree M.Sc. in "European Economics" Final mark: 1.0, 110/110 cum laude Eberhard Karls Universität Tübingen, Germany and Università degli studi di Pavia, Italy
2013	B.Sc. in "Business Administration" Final mark: 1.3 University of Mannheim, Germany Semester abroad at Free University of Bolzano, Italy
2010	University orientation year Major in "Computer Science" and "Psychology" University of Idaho, USA
2009	High school diploma (Abitur) Final mark: 1.1 Gymnasium Nidda, Germany

Working Experience

01/2016 – 01/2017	Allianz Deutschland AG, Munich Allianz Inhouse Consulting
08/2014 – 09/2014	Deutsche Bank AG, Frankfurt a.M Investment Products and Insurances
06/2013 – 09/2013	Commerzbank AG, Frankfurt a.M Group Risk Management, Market Models
02/2013 – 05/2013	University of Mannheim Chair Prof. Dr. Biemann for Personnel management and leadership
09/2011 – 05/2012	University of Mannheim Chair Prof. Dr. Bless for Social psychology
06/2011 – 08/2011	Consultancy Schlegel und Partner, Weinheim Department vehicle engineering
08/2010	Axel Springer AG, Dortmund BILD editorial department

Scholarships and Awards

08/2019	Frontier Proposal Fellowship 2019 of IMT School for Advanced Studies Lucca
07/2016	Award for best laureate 2015 of the Faculty of Economics of Università degli studi di Pavia
06/2016	Werner-Diez-Prize for the best double degree of the University of Tübingen - Faculty of Economics
2011 – 2015, 2019 – 2020	Scholarship of the German Academic Scholarship Foundation

Presentations

1. PhD in Economics - Workshop 2019, *What Drives Market Reactions after R&D Announcements? Empirical Evidence from the Pharmaceutical Industry*, 01-03.10.2019, Scuola Superiore Sant'Anna, Pisa, Italy.
2. 14th annual conference - Warsaw International Economic Meeting 2019, *What Drives Market Reactions after R&D Announcements? Empirical Evidence from the Pharmaceutical Industry*, 02-04.07.2019, University of Warsaw, Poland.
3. NEXT GEN Paper Development Workshop, *Elicitation of investor sentiment on stock market events via machine learning*, 13.06.2019, University of Pisa, Italy.
4. Young Economists of Tuscan Institutions, *Identifying types in contest experiments*, 12.04.2019, University of Siena, Italy.
5. Prague Conference on Behavioral Sciences 2019, *Market return asymmetry of clinical trial outcomes*, 05-06.04.2019, Center For Behavioral Experiments, Prague, Czech Republic.
6. Workshop on "Modelling Scientific Ecosystems: Medicine and Pharmacology", *Clinical trial success prediction: technological foresight via supervised learning*, 14.01.2019, University of Ancona, Italy.
7. Workshop on "Recent Advances in Public Economics and Quantitative Methods", *Learning and Drop Out in R&D contests: An Experimental Approach*, 09.03.2018, IMT School for Advanced studies Lucca, Italy.
8. BRIGHT La Notte dei Ricercatori in Toscana 2017 - Aspettando Bright, *La competitività delle imprese: quali scenari per Lucca?* 26.09.2017, IMT School for Advanced studies Lucca, Italy.

Abstract

The field of economics witnesses a growing interest to better understand how individuals form decisions and how these decisions can be supported with the help of sophisticated data-mining tools. The first half of the thesis analyzes investment decisions in a competitive environment that is free of confounding factors by means of experimental data. The second half focuses on how supervised learning algorithms can be applied to predict the outcome and perceived success probability of pharmaceutical projects for aiding decisions of policy makers, managers and financial investors.

More specifically, chapter two contrasts repeated individual expenditure decisions in different contest treatments by varying the uncertainty of the outcomes and the number of contest opponents. Contests with probabilistic outcomes show decreasing over-expenditures and a higher rate of "drop out". If outcomes become deterministic, expenditures quickly convergence towards equilibrium predictions and a near to full participation. These results are robust to changes in the number of opponents. A learning parameter estimation using the experience-weighted attraction model suggests that subjects adopt different learning modes across different contest structures and helps to explain expenditure patterns deviating from theoretical predictions.

Chapter three explores the presence of latent contestant types by applying the classifier Lasso to two versions of contest experiments, one that keeps the grouping of contestants fixed and one that randomly regroups contestants after each round. Results suggest that there exist three distinct types of players

in both contest regimes. The majority of contestants in fixed groups behaves reciprocal to opponents' previous choices. For experiments in which contestants are regrouped, the share of "reciprocators" is significantly lower. In both cases the reaction of other player types seems to differ from what is expected from a myopic best-reply.

Pharmaceutical drug development can be seen as a real-world example for lottery contest. Assessing drug candidates' odds of success often relies on methods based on historical success rates. However, machine learning offers a more data-driven approach to identify promising projects. To evaluate its usefulness, chapter four assesses the performance of several supervised learning algorithms that are trained and validated on a large database of projects. Using a sizeable list of project characteristics as input data, classification via state-of-the-art supervised learning methods is more accurate compared to more simplistic methods. The chapter aids stakeholders in the pharamceutical industry to make more informed decisions regarding stage-specific project outcomes.

Chapter five extends the study of project outcomes by assessing the relationship between product innovation announcements of bio-pharmaceutical companies and their stock reactions using an event study approach. We hypothesize that financial returns that follow news on product innovation are shaped by a "probability effect", that depends on how investors perceive the product's likelihood of success, and a "portfolio effect" that depends on the relative importance of a product within a company's portfolio. To test for the probability effect, project specific success probabilities are estimated via supervised learning methods. The portfolio effect is measured by the share of the product's net present value. Market reactions are found to be higher when assosciated to projects with high portfolio importance but lower when associated to projects with high expected success probability.

Chapter 1

Introduction

1.1 General Introduction

In January 2018, the multinational company Pfizer announced its withdrawal from the research and development (R&D) intensive Alzheimer market, making hundreds of neuroscience discovery jobs obsolete. It was a setback for the whole sector which permanently suffers to launch successful cures for a market affecting approximately 44 million patients worldwide (Crow, D., 2018). Pfizer is not an isolated case: in late 2016, Eli Lilly released its phase III clinical trial failure which was followed by the announcement of Merck's failure only a couple of months afterwards. In environments characterized by high outcome uncertainty, such as Alzheimer research (Tanzi and Parson, 2001) and more generally drug development, fully rational decisions are likely to be impaired (Dosi et al., 2001), which renders learning from previous unsuccessful attempts often not fruitful. The consequence is an inefficient allocation of efforts in a trial-and-error search regime (Dosi et al., 2001), which in the pharmaceutical drug development can easily result in investment losses in the hundreds of millions. One trivial way to avoid such a dilemma, as in the above-mentioned case, is stopping future ventures which however impedes innovative advancement of the industry. The problem of resource misallocation seems to be pervasive in the bio-pharmaceutical industry,

which has witnessed excessive investment in some and too little investment in other research areas (see, for example Hopkins et al., 2007; Jones and Wilsdon, 2018), resulting in a low R&D productivity during recent years (Lendrem et al., 2015; Pammolli et al., 2011).

Drug discovery in the pharmaceutical industry is one prevalent example of a winner-take-all market in which companies compete for a single prize, the exclusive market rights of a drug, whose assignment ultimately depends on the relative efforts of market participants. Such environments, that feature high outcome uncertainty with current project success independent of investments in past projects, have often been conceptualized as lotteries (Sutton, 1998; Malerba and Orsenigo, 2002; Garavaglia et al., 2013). To better understand how investment choices are characterized in these environments, we thus represent them in chapter 2 and 3 by different forms of (lottery-) contests (Tullock, 1980). Since the empirical investigation of repeated investment choices is plagued by severe methodological limitations in non-controlled applied industrial organization research, such as spurious regression (Bennett and J. Snyder, 2017) and survivor bias (Denrell, 2003), we opt for an experimental setting. Laboratory data allows us to directly observe individual choices free of confounding factors that usually accompany the empirical analysis of field studies. For example, individual investments in contest situations can often not be directly measured and the indirect measure, an individual's performance, could be confounded by other factors such as personal ability (Dechenaux et al., 2015). By characterizing differences in observed individual investment behavior we can draw tentative conclusions that relate to similar real-world settings.

One major difference between contest experiments and real-world contest settings in the pharmaceutical industry is that the probability of a project's success is not simply determined by a company's relative share of R&D investment but rather by a complex interplay of various project characteristics. Since neither the set of project characteristics that impacts the success probability nor their interactions are commonly known, it remains challenging to approximate a project's probability of success. Chapter 4 proposes that machine learning algorithms trained on infor-

mation of completed projects can be used to learn more about success factors and to improve the classification of the outcome of projects.

The steady increase in computational power combined with progress in storing and disseminating large quantities of information has set the stage for operating state-of-the-art machine learning algorithms that can be flexibly used for a multitude of tasks. Prominent examples include face and speech recognition (Yang et al., 2002; Deng and Li, 2013), but novel applications go as far as predicting earthquakes (Rouet-Leduc et al., 2017). In the medical-pharmaceutical area, machine learning tasks already form an integral part (Ziad Obermeyer, 2016), for example, for cancer prognosis (Kourou et al., 2015), for encoding neuronal activity (Bouton et al., 2016), and for screening of new pharmaceutical compounds (L. Zhang et al., 2017). All these applications have in common that they require the assessment of a lot of information to form some sort of decision. The rules to form a decision are however not known a-priori but are rather established by the algorithm itself using the available information. The algorithm is colloquially said to learn.

In the pharmaceutical industry, historic information on successful and unsuccessful projects has been collected for many years and has been made publicly available in large online repositories such as *clinicaltrials.gov*. Yet, it goes beyond current expertise to decide what piece of information is indicative for a project's success and to what degree. We show that supervised learning algorithms, after being trained on an array of observable project characteristics, better distinguish successful from unsuccessful projects than, for example, traditional methods that consider only the historic success rate of similar projects. Shifting from exploring how individuals make investment decisions in situations of high outcome uncertainty in chapter 2 and 3, we present in chapter 4 a data-driven approach to support human decision-making in such environments.

Business environments characterized by high outcome uncertainty do not only create challenges for companies operating in them but also for financial investors that wish to partake at risky endeavors. Especially news that relate to product innovation in the pharmaceutical industry

can cause substantial shifts in companies' stock prices (see for example Urbig et al., 2013). Yet, even when employing modern machine learning algorithms, predicting future stock market returns is only possible to a limited extent (Gu et al., 2018). Prior to the prediction of market movements, one needs to develop a sound understanding of common characteristics that drive price variation as an aftermath of innovation announcements. We hypothesize in chapter 5 that there are two main drivers for such of price changes, which can be approximated by using a detailed database of pharmaceutical projects and a combination of supervised learning techniques. The reported results provide new insights on market reactions after innovation announcements in environments for which the outcome of innovation is highly uncertain.

This thesis addresses open challenges of environments of high outcome uncertainty from different angles. Chapter 2 and 3 approach questions related to how individual decisions can be characterized in such environments via assessing different dimensions of contest experiments. Chapter 4 and 5 use large pharmaceutical data sets and various supervised learning approaches to aid future decision-making in these environments. The next section of the introduction addresses the structure of the thesis by summarizing the methodology and main findings of each chapter separately and in more detail.

1.2 Structure and Summary of the Dissertation

The thesis is organized in six chapters of which the first chapter provides a general introduction of the topic and sketches the content of the following chapters.

Chapter 2 explores how individual expenditures can be characterized in repeated contest experiments under four different experimental treatments. The treatments vary in the level of competition using groups of three or five individuals and differ in the degree of outcome uncertainty by using winner-take-all and proportional prize contest payoff functions. Besides reporting average expenditure levels, chapter 2 focuses on the

share of zero expenditures, a feature that has not received much attention in the existing contest literature. We find that players abstain more often from expending positive amounts in treatments with higher competition as well as in treatments with higher outcome uncertainty. Interestingly, the share of zero expenditures in the "high-competition, high-outcome-uncertainty" treatment rises significantly over the rounds played, which cannot be explained by weighted forms of fictitious play that consider previous opponent effort choices. Estimation of the experience-weighted attraction model (C. Camerer and T. H. Ho, 1999) reveals that participants in winner-take-all contests rather employ reinforcement learning strategies (Roth and Erev, 1995) than evaluating hypothetical payoffs of not chosen strategies. This behavior is in line with the finding that participants in winner-take-all contests decrease their expenditures after a series of losses. The chapter concludes with drawing parallels and implications to real world contest settings.

Chapter 3 breaks with the established tradition of contest experiments to focus on average effects for each treatment. Instead, we adapt the classifier-Lasso (Su et al., 2016) to be used in a fixed effects Tobit model in order to detect player types that differ in how their expenditures relate to information received from previously played rounds. We detect latent player types in two repeated contest regimes. In the first, participants compete against the same group of opponents over all contest rounds. In the second, participants are randomly regrouped after each round. In both contest regimes the optimal number of latent player types is estimated to be three, whereas the type composition changes. The fraction of players that reacts reciprocal to previous opponent expenditures is significantly higher in contests under which group assignment stays unchanged. In these treatments we find that groups with more "reciprocators" display lower average expenditures which hints at their collaborative disposition. The chapter ends by highlighting the importance of considering heterogeneous behavior in competitive situations under outcome uncertainty.

Chapter 4 approaches the challenge of outcome uncertainty in drug development by applying machine learning techniques to classify the out-

come of pharmaceutical projects based on available historical data. After explaining in detail the data set characteristics and the supervised learning approach, the chapter continues by comparing the classification results of different supervised learning methods and common estimation methods. We find that the BART algorithm (Kapelner and Bleich, 2013) delivers the most accurate classification results, significantly different from classification results of simple decision heuristics based on historical success rates of similar projects. The best-in-class approach is then used to predict outcomes of the current development pipeline and to reveal which groups of projects are predicted to be most successful. Further, we analyze which project characteristics are most often selected by different algorithms to be relevant for the outcome classification of the project. As product development in the pharmaceutical industry is a very risky and costly business, possessing insightful machine learning estimations on projects' success propensity is undoubtedly valuable for many stakeholders including company managers, policy makers and financial investors.

The study of project outcomes extends to chapter 5 by assessing the relationship between product innovation announcements of bio- pharmaceutical companies and their stock reactions using an event study approach. After setting up a general company stock reaction model, we derive a set of hypotheses stating that returns around project-specific news are shaped by two effects: a "probability effect", that depends on how investors perceive the project's likelihood of success, and a "portfolio effect" that depends on the relative importance of a project within a company's portfolio. To test for the probability effect, project-specific success probabilities are estimated via a combination of supervised learning methods. The portfolio effect is measured by the product's net present value over the net present value of the company right before the innovation announcements. Various regression specifications and robustness checks confirm that market reactions after product innovation announcements are higher if they are associated to products of high portfolio importance but lower if they are associated to products with a high success probability. The last part of the chapter discusses the results putting

special attention to strategies that could potentially mitigate the risks of negative innovation announcements.

Lastly, the conclusive chapter 6 wraps up the findings of each chapter and puts them into context by discussing their limitations, potential policy implications, and possible future extensions.

Chapter 2

Learning and Drop Out in Contests: An Experimental Approach

2.1 Introduction

Settings for which outcomes are characterized by high uncertainty are likely to undermine fully rational decisions, which can render learning from previous attempts not successful. As a consequence, this process may lead to an inefficient allocation of efforts in a trial-and-error search regime, that can easily results in losses (Dosi et al., 2001). The aim of this study is to identify, by means of experiments, how the payoff structure and the level of competition affect individual efforts in competitive settings with uncertain outcomes.

Our candidate setting is the Tullock rent-seeking contest in which subjects compete for a single prize, whose assignment probability depends on the relative share of subjects' efforts (Tullock, 1980). In rent-seeking contests, subjects persistently deviate from what standard game theoretical models predict. A survey of the experimental contest literature by

Dechenaux et al. (2015) has highlighted that contestants spend on average considerably more than the theoretical equilibrium. Different motivations have been proposed to explain this phenomenon, such as joy of winning, probability distortion and impulsive behavior (see Sheremeta, 2018).

Most studies find an overall decrease in expenditures when subjects repeatedly play this game (e.g. Cason, Sheremeta, et al., 2012) which is usually attributed to learning without further specifying the process behind it. Another empirical regularity that usually remains uncommented is that participants often choose not to spend any resources in the contest. Yet, by looking across a subset of experimental studies on lottery contests, we find that zero expenditures are indeed a frequent, sometimes even modal, choice of participants. The fraction of zero plays is higher for larger competing groups and is increasing in late stages of the experiment. We hypothesize that the frequency of non-bidding participants increases with the perceived uncertainty about winning the prize. The uncertainty of winning is higher in single-prize contests and increases with the number of competitors. Moreover, if the share of zero expenditures increases in the course of the repeated contest experiment, it could be a consequence of individuals learning not to expend, which deserves a closer investigation.

To explore how the uncertainty of outcomes and the number of opponents affect behavioral patterns such as zero expenditures in contest settings, we set up a laboratory experiment in which we compare, over sixty periods, participants' expenditure choices in the standard winner-take-all Tullock contest versus a non-probabilistic equivalent share contest (L. Friedman, 1958). The winner-take-all (lottery) contest allows only for one winner of the prize, whose winning probability is proportional to the share of own investments over the total group investments. In the deterministic (share) contest, contestants receive a fraction of the prize proportional to their percentage in the group investment. Under the assumption of risk-neutrality, both contest settings are equivalent in terms of equilibrium predictions. Varying the contest type and the group size of three and five contestants, we create a 2x2 experimental design. Our

analysis extends to competition under different group size in order to test if the increase in the level of competition (see Huck, Normann, and Jörg Oechssler, 2004) may exacerbate zero expenditures, as suggested by our meta-analysis, linked to lower earnings in share contests or more frequent expected losses in winner-take-all contests.

We find that the average levels of effort observed in share contests under both group sizes are well described by standard game theoretical predictions. Conversely, in the lottery contests we are unable to distinguish total group expenditures between the two group sizes. The decline of average expenditures in the five-player lottery contests coincides with a significant increase in what we label "drop outs", i.e. zero expenditures that are not justified either by the myopic best-response or weighted fictitious play. Drop outs are instead significantly less frequent in share contests and, if anything, decrease over time. The distinct expenditure patterns found between the two settings suggest that differences in the contests' payoff structures affect subjects' learning process.

We test this hypothesis by estimating the experience-weighted attraction model (EWA, C. Camerer and T. H. Ho, 1999). The results reveal that lottery contestants learn significantly more from their own past payoffs than players in the share contests (experiential or reinforcement learning Roth and Erev, 1995). Our results support recent findings by Alós-Ferrer and Ritschel (2018) on subjects' frequent use of the reinforcement heuristic "win-stay, lose-shift" rather than a more reasoned approach based on myopic best-response. The reliance on experiential learning in lottery settings can explain both the decreasing expenditures over time and the increasing propensity of zero expenditures choices. The more often a lottery participant loses, the more she will discourage positive expenditures up to non-participation. Further regression analyses confirm that expenditures decline significantly with an increase in prior accumulated losses. Since experienced losses are more frequent in larger competing groups, expenditures are expected to decrease at a faster rate. This line of thought explains the absence of differences in total rent-seeking between groups with signficantly different size in lottery contests, contrary to theoretical predictions.

Previous experimental studies have explored subjects' behavior in contests, whose design or methods partially overlap with ours. For example, Lim et al., 2014 investigate how lottery contestants' choices differ in a ten-period setting by varying the group size whereas Parco et al., 2005 use a reduced version of the EWA to compare simulated choices with mean expenditures in a two-stage lottery contest. Differently from them, we aim at comparing contestants' behavior using the share contest as a benchmark, as well as highlighting how different contest structures may lead to different drop out rates and learning modes over a longer experimental set-up (60 periods). Another strand of research has compared subjects' behavior in contests with different payoff structures (e.g. Fallucchi, Renner, and Sefton (2013), Chowdhury, Sheremeta, et al. (2014), Ghosh and Hummel (2018)), but none has so far explored their interaction with different group size. Fallucchi, Renner, and Sefton, 2013 propose that different behavior across contests with different feedback structure stems from differences in learning methods, yet do not test this suggestion. Similarly, the high fraction of zero expenditures has often been attributed to myopic best-reply without proper analysis. Recently, Pangallo, Heinrich, et al. (2019) show that in repeated competitive settings equilibrium predictions are usually an unrealistic assumption. Moreover, Pangallo, Sanders, et al. (2017) show in games with a stochastic payoff component, under certain assumptions, models of learning such as EWA do not converge to the equilibrium. In this chapter we investigate experimentally how expenditure dynamics in general, and zero expenditures in particular, can be explained by distinct learning mechanisms across contest structures.

The chapter is organized as follows: section 2.2 introduces the contest forms and offers a brief review of the related experimental literature on contests, a meta-analysis on zero expenditures in lottery contests, and a review of learning modes in related situations. Section 2.3 presents the experimental design and procedures. The experimental results are presented and discussed in section 2.4. The first, descriptive, part of our result section highlights differences in group expenditures and in the fraction of zero expenditures. The second part presents the EWA model

estimations and adds support for different learning modes across contest games. We conclude in section 2.5 with a discussion of our findings and argue how these results may well represent some empirical regularities in winner-take-all settings outside the experimental literature.

2.2 Theoretical Background and Experiments

The Tullock model of rent-seeking (Tullock, 1980) has been extensively applied for modeling competitive situations such as lobbying, patent races, litigation lawsuits, grant seeking, etc. (Konrad, 2009). In the simplified model, often referred to as lottery or winner-take-all contest, N agents compete for a prize of size V, where x_i is the amount of expenditure of agent i and X is the aggregate expenditure. The individual profits π_i depend on all agents' expenditures, the prize assignment, and a homogeneous initial endowment denoted by e:

$$\pi_i = \begin{cases} e - x_i + V & \text{with probability} \quad x_i/X \\ e - x_i & \text{otherwise.} \end{cases} \tag{2.1}$$

Therefore, the probability of one agent to receive the prize increases with own expenditures but decreases with the expenditures of others. In an alternative version of the contest, also known as share or proportional prize contest, the prize is not assigned to one agent only, but it is divided across all Nagents proportionally to the fraction of own expenditures x_i and aggregate expenditures X. Thus, each agent with positive expenditure receives a share of the prize. The payoff function in this case is equal to:

$$\pi_i = e - x_i + V(x_i/X) \tag{2.2}$$

The two contests share the same expected payoffs and, under the assumption of risk neutrality, the same individual effort equilibrium predictions, where $x_i^* = V(N-1)/N^2$. However, the realized payoff in the

lottery contest differs from the one in the share contest plausibly due to the stochastic winner-take-all nature of the game.

Winner-take-all and share contests have frequently been used to model a vast array of competitive situations. The early work by L. Friedman, 1958 constitutes a first attempt to use the share contest to model the allocation of advertisement budget across media. Proportional-prize assignments are also observed in electoral schemes (Schram and Sonnemans, 1996), lobbying (Krueger, 1974) and labor compensation (Kruse, 1992). Winner-take-all contests are used to model situations in which the choice of the winner does not depend solely on efforts. Applications range from the seminal rent-seeking hypothesis by Tullock (1980), to political polls (J. M. Snyder, 1989), sport tournaments (Szymanski, 2003), and patent races (Fudenberg et al., 1983; Harris and Vickers, 1985).

As it is often difficult to capture expenditure dynamics with field data, laboratory experiments have become increasingly popular in recent years to characterize behavior in different contest settings.[1] Many experiments support pervasive over-dissipation in lottery contests paired with high heterogeneity in effort levels across contestants (e.g. Millner and Pratt, 1989; Sheremeta and J. Zhang, 2010; Mago, Samak, et al., 2016). Based on a sample of thirty studies, Sheremeta, 2013 report a median overbidding rate of 72% compared to the equilibrium predictions. Contrary to the lottery contests, share contests display less variation in individual spending behavior and a quicker convergence over time towards to the predicted equilibrium level (Fallucchi, Renner, and Sefton, 2013; Chowdhury, Sheremeta, et al., 2014; Cason, Masters, et al., 2018).

Since it is common to focus on mean expenditures when analyzing overbidding in contests, the choice of zero expenditures has often been overlooked. We summarize the data of seven contest experiments, considering in total ten independent standard repeated lottery treatments (as specified in equation 2.1). Table 2.1 shows that zero expenditures are

[1] The following review of the experimental contest literature puts our research in context, while acknowledging that, given the volume of research in this area, it is only partial and suitable for the scope of the chapter. See Dechenaux et al. (2015) for a broad overview of contest experiments.

Table 2.1: Meta-analysis of zero expenditures plays in lottery contests

Group size	2	4
Number of observations	996	9000
Mode expenditures equal to zero	No	Yes
Share of zero expenditures	3.9%	12.0%
Share of zero expenditures that are not myopic best-response (drop out)	3.5%	6.0%
Share of drop out over share of zero expenditures	89.7%	50.0%

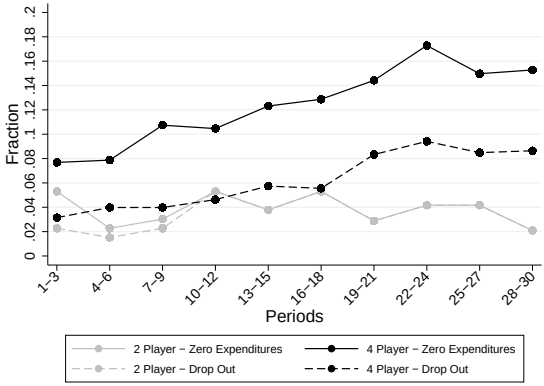
The meta-analysis incorporates data from ten treatments of seven publications using repeated standard lotteries. Two players: Abbink et al., 2010, Cason, Sheremeta, et al., 2012. Four players: Mago, Samak, et al., 2016, Sheremeta, 2010, Sheremeta and J. Zhang, 2010, Price and Sheremeta, 2011, Sheremeta, 2011.

the modal choice in four-player lottery contests making up 12% of the total choices and are increasing over time (see figure 2.1). In the two-player settings the share of zero expenditures is lower (3.9%) and stable over time. Yet, most of them, especially in later periods, are not a myopic best-response to previous opponent choices. We refer to these sort of zero expenditures as "drop out". In four-player treatments on average 50% of the zero expenditures are drop outs - a share that is increasing over time.

The contest literature has focused on explanations for overbidding (e.g. bounded rationality (Lim et al., 2014), heterogeneous preferences (Shupp et al., 2013), and utility from winning (Schmitt et al., 2004)) and treats zero expenditures simply as best-repsonse to the well-studied over dissipation. Given the collected evidence from prior studies, this cannot be an exclusive explanation. An alternative motivation, that we explore in this chapter, is that lottery contestants in larger groups choose zero investments with increasing frequency because it is in line with their applied learning rules.

We are aware of a handful of studies that analyze learning in repeated games with stochastic outcomes. They differ from our experimental setting and learning identification strategy in many aspects. Yet, they sup-

Figure 2.1: Fraction of zero expenditures over time from meta-analysis



port the use of simple learning heuristics by decision makers. Gunnthorsdottir and Rapoport, 2006 find that reinforcement learning (Roth and Erev, 1995) explains aggregate efforts in a two-stage group game with an inter-group lottery in the first stage. Reinforcement learning combined with directional learning (Selten and Stoecker, 1986) describe well individuals' behavior of a Tullock contest with group size uncertainty (Boosey et al., 2017). In addition, learning spillovers between share and lottery contests are found in a within-subjects experiment by Masiliūnas, 2019. Even though learning behavior in the share contests has so far only received minor attention in the literature, the expected payoff structure provides a useful benchmark to observe behavior in the standard contest, and allows us to treat them as a special case of the more commonly studied Cournot oligopoly.[2] Evidence on learning from oligopoly experiments suggests that players employ a mix of sophisticated and imitative learning. For example, Bigoni and Fort, 2013, with an application of a modified EWA model to a Cournot game under endogenous information disclosure, find that participants use a mixture of

[2]The payoff structure of the share contest resembles a Cournot oligopoly with iso-elastic inverse demand function plus a constant. See Engel (2007) for a meta-analysis of over 500 experimental studies of oligopoly settings.

reinforcement, imitation and belief learning, with the latter accounting for the major share.[3]

Lastly, learning models have been used to explain behavior in repeated auction experiments. From the bidders' perspective, auctions look similarly stochastic to lotteries, since the value of the prize is usually drawn for each bidder from a random distribution so that bids submitted by rivals appear uncertain. In addition, overbidding is commonly observed in first-price auctions (Filiz-Ozbay and Ozbay, 2007). Reinforcement and observational learning is found to reduce overbidding in first-price common value auctions (Garvin and Kagel, 1994) while directional learning can explain repeated individual bids (Neugebauer and Selten, 2006).

2.3 Experimental Design and Procedures

The experiment was designed based on the software z-tree (Fischbacher, 2007) and carried out at the University of Nottingham.[4] A total number of 140 students, recruited from various field of studies via the online recruiting system ORSEE (Greiner, 2015), acted as participants. No participant took part in multiple sessions or participated previously in experimental contest settings.

With the start of each session, the experimenter walked the participants through the experimental instructions (see Appendix A.3) and privately answered their possible questions. To ensure that information passes only via the intended channels of the contest experiment, other forms of communication between participants were prohibited. Participants were randomly assigned to an anonymous group that did not change over the

[3]Other studies on learning in oligopolies find evidence for imitative learning (Vega-Redondo, 1997) (such as Huck, Normann, and Jorg Oechssler, 1999; Jörg Oechssler et al., 2016; D. Friedman et al., 2015), myopic best-response dynamics (Bigoni, 2010) and reinforcement learning (Jiao and Nax, 2017).

[4]The experiment was administered by my co-author, Francesco Fallucchi. Its underlying characteristics and instructions, apart for the treatment assignment and the handling of information feedback, follow the experimental set-up outlined in Fallucchi, Renner, and Sefton, 2013. The set-up of the experiment is detailed since it adds to the understanding of the subsequent results.

course of the experiment and for which individual identities were kept concealed.

Participants took part in the contest experiment for a total of sixty periods.[5] At the beginning of each period, each participant received an endowment of 1000 points. Then, participants simultaneously decided how many points to use in the prize competition with total worth of 1000 points. Points that were not spent were accumulated to each individual balance. After each period, points that were possibly won in the prize competition were added to the this balance. In the case that no group member decided to participate in the prize competition, the prize was not disbursed. We adopt a 2x2 design where treatments differ in the group size, with three (3) or five (5) contestants, and the contest structure, share (S) or lottery (L). 3S, 5S, 3L and 5L are the treatment notations used throughout the rest of the chapter. We conducted two sessions for each treatment, either with 15 or 20 subjects, resulting in ten independent observations in treatments with three-player groups and eight independent observations in treatments with five-player groups. A summary of the treatments is reported in table 2.2.

After each period, participants were reminded of their own choice and informed about the total expenditures of the other members of the group to which they belong and their own earnings. We opted for this "partial" feedback disclosure to rule out imitative behavior among contestants, although we could not rule out the other simple behavioral rule of imitating the average expenditures by opponents in the previous round.[6] The points that participants earned over the experiment were added and converted into a cash equivalent, which was privately paid out after each session. On average, participants received £9.30 for an hour of participation time. At the end of the experiment, we conducted a questionnaire that encompassed demographic questions and an individual risk

[5]The period length in most reviewed experimental studies does not exceed sixty periods. Exceptions are the studies by D. Friedman et al., 2015 and Jörg Oechssler et al., 2016 of continuous time games for 1200 periods.

[6]We check the fraction of players whose choice imitates average opponents' expenditures of previous rounds in Appendix A.2 table A.2. The fraction of imitation is significantly lower in lottery treatments and not increasing over time.

attitude assessment using a survey measure previously tested for a representative pool of subjects (see Dohmen et al., 2011).[7] The two contest structures share the same expected payoff and, under the assumption of risk neutrality, the same Nash equilibria. Introducing risk aversion could potentially alter theoretical predictions, however the direction and the extend of the effect of risk aversion on contest expenditures remains ambiguous under general conditions (Skaperdas and Gan, 1995; Konrad and Schlesinger, 1997). Also in the experimental literature there seems to exist no consensus regarding the effect of risk attitude on contest expenditures (see for example Shupp et al., 2013; Mago, Sheremeta, et al., 2013). To be able to compare both contest structures, we thus maintain the assumption of risk neutral contestants. The theoretical prediction of symmetric group expenditures under risk neutrality, given by $x^* = NV(N-1)/N^2$, corresponds to $666.\bar{6}$ points for three-player contests and to 800 points for five-player contests.[8] Hence, predicted equilibrium expenditures at individual level are 222 and 160, respectively.

Table 2.2: Experimental 2x2 composition

		Group Size	
		N=3	N=5
Contest	Share	3S 10 groups	5S 8 groups
	Lottery	3L 10 groups	5L 8 groups

[7] Subjects answered on a Likert scale from 1 to 7: "How willing are you to take risks, in general? Unwilling to take risks (1) Fully prepared to risks (7)". We do not find significant differences of risk scores across treatments. Appendix A.1 addresses the relevance of risk attitudes on expenditures in more detail.

[8] See Appendix A.5 for a derivation of the Nash equilibrium.

2.4 Results

We lay out the results in two sub sections. In section 2.4.1 we illustrate the spending- and participation behavior of contestants of all treatments. In section 2.4.2 we illustrate how EWA estimation results differ across treatments and analyze expenditures using Tobit mixed effect regressions. Reported p-values (p) for within-treatment comparisons between the two halves of the experiment rely on Wilcoxon matched-pairs signed-rank tests. For comparisons between treatments we report p-values of two-sided Wilcoxon rank-sum tests, for which the independent observations are formed on the group level.

2.4.1 Group Expenditures and Participation

Result 1. *(a) Average group expenditures in share contests increase significantly with an increase in the group size. (b) Average group expenditures in lottery contests, contrary to predictions, do not significantly increase with an increase in the group size.*

Figure 2.2 shows the average group expenditure patterns of all treatments relative to their predicted theoretical equilibria. In all cases, the initial expenditures lie substantially above the Nash equilibrium predictions. In the share contests, mean expenditures are higher in larger groups and decline quickly to a level close to the equilibrium where they exhibit no noticeable time trend thereafter. This result is in line with previous experimental evidence. Instead, we find the results of the lottery contest surprising: over-expenditures compared to the predicted equilibrium levels are persistent throughout the experiment, but average expenditures do not differ across different group size. Moreover, over the longer horizon, total expenditures are lower in larger groups, sometimes even below the level predicted by the Nash Equilibrium. We report in table 2.3 the average group expenditures for the first, second half, and overall periods and p-values of within- and between-treatment comparisons. In the share contests we find that average total expenditures are significantly higher in 5S than in 3S for all intervals considered

(all $p \leqslant 0.01$). Contrary to the share contests, between-treatment comparisons of lottery contests at group level confirm the pattern observed in figure 2.2, with an overall similar level of expenditures for all intervals considered (all $p \geqslant 0.25$). Hence, we cannot reject the hypothesis that group expenditures of winner-take-all contests under different contest size are the same.[9]

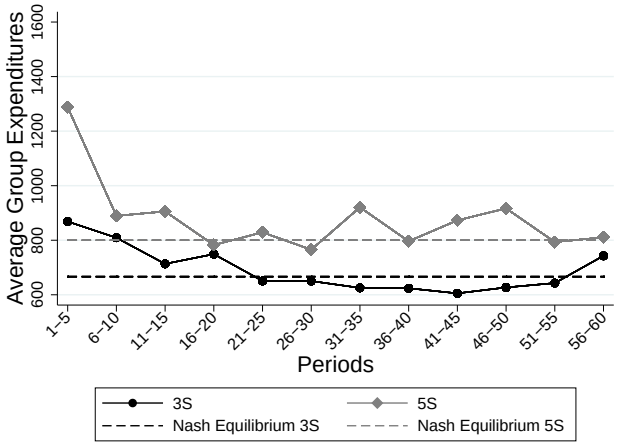
Table 2.3: Average group expenditures, group comparisons for 1st half, 2nd half and all periods between and within treatments.

Average Group Expenditures (Standard Errors)	3S	5S	3S vs 5S p-value	3L	5L	3L vs 5L p-value
All periods	692.41 (11.73)	880.90 (12.67)	0.01	960.80 (21.74)	1055.94 (38.94)	0.53
Period 1-30	740.25 (20.43)	909.99 (13.12)	0.01	1039.05 (27.67)	1243.23 (46.52)	0.25
Period 31-60	644.57 (11.91)	851.81 (16.54)	0.01	882.54 (20.65)	868.65 (35.50)	0.72
1-30 vs. 31-60 p-value	0.24	0.12		0.09	0.01	

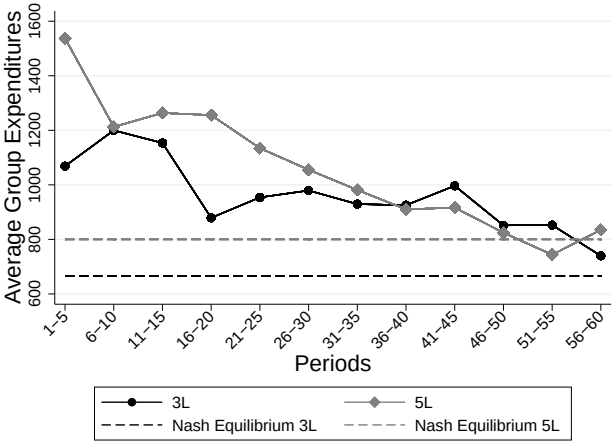
Our finding contradicts theoretical results that predict higher group expenditures for larger groups. Previous studies that explored behavior in contests with different group size have supported the theoretical claim, yet considered only one-shot decisions (Anderson and Stafford, 2003) or

[9]For completeness, we report within-treatment comparison of expenditures in the two halves of the experiment. Results differ across the different contest structures. In both share contests, expenditures are not significantly different in the second half of the experiment compared to early periods (3S: 740 vs. 645 points, $p = 0.24$; 5S: 910 vs. 852 points, $p = 0.12$). In lottery treatments we observe a significant decrease in group expenditures over time, sharper for larger groups, which consequently leads to the similar expenditure levels observed between 3L and 5L (3L: 1039 vs. 883, $p = 0.09$; 5L: 1243 vs. 869, $p = 0.01$).

Figure 2.2: Average group expenditure across treatments over time



(a) Share Contest



(b) Lottery Contest

ten repetitions (Lim et al., 2014).[10] Over a longer time horizon, average group expenditures seem to show different dynamics.

Result 2. *(a) The fraction of zero expenditures is significantly higher in the lottery contests than in the share contests and increases significantly over time for larger lottery groups.*

Our results question the hypothesis that expenditures in the lottery contests converge towards group size dependent equilibria. We thus look for justifications that explain the decrease of expenditures in another common finding from contest experiments: the zero expenditures. To get a first glimpse of the prevalence and dynamics of zero expenditures in contests, we compute the total fraction of zero expenditures across treatments. The findings confirm our initial hypothesis that zero expenditures are more frequent for contest treatments with higher payoff uncertainty. The total share of zeros increases with group size (3S vs. 5s: $p = 0.01$, 3L vs. 5L: $p = 0.01$) and is more pronounced in the lottery contest (3S vs. 3L: $p < 0.01$, 5S vs. 5L: $p < 0.01$) reaching up to 40% in the late game of 5L. Moreover, as shown by the black lines in figure 2.3, the fraction of zero expenditures is stable across time in 3L (periods 1-30 vs. 31-60 $p = 0.92$) and increases in 5L (periods 1-30 vs. 31-60 $p = 0.09$). Result 1(b) can thus be explained by the increasing fraction of zero expenditures in 5L which lead to a faster decrease in average group efforts than in 3L.

An intuitive justification for the pronounced fraction of zero expenditures in the lottery contests is that players expect their opponents to overbid.[11] Since we cannot observe players' expectations on future opponent expenditures, we assume that expectations are formed based on past opponent behavior. Thus, we assess if zero expenditures are a best response (BR) given the history of opponents' decisions using two forms

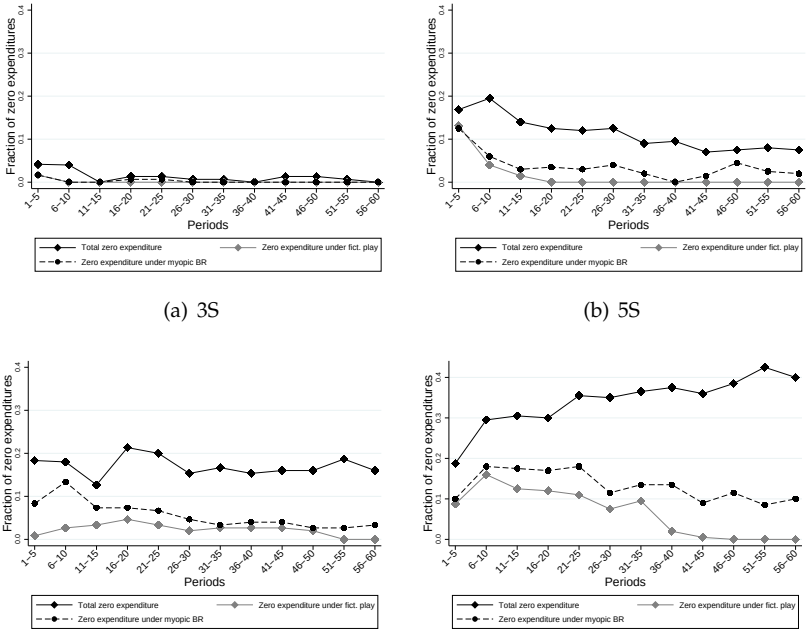
[10] Another experiment, by Morgan et al., 2012, spans over fifty periods. However, the group size formation in their case is endogenous and therefore not comparable to ours where all players are active.

[11] If a player expects that opponents will spend in total 1000 or more, it is individually payoff maximizing to abstain from bidding. A mathematical demonstration can be found in Appendix A.5.

of "weighted-fictitious play". A choice j of player i in period $t+1$ is justified under weighted fictitious play if j maximizes the following expression:

$$\underset{j}{\text{argmax}} \left(\frac{\phi^{t-1}\pi_i\big(s_i^j, s_{-i}(1)\big) + \ldots + \phi^1\pi_i\big(s_i^j, s_{-i}(t-1)\big) + \pi_i\big(s_i^j, s_{-i}(t)\big)}{\phi^{t-1} + \ldots + \phi + 1} \right) \tag{2.3}$$

Figure 2.3: Fraction of zero expenditures across treatments over time



(a) 3S

(b) 5S

(c) 3L

(d) 5L

The parameter ϕ acts as a discount factor. If $\phi = 0$, then the expression reduces to the myopic BR case of $\text{argmax}_j \left(\pi_i(s_i^j, s_{-i}(t))\right)$ which denotes the maximal hypothetical payoff of player i choosing from expenditure levels j given the choices of its opponents s_i at time t (reported as the

dashed lines in figure 2.3). At the other extreme, when $\phi = 1$, all hypothetical past payoffs from strategy j are weighted equally for each period. In this case the best choice is the one that would have resulted in the highest average payoff across all rounds played, also known as "fictitious play" (reported as the gray lines in figure 2.3).

Most zero choices in lottery contests can neither be justified by myopic best-responses nor by fictitious play (average fraction of zeros not justified by myopic best-responses: 66% in 3L, 62% in 5L; by fictitious play: 87% in 3L, 81% in 5L).[12] Yet, myopic best-responses account for more choices than fictitious play, consistent with the previous findings by Rockenbach and Waligora (2016) that lottery contestants hold myopic beliefs. We hence focus on the zero expenditures that are not explained by a myopic best-response and define them as drop outs. In case of a drop out a player decides to spend nothing even if it is payoff maximizing to bid a positive amount based on a myopic best-response. The average fraction of drop out per round can be assessed from figure 2.3 as the difference between the fraction of total zero expenditures and the fraction of zero expenditures under a myopic best-response.

Result 2. *(b) The fraction of drop outs is higher in lottery contests and increases in the five-player lottery.*

Drop outs in share contests are more frequent in larger groups (3S vs. 5S: $p = 0.04$) yet their fraction is relatively low when compared to lottery treatments and stable over time (average fraction of drop out over all choices in 3S: 1%, 5S: 8%, 3L: 11%, 5L: 21%). In the lottery contests the drop out fractions are higher and differ not only in group size but also with respect to the share contests (3S vs. 3L: $p < 0.01$, 5S vs. 5L: $p = 0.01$, 3L vs. 5L: $p = 0.02$). The quota of drop out is highest in 5L and increases significantly over time (periods 1-30 vs. 31-60 $p = 0.04$) while we do not find such an increase in 3L (periods 1-30 vs. 31-60 $p = 0.22$).

[12] The fraction of zeros that would be justified by weighted fictitious play with discount factors different from 0 or 1 lies between the fractions of the extreme cases.

From the previous analysis we deduce that the common decreasing pattern in average expenditures, which we observe in all treatments, may be driven by different behavior, depending on the contest structure. Although in share contests we are far from having the whole contestants managing to achieve the equilibrium level, the decrease in average expenditures hints towards a process of learning to play optimal strategies.[13] Conversely, the decrease of expenditures in the winner-take-all contests can largely be attributed to an overall drop out effect, that is not explained by forms of fictitious play.[14] We hypothesize that the differences in the contests' payoff structures have non-negligible effects on the learning process. The results from the EWA learning model in the next section further explore this thought.

2.4.2 EWA Model Estimation and Interpretation

In the previous section we have shown that the structure and group size of the contest affect total expenditures and expenditure dynamics. Differences in expenditures over time suggest that individual behavior is driven by different learning paths across contests. To test for aggregate learning differences between treatments, we estimate for each treatment the EWA model (C. Camerer and T. H. Ho, 1999). In our estimations, we group the 1001 expenditure choices into $K = 11$ bins of equal distance and round all choices to the closest bin to facilitate comparability.

Every contestant i forms each period t a set of j "attractions" denoted by $A_i^j(t)$, which get recursively reinforced or weakened after every round. Attractions are updated as follows:

[13] The heterogeneity of plays are in line with the previous findings in oligopoly experiments: e.g. Rassenti et al. (2000).

[14] In addition, we evaluate a simple reinforcement rule that assumes participants choose a lower effort than in the previous period if their payoff was negative and the same effort otherwise. Figure A.1 in Appendix A.2 shows that the fraction of zero expenditures are quite well captured by this decision criterion, which motivates the use of the EWA model that combines weighted fictitious play with reinforcement learning.

$$A_i^j(t) = \frac{\phi \cdot N(t-1) \cdot A_i^j(t-1)}{N(t)} + \frac{\delta \cdot E[\pi(s_i^{-j}(t), s_{-i}(t))]}{N(t)} + \frac{(1-\delta) \cdot \pi(s_i^j(t), s_{-i}(t))}{N(t)} \tag{2.4}$$

where $s_i^j(t)$ refers to the actual strategy j chosen by player i in period t while $s_i^{-j}(t)$ denotes all the possible strategies in the same period. Defining $s_{-i}(t)$ as the strategy vector chosen by all other players, the payoff of player i choosing j in t is given by $\pi_i(s_i^j(t), s_{-i}(t))$. Similarly, $E[\pi_i(s_i^{-j}(t), s_{-i}(t))]$ denotes the hypothetical payoff of player i that would have been expected for any possible strategy j given the strategies of all other players. Since the prize assignment in the share contest is deterministic, this expression simplifies to $\pi_i(s_i^{-j}(t), s_{-i}(t))$ and is equivalent to the expected hypothetical payoff of the lottery contest given risk neutrality. $N(t)$ is a weight on past experience. The faster $N(t)$ increases in t, the less players focus on immediate current payoffs at time t. The weights update using the following rule, where the parameter κ determines the growth rate of attractions, which reflects how quickly players lock into a strategy:

$$N(t) = (1-\kappa) \cdot \phi \cdot N(t-1) + 1, \qquad t \geq 1. \tag{2.5}$$

The current attractions of the array of possible strategies J determine the probability of player i choosing strategy j in the next period $t+1$. A logistic transformation links previous attractions to the choice probabilities (equation 2.6). Thus, the higher a contestant's past attraction for a specific strategy, the higher the probability that this strategy will be pursued in the next round.

$$P_i^j(t+1) = \frac{e^{\lambda \cdot A_i^j(t)}}{\sum_{k=1}^{J} e^{\lambda \cdot A_i^k(t)}} \tag{2.6}$$

Table 2.4: Description of estimated EWA parameters

Parameter, Domain	Description
$\delta \in [0, 1]$	Weight of hypothetical payoffs. The higher δ, the more (less) weight is assigned to own hypothetical (realized) payoffs. $\delta = 0$ ($\delta = 1$) corresponds to pure reinforcement (belief) learning.
$N0 \in \left[0, \frac{1}{1-(1-\kappa)\phi}\right]$	Weight of pre-game attractions. Indicates how many periods of experience are required to offset pre-game attractions. The boundaries ensure that the weights of $N(t)$ are increasing in t.
$\kappa \in [0, 1]$	Growth property of attractions. If $\kappa = 0$, then current attractions are a weighted average of past attractions and past payoffs. If $\kappa = 1$, attractions accumulate and can grow larger than present payoffs.
$\phi \in [0, 1]$	Decay rate of attractions from past rounds. The smaller ϕ, the faster are past attractions discounted and the more weight is assigned to attractions of the present round.
$\lambda \in [0, \infty)$	Sensitivity measure of attractions. The higher λ, the more do present attractions matter to determine choice probabilities of future actions. $\lambda = 0$ implies that choice probabilities are not influenced by attractions.

Attractions for each strategy are updated via weighting the three summands in equation 2.4: first, the previous experience discounted by factor ϕ, second the current forgone payoffs, and third the current received payoff. Current forgone payoffs are the payoffs that could have been expected if the contestant had chosen differently by keeping the opponents' strategies fixed. Formation of attractions via evaluating forgone payoffs is equivalent to belief learning (a version of weighted fictitious play, Brown, 1951) whereas focusing exclusively on realized payoffs (i.e. when delta (δ) is zero) reduces the model to reinforcement learning (Roth and Erev, 1995). Thus, the EWA model incorporates two canonical learning models via the parameter delta.[15]

We show in table 2.5 for each treatment the simulation results of the EWA expenditures distribution using varying δ (0, 0.5, 1) and our observed expenditures frequencies. As δ increases towards belief learning the simulated choices show less variation and roughly resemble the true expenditures frequencies in the share contests. The true distribution of expenditures in the lottery contest is more disperse which is why we assume that lottery contestants rely less on belief learning than share contestants.

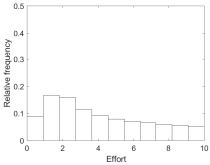
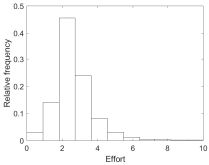
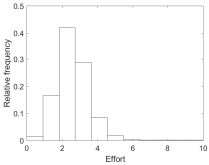
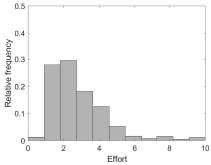
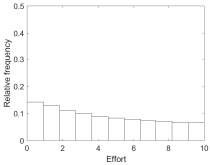
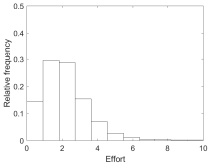
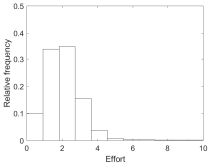
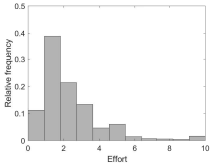
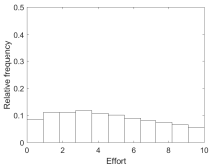
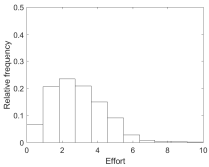
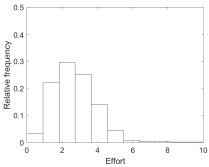
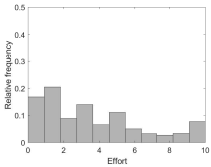
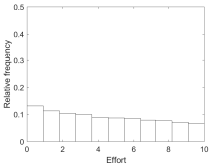
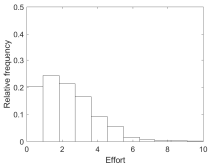
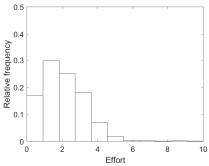
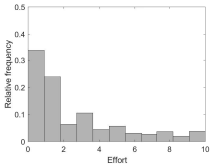
Result 3. *Share contests allow for a mixture of reinforcement and belief learning. In winner-take-all contests, learning is mostly driven by previous own payoffs.*

For each treatment the EWA model is estimated over the first half and the complete sample via a Maximum Likelihood procedure.[16] Estimated parameters and their theoretical domains are summarized in table 2.4. We refrain from freely estimating initial attractions by following the ap-

[15] We use a parameterized version of EWA since it allows us to explicitly estimate δ for each treatment. Whereas a single factor EWA (T.-H. Ho et al., 2001) is suitable to model average contest choices as in Parco et al., 2005.

[16] In the EWA model we abstain from estimating separately the second half of the game since the estimation of the second half only would require that initial attractions of period 31 reflect the complete knowledge formed in the first half of the game and thus possibly results in information loss.

Table 2.5: Simulated EWA expenditures distribution and observed expenditures distribution

	$\delta = 0$	$\delta = 0.5$	$\delta = 1$	Observed
3S				
5S				
3L				
5L				

The first three columns show the relative frequency of expenditures for the EWA simulations, the forth column shows the true expenditures distribution. For each setting 600 players were simulated and experimental specifications were kept unchanged. Each treatment is simulated for $\delta = 0$ (reinforcement learning), $\delta = 0.5$ (equal mix between reinforcement- and belief learning), and $\delta = 1$ (belief learning). All other EWA parameters were kept constant at moderate levels for each simulation ($\lambda = 1$, $\phi = 0.8$, $N0 = 0$, $\kappa = 0.8$, initial attractions=0).

proach used in T. H. Ho et al., 2008 and choose the initial attractions to maximize the likelihood of observing first period choice frequencies.[17]

We report in table 2.6 the estimated parameters with their clustered standard errors and confidence intervals in parentheses.[18] The main parameter of interest in our estimation analysis is δ, that indicates the degree of belief learning used in the game. In the share contests δ are significantly greater than zero, between 0.59 and 0.67, suggesting that players adopt a mixture between reinforcement learning (considering own realized payoffs) and belief learning (considering all own hypothetical payoffs). On the contrary, players in lottery contests mostly rely on reinforcement (experiential) learning, as δ are significantly closer to zero. This is especially true for the early stage of the game (0.02 for 3L and 0.13 for 5L). As players get more familiar with the game, they shift slightly towards belief learning in all treatments. Nevertheless, the difference between share and lottery contests remains substantial.[19] One can argue that belief learning is more complex, since it requires to evaluate for each possible strategy the expected payoffs given the opponents' set of strategies. The evaluation of hypothetical scenarios could be more difficult in lottery contests due to the discrepancy between expected and realized payoffs. It has been shown that individuals lock-in expenditures at inefficient levels in low δ scenarios (Pangallo, Sanders, et al., 2017).

[17] The traditional EWA model considers only choices that are shaped by past attractions while later versions add sophistication to allow for choices based on expectations of opponents behavior (e.g. C. F. Camerer et al., 2002). A sophisticated player forms a best-response from forecasting the actions of all other players. Since including additional free EWA parameters increases the danger of overfitting the model, we instead check for sophistication via the Quantal Response Equilibrium (QRE) model of McKelvey and Palfrey, 1995. In the QRE model, sophistication is captured via the precision parameter λ^Q. The higher λ^Q, the more sophisticated are the overall choices. We find significantly lower λ^Q for lottery treatments. The estimation results of the QRE model can be found in Appendix A.4.

[18] Standard errors are clustered at individual level. Group level clustering is not necessary, due to random assignment of individuals to groups (see Abadie et al., 2017). A set of less conservative estimations without clustering adjustments at individual level is proposed in Appendix A.2 table A.3.

[19] To rule out the possibility that the reported delta estimates are due to our specific EWA estimation set up, we run three additional EWA specifications in which we limit the number of freely estimated parameters. The first specification assumes $N0 = 0$, the second adds that all initial attractions are zero, and the third additionally sets $\kappa = 0$. For all three specifications we obtain delta estimates similar to our reported results. See Appendix A.2 table A.4.

Table 2.6: EWA estimation results across treatments

Treatment	3S		5S		3L		5L	
Periods	1-30	All	1-30	All	1-30	All	1-30	All
$N0$	0.26**	0.19	0.14	0.14	0.53***	0.18***	0.40*	0.08
	(0.11)	(0.12)	(0.11)	(0.12)	(0.12)	(0.07)	(0.23)	(0.12)
	[0.10 0.53]	[0.05 0.53]	[0.03 0.50]	[0.02 0.53]	[0.30 0.75]	[0.08 0.35]	[0.10 0.81]	[0.00 0.68]
κ	1.00***	0.95	0.55***	0.58***	1.00***	0.44***	1.00***	0.64
	(0.00)	(0.29)	(0.16)	(0.17)	(0.00)	(0.15)	(0.00)	(0.52)
	[1.00 1.00]	[0.00 1.00]	[0.25 0.81]	[0.27 0.84]	[0.36 1.00]	[0.19 0.72]	[1.00 1.00]	[0.02 0.99]
ϕ	0.73***	0.75***	0.75***	0.75***	0.71***	0.79***	0.79***	0.82***
	(0.04)	(0.05)	(0.04)	(0.03)	(0.05)	(0.03)	(0.03)	(0.02)
	[0.64 0.81]	[0.63 0.84]	[0.66 0.83]	[0.68 0.80]	[0.60 0.80]	[0.72 0.85]	[0.73 0.83]	[0.77 0.86]
λ	0.80***	0.77***	0.70***	0.69***	0.88***	0.83***	0.88***	0.86***
	(0.04)	(0.03)	(0.03)	(0.03)	(0.02)	(0.03)	(0.01)	(0.07)
	[0.73 0.87]	[0.72 0.84]	[0.64 0.76]	[0.63 0.76]	[0.85 0.92]	[0.77 0.88]	[0.86 0.91]	[0.72 1.01]
δ	0.61***	0.67***	0.59***	0.62***	0.02	0.11	0.13*	0.18***
	(0.07)	(0.05)	(0.04)	(0.03)	(0.11)	(0.08)	(0.07)	(0.07)
	[0.47 0.73]	[0.57 0.75]	[0.51 0.66]	[0.55 0.69]	[0.00 1.00]	[0.02 0.39]	[0.04 0.35]	[0.07 0.35]
$Obs.$	900	1800	1200	2400	900	1800	1200	2400
$-LL$	1494.0	2666.6	1766.6	3131.6	1444.3	2768.7	1824.0	3210.6
AIC	2998.0	5343.3	3543.2	6273.2	2898.6	5547.3	3658.0	6431.1
BIC	3022.0	5370.8	3568.6	6302.2	2922.6	5574.8	3683.5	6460.1

EWA Maximum Likelihood estimation with restricted parameters reported. Standard errors of the restricted metric, in parenthesis, are clustered at individual level; 95% confidence interval of restricted estimates in brackets. P-values ($H_0 = 0$): *$\leqslant$ 0.10, **$\leqslant$ 0.05, ***$\leqslant$ 0.01, Obs. is the number of observations, -LL is the negative log-pseudolikelihood, AIC and BIC are the Aikaike- and Bayesian information criterion, respectively.

Other parameter estimates are similar across various estimations. We find $N0$ below one in all estimations, which indicates that pre-game attractions are offset completely by first period attractions.[20] Kappa (κ), which measures the growth property of attractions, is significantly different from zero in the first half of the game, indicating that the importance of past attractions grows over time. The decay rates of past attractions indicated by phi (ϕ) are significantly different from zero, but similar across and within treatments, on average between 75% in the share contest and 81% in the lottery contest. Although the difference is negligible, this indicates that lottery contestants may rely more on their past experience. This result is consistent with the idea that present payoffs in lottery contests reveal less useful information. Finally, lambda (λ) is significantly different from zero for all specifications, indicating that contestants' choices are influenced by attractions formed in past rounds.

The EWA parameter estimation uncovers a noticeable difference in learning between the share and lottery contests. A strong reliance on own realized payoffs by subjects in lottery contests conveys that current decisions depend on the success of the previous ones. Victories reinforce the probability of playing the corresponding level of expenditure, while losses make the corresponding expenditure level less attractive in forthcoming periods. If a subject experiences frequent losses, positive expenditures levels will, over time, become less appealing to the advantage of zero expenditures. Following this thought, the fraction of zero expenditures should be higher in contests with a higher share of non-winners. This reconciles with our results in section 2.4.1 that show that the fraction of drop outs is significantly higher in lottery treatments than in share treatments and increases significantly over time in larger groups. If lottery contestants rely indeed on previous own experiences, then we should observe a drop of expenditures after a series of losses.

[20] An influence of pre-game attractions is not expected from the experimental design, since none of the participants was previously familiar with the contest setting.

Result 4. *(a) Lottery contestants significantly decrease expenditures after the accumulation of losses. (b) The effect is more pronounced for bigger groups.*

We expect a negative relationship between the length of the series of losses prior to time t and the expenditure level at time t. We assess this relationship, using a set of Tobit mixed effect models.[21] The model assumes that expenditures levels are left censored at 0 with random effects at individual and group level. In all models we regress the expenditures at t on previous own expenditures, previous opponents' expenditures (linear and squared) and a variable capturing the time trend.[22] We check the relationship between prior losses and current expenditures via the variable *loss streak* defined as the accumulated, (negative) payoff from consecutive losses prior to time t relative to contestants' endowment. After each incurred loss, the variable decreases by the foregone profits that would have been received by choosing to not spend the endowment. Consequently, a *loss streak* remains unchanged for zero expenditures. If the contest has been won in the previous period, the contestant's *loss streak* is reset to zero.

Model (1), for 3L, and (4), for 5L, in table 2.7 contain only control variables. In both lottery treatments the influence of prior expenditures on current choices is positive and significant, while the effect of period advancement is negative and significant. In models (2) and (5) we find a positive and significant effect of the *loss streak*, indicating that the accumulation of losses indeed leads to a decrease in expenditures. The effect is more significant in 5L for which individual loss accumulation is on av-

[21] It might seem suitable to use a hurdle model (Moffatt, 2015), as individuals first choose whether to participate and subsequently decide their expenditure level. However, its requirements for panel data are too rigid given our setting. In particular, hurdle models classify a subject as "zero contributor" if own expenditures remain zero over all periods. Since we do not observe "zero contributors" per se, but rather an increasing tendency of contestants to choose zero expenditures over time, we opt for the more flexible Tobit mixed effect model.

[22] The use of lagged expenditures as explanatory variable requires stationary panels, which we test for using the Levin-Lin-Chu unit root test for panel data (Levin et al., 2002). We reject the hypothesis that lottery expenditures follow a unit root process ($p<0.00$). We include the square term of other expenditures in $t-1$ since the myopic best response function is concave and thus better approximated by a quadratic polynomial.

erage higher given the lower unconditional probability to win the contest. We further analyze the *loss streak* effect in (3) and (6), where we additionally control for the contestants' gender. As an exploratory result, we find that women spend on average more than men.[23] From (3) we find that splitting the loss streak variable with respect to gender does not lead to significant effects on expenditures. Yet when increasing the group size (6), the accumulation of prior losses significantly decreases the expenditures for both genders, for women more sharply than for men. This last result has been similarly observed in experimental tournaments (Buser, 2016).

The regression results support the claim that decreasing expenditures in the lottery contest are driven by previous lottery outcomes. With the cumulation of losses, contestants tend to lower their expenditures and may, over time, drop out of the contest.[24] Since individual losses accumulate longer when facing more opponents, the decrease of expenditures is more pronounced in 5L.

2.5 Final Discussion

Our contribution to the literature is two-pronged. First, we offer a clean comparison of how subjects behave across different contest structures. The witnessed behavioral discrepancy between contest types might be connected to the probabilistic prize assignment in the lottery that influences how contestants form their choices. We hypothesize that share and lottery contestants use distinct learning strategies that may also explain another expenditure peculiarity, that tends to be overlooked: the modal choice in lottery contests is often zero.

We find that, in discord with theoretical predictions, group size does not affect total investments in lottery contests. The decrease of investments

[23] In line with findings from prior contests experiments such as Price and Sheremeta, 2015 and Brookins and Ryvkin, 2014.

[24] To analyze the effect of loss cumulation on the likelihood of zero expenditures we report a random effects logit regression in Appendix B table A.5. We find that a prior loss streak significantly increases the probability of zero expenditures in 5L. For 3L we find that the effect goes in the same direction but is not significant, which might be due to the lower loss cumulation in 3L.

Table 2.7: Tobit mixed effects regression on lottery expenditures.

Dependent variable:	3L			5L		
Expenditures_t	(1)	(2)	(3)	(4)	(5)	(6)
$\text{Expenditures}_{t-1}$	0.40***	0.40***	0.40***	0.20***	0.16***	0.116***
	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
Other	0.03	0.05	0.06	-0.03	0.02	0.02
$\text{Expenditures}_{t-1}$	(0.05)	(0.05)	(0.05)	(0.04)	(0.04)	(0.04)
Other	-0.00	-0.00	-0.00	0.00	-0.00	-0.00
$\text{Expenditures}^2_{t-1}$	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Period	-1.41***	-1.37***	-1.35***	-3.71***	-3.43***	-3.36***
	(0.42)	(0.42)	(0.42)	(0.46)	(0.45)	(0.45)
Loss streak		0.02*	0.00		0.07***	0.04*
		(0.01)	(0.02)		(0.01)	(0.02)
Female			85.00*			83.26
			(49.57)			(85.71)
Female×Loss streak			0.03			0.05**
			(0.02)			(0.02)
Constant	197.48***	199.64***	145.54***	192.06***	223.29***	161.54**
	(34.14)	(34.44)	(45.50)	(45.30)	(46.02)	(77.05)
Obs.	1770	1770	1770	2360	2360	2360

Multilevel Tobit mixed effects model; Standard error in parenthesis. P-values: *$\leqslant 0.10$, **$\leqslant 0.05$, ***$\leqslant 0.01$, Obs. is the number of observations.

in large-group lottery contests is influenced by an increasing fraction of zero expenditures. A substantial number of these zero expenditures is not justified by myopic best-responses (or other forms of fictitious play) and defined by us as "drop out". Even though the drop out rate is lower for share contests, the average expenditures converge to theoretical predictions which suggests that spending behavior across contests is formed by distinct learning processes.

A parameter estimation of the EWA model in all treatments indicates that lottery contestants decide mostly based on the information gathered from their own realized payoffs. Since success in the lottery contest is stochastic, subjects who base their investment decision entirely on

their previous decisions are less able to adapt their strategies in a payoff-optimizing fashion. On the contrary, participants in the share contests rely on a mixture of own realized payoffs as well as foregone payoffs. This may be facilitated by the deterministic nature of the payoffs in the share contest. The distinct learning patterns estimated in the two contest environments do not significantly change over time and are robust to changes in the number of players.

Repeated losses that subjects face in the lottery contests decrease over time the reinforcement of positive expenditure levels and consequentially zero expenditures, irregardless of being a myopic best-response, become more appealing. Our regression results add to this thought by showing that the cumulation of prior losses leads to a significant decrease in expenditures which is more pronounced in bigger groups. The higher drop out rate in the five-player lottery is presumably driven by the faster accumulation of individual losses. As a consequence, an increase in the group size does not necessarily increase total rent-seeking.

In a society full of embedded lottery contests our results obtain practical relevance. First, it may be beneficial in rent-seeking situations, such as lobbying, that an increase in contestants does not significantly change total rent-seeking effort. Contrary, in case a high sum of efforts is favored, such as in a philanthropic fund-raising lottery, increasing the pool of participants might not lead to the desired effect if the individual loss probability is increasing simultaneously. Second, decisions not to invest can be aggregated on a macro-level to the so called "industry shakeout", i.e. a significant reduction in the number of active firms during the expansion of new industries (see Gort and Klepper, 1982; Klepper, 1996; Klepper, 1997). One traditional explanation for firm exit dynamics postulates that market participants use Bayesian updating to learn their true ability of operating in the market (Jovanovic, 1982; Jovanovic and Nyarko, 1995). Thus, firms decide to exit assessing their past performance. Our finding, that participants invest less after losses and potentially drop out of the contest due to non-reinforcement of positive payoffs, takes a similar line. Third, similar to firms that misallocate R&D efforts in environments with random outcomes (Ahuja and Novelli, 2017), we stress the diffi-

culty of players to form rewarding learning strategies in highly uncertain domains. A well-known domain that is characterized by high outcome uncertainty is the drug discovery in the pharmaceutical industry, which has been conceptualized as la lottery (Sutton, 1998). Understanding how individuals form decisions may help to better address challenges that companies in such environments face.

The presented work calls for a better understanding on how contestants learn in highly uncertain environments, such as winner-take-all contests with many participants, and how their decision-making abilities can be improved. In addition, the observed drop out effect and its relationship to group size and possibly other contest characteristics deserves increased attention to better bridge the gap between experimental findings and theoretical explanations of contestants' behavior.

Chapter 3

Identifying Types in Contest Experiments

3.1 Introduction

The characterization of heterogeneous behavior has a long tradition across economics disciplines. Various approaches, from revealed preferences to newly developed econometric models, have been used to define subjects' behavior in parsimonious and yet tractable ways. These methods receive notable attention by behavioral economists, who have a long tradition in designing experimental paradigms to pursue this scope (e.g. Houser et al., 2004). One eminent example of aforesaid research is the typology in public good settings proposed by Fischbacher et al., 2001 who use a variant of the strategy method (Selten, 1965). If we look at competitive settings, however, little progress has been made in defining a meaningful typology.

An early suggestion comes from Potters et al., 1998 in the experimental analysis of the rent-seeking contest (Gordon Tullock, 1967) . The authors acknowledge that, from the remarks left by participants, it is possible to classify three types of players: the "gamesmen", who seem to understand the strategic nature of the game; the "adapters", who adapt to the

outcomes of the previous rounds; and the "confused", who randomize effort. More recently, Herrmann and Orzen, 2008 find support for the existence of different types in a variant of the Tullock contest using the strategy method.

Modeling heterogeneity in competitive situations such as contests is an important step to understand conflict resolution, as it facilitates further research on the different motives that drive the behavior of individuals and aid comprehension of group dynamics caused by different type compositions. In this chapter we propose a classification of types in the Tullock contest using the classifier-Lasso (C-Lasso, Su et al., 2016), which identifies and estimates latent group structures in panel data.

We apply the C-Lasso to a data set of six repeated contest experiments that differ in whether the groups of contestants stay fixed or are randomly re-matched after each round. Other studies have highlighted differences in average behavior under certain circumstances between fixed and randomly rematched contestants (e.g., Baik et al., 2015; Fallucchi and Renner, 2016). Instead of focusing on the overall level of rent-seeking, we look at the strategies that players adopt.

Results from our estimations suggest that the optimal number of behavioral types in both contest regimes is three, in line with what has been previously suggested. We label more than half of contestants in fixed matching as *reciprocators* since they show an increasing response function with respect to the effort based on the opponents' previous choices. A second group, only formed by a tenth of players, adapts their effort to the past opponents' choices in a concave fashion (*gamesmen*). For the remaining third, the *others*, opponent effort on current choices seems to matter in a non-standard way.

In experiments with random matching we also find *reciprocators*. However they represent a significantly lower fraction of players (around 25%). These findings are coherent with the experimental evidence in other settings that fixed matching protocol induces a higher orientation to reciprocity (e.g. Schmidt et al., 2001). Interestingly,in treatments with a fixed matching protocal we find that groups with more *reciprocators* are associ-

ated to lower average group efforts hinting at the collaborative nature of this type of contestant.

The remainder of the chapter is structured as follows. The next section offers a short review on the literature of player type identification in experimental settings. In section 3.3 we describe the contest data and present the C-Lasso mechanism. In section 3.4 we show the results of the C-Lasso estimations for both contest with fixed groups and randomly re-matched groups. We conclude in section 3.5 with a discussion of our results.

3.2 Type Identification in the Experimental Economics Literature

Across many economic situations we observe heterogeneity in individual behavior. In laboratory settings, behavioral heterogeneity should be largely due to personal differences in information-processing and decision-making. Game theoretical models have long built on these individual differences, as for instance the Level-k (Stahl, 1993; Stahl II and Wilson, 1994) or cognitive hierarchy models (C. F. Camerer et al., 2004). In this section we review different approaches used in the experimental economics literature to group individuals given their observed differences and thus to identify "what types are there?" (Cosaert, 2019).

One simple way to categorize individuals is by visual examination of their characteristics. For example, Fischbacher et al., 2001 classify participants in a public good experiment based on how their contribution choice depends on (hypothetical) choices of other players. In another public good experiment, Kurzban and Houser, 2005 regress the individual contributions of each participant on the average observed contribution of group members in the last round, and develop a player taxonomy based on the differences in individual regression coefficients. A similar classification approach has been used in a tragedy of the commons experiment by Casari and Plott, 2003.

More sophisticated approaches to form taxonomies in experimental settings frequently use Bayesian estimation of type-specific decision rules or

finite mixture models. The former models apply Bayesian updating rules to estimate the probability of a participant sharing a common set decision rules. For example Houser et al., 2004 analyze behavior in an optimization task, in which participants choose successively between two options whose payoffs are stochastic and depend on the trajectory of own prior choices. Participant types are identified via the posterior probability to apply a certain set of decision rules. Different to early work on type identification using Bayesian updating (El-Gamal and Grether, 1995), the number and form of decision rules is not a-priori given but derived using a common set of candidate variables that potentially affect future choices. Due to computational reasons the procedure is well-suited to identify types in situations for which the candidate set can be restricted. The number of latent types is determined by estimating different models and comparing their fit to the experimental data. Other exemplary adaptations of Bayesian estimation for type identification are used in auction settings (Shachat and Wei, 2012) and normal form games under time constraints (Spiliopoulos et al., 2018).

The second group of models, the finite mixture models, assume that observed data is aggregated from distinct sub distributions, one for each type (Fraley and Raftery, 2002). The number of types and their composition can be evaluated using commonly used model selection techniques. The use of finite mixture models for categorizing individuals has become standard in repeated experiments. For example, Polonio et al., 2015 use finite mixture models to cluster participants in two-player normal-form games based on their eye movement patterns. Other applications that involve finite mixture models are found in public good games (Bardsley and Moffatt, 2007), beauty contests (Bosch-Domènech et al., 2010), lottery choices (Conte et al., 2011), and private information games (Brocas et al., 2014).

In this work, we propose with C-Lasso (Su et al., 2016) an alternative approach to the traditional ones, that does not only cluster individuals but simultaneously estimates their type specific characteristics. C-Lasso has recently been applied for panel data estimations across economic settings (Lu and Su, 2017; Wang et al., 2019). The method shrinks the set of

individual regression coefficients to a smaller set of group-specific coefficients and assigns each participant to one of the identified groups. From an econometric point of view, C-Lasso avoids a series of drawbacks of mixture models. In particular, the likelihood function of a finite mixture model can show irregularities such as multimodality (Lehmann and Casella, 2006; Spiliopoulos et al., 2018), and therefore introduces a complication in detecting the local maximum point that corresponds to the efficient root.[1] On the contrary, the application of the penalized likelihood maximization of C-Lasso produces a unique solution for any choice of parameters.

3.3 Data and Methods

Tullock contests are frequently used to model competitive situations under uncertainty (Konrad, 2009) and form an essential part of the experimental economics literature (Dechenaux et al., 2015). In a standard specification of the contest N players compete for a prize of size P, whose assignment is probabilistic. The chance of contestant i to win the prize increases with his own efforts x_i but decreases with the total sum of efforts over all contestants $\sum_{i=1}^{N} x_i$. The individual profit π_i depends thus on all contestants' expenditures, the prize assignment and a homogeneous initial endowment e:

$$\pi_i = \begin{cases} e - x_i + P & \text{with probability} \quad x_i / \sum_{i=1}^{N} x_i \\ e - x_i & \text{otherwise.} \end{cases} \tag{3.1}$$

Our dataset consists of six experimental treatments.[2] In two of them (Savikhin and Sheremeta, 2013; Mago, Samak, et al., 2016) contestants repeatedly play the contest against the same three opponents. In the remaining four treatments (Sheremeta, 2010; Sheremeta and J. Zhang, 2010;

[1] Alternative approaches to deal with these problems have been proposed in the literature without a general consensus (e.g. McLachlan and Peel, 2004, Mercatanti, 2013 and Feller et al., 2018).

[2] We warmly thank the authors for providing us with their experimental data

Price and Sheremeta, 2011; Chowdhury, Sheremeta, et al., 2014) contestants are randomly regrouped with three opponents after each round. We refer to the two treatment types as fixed matching (FM) and random matching (RM), respectively.[3] For each round contestants anonymously compete against their opponents by deciding their effort level that can take values between zero and the value of the prize P.[4] After each round, they are reminded of their own effort and receive information on their opponents' total effort and whether they have won the prize. At the end of the experiment the accumulated earnings are converted and paid out in cash.

Consistent with evidence from other experimental contests (Dechenaux et al., 2015), the sample shows substantial variation in efforts. Part of the observed heterogeneity may be explained by divergent contestants' reactions to the information provided after each round of the contest. For example, prior effort of opponents could lead to imitative behavior (e.g., Apesteguia et al., 2007) or to adapt own efforts in line with the (myopic) best response function of the contest. Further, how information of the previous round affects effort decisions might depend on whether opponents stay the same or potentially change after each round. We thus are interested in how contestants can be characterized by their reactions to prior information within FM and RM treatments and how contestants differ between fixed and random matching regimes.

We analyze the data of FM treatments, $N = 96$ contestants observed for $T = 19$ periods, and of RM treatments, $N = 183$ contestants observed for $T = 29$, into two panel data sets ($\boldsymbol{\omega}_{it} = (y_{it}, \boldsymbol{x}_{it})$).[5] The variables of interest are y_{it}, the effort of contestant i at time t, and the vector of

[3] For all treatments, different subjects are recruited from a pool of students (Chowdhury, Sheremeta, et al. (2014) recruits from University of East Anglia, the other studies from Purdue University). We test for differences in average group efforts within the experiments of both FM and RM treatments using Kruskal-Wallis rank tests but do not detect significant deviations (p-value FM: 0.222, p-value RM: 0.280).

[4] In the course of the empirical analysis, we normalize the range of possible effort values to $[0; 1]$ for all experimental treatments to facilitate the comparison of results.

[5] From the original data set we drop the observations of four contestants in FM treatments and nine contestants in RM treatments, whose prize assignments are time invariant and therefore not possible to assess with C-Lasso. We also omit the first period due to lag effects in the model.

covariates $\boldsymbol{x}_{it}$: the sum of normalized expenditures of the contestant's opponents at $t-1$ (*l.othereffort*), the squared sum of normalized expenditures of the contestant's opponents at $t-1$ (*l.othereffort*^{2}), an indicator of the contest outcome for contestant i at $t-1$ (*l.win*) and a time indicator (*period*). We include *l.othereffort*^{2}, since the (myopic) best response function of the contest is concave and thus better approximated by including a square term in the regression ($BR(X) = \sqrt{X} - X$, where X is the sum of opponent expenditures, see Appendix A.5 for a formal derivation).[6] Given the censored nature of the outcome, a standard Tobit panel with fixed effects can plausibly be assumed to model the contests:

$$y_{it} = max(0, \mu_i + \boldsymbol{x}_{it}\boldsymbol{\beta}_i + \epsilon_{it}) \qquad \text{with} \quad \epsilon_{it} \sim N(0, \sigma_\epsilon^2), \tag{3.2}$$

where μ_i is the individual fixed effect; $\boldsymbol{\beta_i}$ the vector of individual coefficients; ϵ_{it} the idiosyncratic error term. We use a Tobit model with lower bound, since zero expenditures are frequent whereas maximum expenditures are never optimal and therefore rare choices. We estimate the optimal number of behavioral types using C-Lasso. Since we are not aware of other studies that have adapted the use of C-Lasso in combination with a Tobit model, we provide a step-by-step description of our estimation approach in the remaining part of this section.[7]

C-Lasso estimates group affiliation via shrinking the set of individual coefficients $\boldsymbol{\beta}_i$ to a smaller set of latent, group-specific coefficients $\boldsymbol{\alpha}_k$ by minimizing the following penalized nonlinear likelihood (PNL) function:

$$\underset{(\boldsymbol{\beta}_i, \mu_i, \boldsymbol{\alpha}_k, \sigma_\epsilon^2)}{min} \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \Psi(\boldsymbol{\omega}_{it}; \mu_i, \boldsymbol{\beta}_i, \sigma_\epsilon^2) + \frac{\lambda}{N} \sum_{i=1}^{N} \prod_{k=1}^{K} ||\boldsymbol{\beta}_i - \boldsymbol{\alpha}_k||. \tag{3.3}$$

[6] In addition, we perform likelihood ratio tests on FM and RM data panels to determine whether to include the squared efforts of opponents in t-1, the efforts of opponents in t-2, and the squared efforts of opponents in t-2 as additional regressors. Only the inclusion of *l.othereffort*^{2} leads to a significant increase in the model's log likelihood. See table B.1 in Appendix B.1.

[7] We warmly thank Zhentao Shi, co-author of the C-Lasso reference paper, for his useful comments regarding our implementation.

N, T, and K denote the number individuals, the number of periods, and the number of groups, respectively. The first term of the PNL equation (3.3), $\Psi(\boldsymbol{\omega}_{it}; \mu_i, \boldsymbol{\beta}_i, \sigma_\epsilon^2)$, denotes the log likelihood function of the Tobit model given the data, the individual fixed effect, the vector of individual coefficients, and the variance of the idiosyncratic error term. We transform the Tobit log likelihood function (Olsen, 1978) to make convex optimization methods feasible:

$$-\Psi(\boldsymbol{\omega}_{it}; \boldsymbol{\beta}_i, \mu_i, \sigma_\epsilon^2) = \sum_{y_{it}>0} \frac{1}{2}\big[ln(2\pi) - ln\theta^2 + (\theta y_{it} - \boldsymbol{x}_{it}'\boldsymbol{\delta}_i - c_{it}'\eta_i)^2\big] + \sum_{y_{it}=0} ln\big[1 - \Phi(\boldsymbol{x}_{it}'\boldsymbol{\delta}_i + c_{it}'\eta_i))\big] \tag{3.4}$$

Once the minimization procedure has been carried out, the original parameters are retrieved by $\sigma_\epsilon = 1/\theta$, $\boldsymbol{\beta}_i = \boldsymbol{\delta}_i/\theta$, and $\mu_i = \eta_i/\theta$.

The second term of the PNL function (3.3) reflects the C-Lasso penalization that shrinks the individual-specific vector of explanatory coefficients $\boldsymbol{\beta}_i$ to a vector of group-specific coefficients $\boldsymbol{\alpha}_k$. Every individual is assigned to a group for which a set of Tobit coefficients is estimated. The estimation of the group coefficients $\boldsymbol{\alpha}_k$ depends on the tuning parameter λ and on the number of groups K. If the number of groups K is not known ex-ante, it has to be estimated using an information criterion (IC) function. We select K and λ according to the IC that minimizes:

$$\underset{(K,\lambda)}{min} \frac{2}{NT} \sum_{k=1}^{K} \sum_{i:\hat{\boldsymbol{\beta}}_i = \hat{\boldsymbol{\alpha}}_k} \sum_{t=1}^{T} \Psi(\boldsymbol{\omega}_{it}; \hat{\boldsymbol{\alpha}}_k, \mu_i, \lambda) + \nu(NT)^{-0.5} qK \tag{3.5}$$

The higher the number of groups K, the more free parameters are contained in the model and thus the smaller the negative function $\Psi(.)$. To counterbalance overfitting, the number of groups is penalized by

$\nu(NT)^{-0.5}qK$, where q is the number of explanatory variables and $\nu \in (0,1)$ a penalization parameter set to $\nu = 0.22$. This parameter is obtained using simulated data for which we fix a-priori the number of groups ($K_{\text{known}} = 1,2,3,4$), and assess the correctly specified number of groups over all K_{known} and a sequence of λ.

We evaluate equation (3.5) for all combinations of $K = 1,...,4$ and c_λ in a geometrically increasing sequence of 10 points and choose the combination (K, c_λ) that minimizes the IC function. The tuning parameter λ is obtained via $\lambda = c_\lambda var(y) T^{(-1/3)}$, where the used sequence of c_λ depends on the underlying econometric model. Su et al., 2016 use a geometrically increasing sequence of ten c_λ from 0.01,...,0.1 for a linear model and from 0.2,...,2 for a probit model. As a Tobit model resembles a combination of these two models, we choose with $0.1 \cdot 10^{(k-1)/9}$ $(k = 1,...,10)$ a geometrically increasing sequence that overlaps with both of the authors' original sequences.

The resulting values for the IC function are reported for the FM and RM panel in table 3.1 and table 3.2. The minimum value of the IC function is equal to 0.021 in the FM panel and equal to 0.120 in the RM panel, corresponding to a number of groups $K = 3$ in both panels. Given the optimal combination of tuning parameter and number of groups, one can estimate equation (3.3) and retrieve the C-Lasso group estimates. Estimates are unbiased using a half-panel jackknife procedure (Dhaene and Jochmans, 2015). Lastly, we do not detect signs of model misspecification using RESET tests (Ramsey, 1969), that have been used in Tobit models (Peters, 2000).[8]

3.4 Results

Confirming the suggestions of Potters et al., 1998, our estimation results show that the optimal number of latent groups in the FM sample is in-

[8]We use RESET specifications that include quadratic and cubic terms of linear predictions, as they are found to have high statistical power in Monte Carlo simulations (E. A. Ramalho and J. J. Ramalho, 2012; Lechner, 1995). P-values for the quadratic and cubic terms in FM (RM) treatments: 0.67, 0.43 (0.48, 0.20). Hence, one cannot reject the Null hypothesis that the model is correctly specified.

Table 3.1: FM panel - values of the information criteria (IC) function for alternative number of groups K, and tuning parameter c_λ selection

Tuning Parameter	K: Number of Groups			
c_λ	1	2	**3**	4
0.100	0.048	0.031	0.038	Inf
0.129	0.048	0.035	0.044	0.038
0.170	0.048	0.034	0.046	0.035
0.215	0.048	0.039	0.043	0.035
0.278	0.048	0.043	0.052	0.033
0.359	0.048	0.042	0.039	0.033
0.464	0.048	0.049	0.037	0.051
0.600	0.048	0.050	**0.021**	0.041
0.774	0.048	0.053	**0.021**	0.027
1.000	0.048	0.055	0.028	0.040

Table 3.1 contains for each (K, c_λ) combination the result of the IC function in equation 3.5 using the FM panel.

Table 3.2: RM panel - values of the information criteria (IC) function for alternative number of groups K, and tuning parameter c_λ selection

Tuning Parameter	K: Number of Groups			
c_λ	1	2	**3**	4
0.100	0.197	0.191	0.155	0.166
0.129	0.197	0.190	0.159	0.178
0.170	0.197	0.194	0.158	0.172
0.215	0.197	0.195	0.153	0.175
0.278	0.197	0.194	0.146	0.179
0.359	0.197	0.193	0.120	0.179
0.464	0.197	0.190	**0.120**	0.179
0.600	0.197	0.187	0.122	0.174
0.774	0.197	0.185	0.132	0.145
1.000	0.197	0.180	0.140	0.155

Table 3.2 contains for each (K, c_λ) combination the result of the IC function in equation 3.5 using the RM panel.

deed three. 58% of the contestants form type 1, 9% fall into type 2 and the remaining contestants account for type 3 (see Table 3.3). We test whether contestants that belong to the least frequent type would be equally well represented in one of the other two types. However, likelihood ratio tests reveal that re-assigning type 2 contestants significantly reduces the model fit.[9] Further, we test whether individual attributes such as experimental study or gender are proportionally distributed over the identified types. We do not find that the distribution of static contestant characteristics is significantly different across C-Lasso groups, suggesting that type assignment rather depends on differences in how contestants react to information received of the past round played.[10] We thus turn to the C-Lasso regression results that characterize the differences in types.

Table 3.3 reports the estimates of a standard Tobit model with pooled observations as well as C-Lasso coefficients for each of the three types. The results of the pooled Tobit regression imply that, over all contestants, the effect of past opponents' effort on own effort is concave given the positive (negative) significant effect of *l.othereffortsl* (*l.othereffort*^{2}). Winning the contest in the previous round (*l.win*) significantly increases the efforts of contestants in the subsequent round. Moreover, we do not find a significant time trend (*period*) when pooling observations. Yet, the pooled estimation results mask the heterogeneous behaviors observed in the three groups identified by C-Lasso.

For all three contestant types, opponent expenditures in the previous round matter for own effort choices, but how they matter varies across types.[11] For the first type, contestants show an increasing response function with respect to *l.othereffortsl*, and further increase (decrease) their ef-

[9] P-values are smaller than 0.01 so that we reject the hypothesis that assigning type two contestants to either one of the other groups does not affect the log likelihood of the model.

[10] P-values of Pearson's Chi squared test on C-Lasso type assignment in FM treatments and other time-invariant characteristics: experimental study: 0.496, gender: 0.358, university major: 0.06, region of origin 0.680. For RM treatments: experimental study: 0.926, gender: 0.551, university major: 0.551, region of origin 0.144.

[11] Across C-Lasso groups, we reject the null hypotheses that the coefficients of *l.othereffort* and *l.othereffort*^{2} are jointly equal to zero ($H_0 : \beta_1 = \beta_2 = 0$) and the hypothesis that the average partial effect of *l.othereffort* on own efforts is zero ($H_0 : \beta_1 + 2\beta_2\overline{l.othereffort} = 0$, where $\overline{l.othereffort}$ is fixed to be the mean value of *l.othereffort*) on the 1% significance level using F-tests.

Table 3.3: C- Lasso Tobit regression results for FM treatments

	Pooled Tobit	C-Lasso Tobit		
Dep. variable: *effort*		Type 1 *Reciprocators*	Type 2 *Gamesmen*	Type 3 *Others*
l.othereffort	0.504***	0.113**	1.000***	-0.282***
	(0.030)	(0.051)	(0.237)	(0.085)
$l.othereffort^2$	-0.151***	-0.012	-0.347***	0.101***
	(0.015)	(0.020)	(0.094)	(0.030)
l.win	0.126***	0.031**	-0.009	-0.004
	(0.016)	(0.014)	(0.041)	(0.026)
period	0.001	0.000	-0.005	-0.003*
	(0.001)	(0.001)	(0.003)	(0.002)
σ_ϵ	0.307***	0.200***	0.263***	0.274***
	(0.005)	(0.006)	(0.012)	(0.009)
Obs;N;%	$1824; 96; 100\%$	$1064; 56; 58\%$	$171; 9; 9\%$	$589; 31; 32\%$

Standard errors (in parenthesis); p-values: *$\leqslant 0.10$, **$\leqslant 0.05$, ***$\leqslant 0.01$; Obs. is the total number of observations; N is the number of contestants of each type; % is the relative share of each type with respect to the full sample.

fort after a win (loss). Being faced with lower (higher) bidding opponents leads these contestants to bid low (high) themselves, suggesting a "good for good and evil for evil" reply to previous opponents' choices. We refer to contestants of type one as *reciprocators*. The second type does not significantly respond to previous contest outcome, but own efforts relate positively to other efforts and negatively to its square term. The inverted u-shape that relates *l.othereffort* to current choices, suggests that these players decrease their effort when competition gets fierce, which is line with what would be expected under standard game theoretical considerations. Following the taxonomy of Potters et al., 1998, we label this type as *gamesmen*. For the third type this relationship appears to be inverted, and moreover, efforts are decreasing over time. We do not find an intuitive explanation for this behavior in the contest literature, which is why we simply label them *others*.[12] Due to the non-linearity of Tobit models, we graphically present for each type the marginal effects of *l.otherefforts* on *efforts* for different levels of *l.otherefforts* in figure B.1 in Appendix B.1. Analyzing the shape of the marginal effects of *l.otherefforts* on *efforts* for each group, we find our above reported considerations confirmed. Further, for every C-Lasso regression the estimated standard error of the idiosyncratic error term σ_ϵ is lower than in the pooled Tobit regression, suggesting a better fit of the group estimation due to lower heterogeneity within groups than in the overall sample.

Result 1. *C-Lasso identifies three contestant types for contests with fixed opponents, which we refer to as reciprocators* ($\approx 58\%$)*, gamesmen* ($\approx 9\%$)*, and others* ($\approx 32\%$) *based on how own efforts are explained by opponent efforts in the previous contest round.*

For *reciprocators*, that form the majority of contestants, we want to understand whether their reciprocal bidding behavior reduces average in-

[12] If contestants of type *others* expect their opponents to alternate effort levels, then responding to previously low opponent efforts with own high efforts can be reasonable. In that case we should find that current efforts relate to efforts of opponents in $t-2$ in a concave fashion. In table B.2 in Appendix B.1, we provide exploratory evidence using pooled Tobit regressions on the *others* type that include second lag effects.

Figure 3.1: Relationship between the share of *reciprocators* and average effort - FM treatments

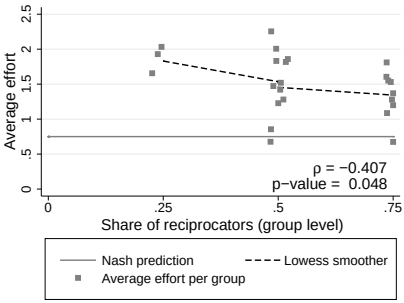
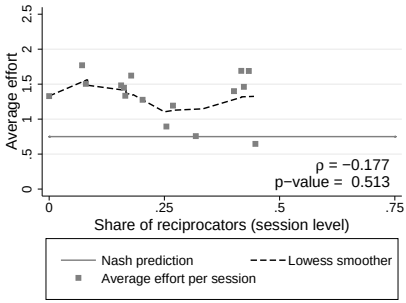


Figure 3.2: Relationship between the share of *reciprocators* and average effort - RM treatments



dividual efforts or exacerbates competition. We thus examine how the share of *reciprocators* in a contest group influences the average group effort. Remember that for all treatments each group consists of four contestants. If reciprocal behavior results in collaboration, then groups with a higher fraction of *reciprocators* should be associated to lower average effort levels. On the contrary, if reciprocity increases competition then we should witness higher average effort levels in groups formed by many *reciprocators*. Figure 3.1 shows how the share of *reciprocators* per group relates to the average group effort. As suggested by the direction of the lowess smoother, we find that average group effort is decreasing with the fraction of *reciprocators* per group (spearman's $\rho = -0.407$ p-value $= 0.048$), suggesting that *reciprocators* reduce efforts conditional on the cooperation of opponents.[13] Comparing the average effort levels to the Nash prediction, we observe substantial rent dissipation across groups, in line with the consensus in the contest literature (see Dechenaux et al., 2015).

[13] A low share of *reciprocators* implies a high share of contestants assigned to other types. We thus examine if a decrease in average group efforts is not only associated with an increase in the share of *reciprocators* but also associated with a decrease in the share of *gamesmen* or *others*, but we do not find evidence for such an association.

Result 2. *Average group level expenditures are lower for contest groups with a higher share of reciprocators.*

Contestants of the FM treatments know that their opponents stay the same over the course of the experiments, so that cooperative tendencies seem to occur naturally. Surprisingly, the *reciprocator* as player type has not received much attention in contest experiments. This might be because in many contest experiments, including Potters et al., 1998, contestants are randomly regrouped after each round, making reciprocal actions less effective. Using data on four different RM contest studies, we explore whether *reciprocators* also appear in RM treatments.

Table 3.4: C- Lasso Tobit regression results for RM treatments

	Pooled Tobit	C-Lasso Tobit		
Dep. variable: *effort*		Type 1 *(Reciprocators)*	Type 2	Type 3
l.othereffort	0.350***	0.167***	-0.058***	-0.551***
	(0.017)	(0.048)	(0.021)	(0.053)
$l.othereffort^2$	-0.095***	-0.035*	0.033***	0.208***
	(0.009)	(0.021)	(0.010)	(0.021)
l.win	0.209***	0.119***	0.020**	0.146***
	(0.010)	(0.017)	(0.008)	(0.016)
period	0.001*	-0.006***	-0.006***	-0.001
	(0.001)	(0.001)	(0.000)	(0.001)
σ_ϵ	0.321***	0.271***	0.188***	0.309***
	(0.003)	(0.005)	(0.003)	(0.007)
Obs;N;%	5307; 183; 100%	1305; 45; 25%	2784; 96; 52%	1218; 42; 23%

Standard errors (in parenthesis); p-values: *$\leqslant$ 0.10, **$\leqslant$ 0.05, ***$\leqslant$ 0.01; Obs. is the total number of observations; N is the number of contestants of each type. % is the relative share of each type with respect to the full sample.

Result 3. *C-Lasso identifies three contestant groups for contests with randomly matched opponents. Only 25% can be identified as reciprocators, significantly less than for fixed matched treatments.*

As in the fixed matching panel, C-Lasso categorizes the 183 RM contestants into three types. Table 3.4 reports the C-Lasso estimates and compares them to the pooled Tobit regression estimates, whose signs and significance level resembles the estimates of the FM panel. The first type contains 45 contestants, whose regression estimates and marginal effect of *l.othereffort* on effort resemble the ones of *reciprocators* in FM treatments. As for FM treatments we report the marginal effects of *l.othereffort* on *effort* in figure B.2 of Appendix B.1. Again we analyze the relationship between the share of *reciprocators* and average effort. Since in RM treatments the share of *reciprocators* per group may change every round, we look at the relationship on the session level for which contestants stay fixed. The RM treatments contain in total 16 sessions. The lowess smoother in figure 3.2 suggests no directional relationship between the share of *reciprocators* in a session and the average session effort. This is in line with the spearman correlation coefficient, which we find not significant (spearman's $\rho = -0.177$ p-value $= 0.513$). Similar to FM treatments, the average session efforts mostly exceed the theoretical equilibrium level.

The share of *reciprocators* across RM contestants reaches only 25% which is significantly lower than in treatments where opponents remain fixed (about 58%).[14] Presumably we observe less *reciprocators* in randomly matched treatments, since there is no direct feedback on the effectiveness of own collaboration attempts. The other contestant types two (52%) and three (22%) increase (decrease) their efforts after previously won (lost) rounds and show a u-shaped relationship between *l.othereffort* and current own efforts. The magnitude of the regression coefficients suggests

[14] A Pearson's chi-squared test, conducted on the contingency table of the empirical distribution of the number of *reciprocators* in the two samples (FM and RM), rejects the null hypothesis that the samples are drawn from the same super-population (p-value < 0.001).

that previous round information affects current efforts to a lesser extent for type two contestants. Similar to the *others* type in FM treatments, both types differ from what is expected under myopic best response considerations, presumably because contestants do not expect new opponents to behave as their last round opponents did.

3.5 Discussion and Conclusion

In many competitive situations the development of a contestant taxonomy is impeded because one cannot directly assess the motivation behind individual choices. We use the C-Lasso methodology that identifies latent group structures in panel data and thus categorizes contestants based on how their effort choices relate to information they received about the previous round.

Using data from six different studies, we find that contestants react heterogeneously to information of the previous contest round. C-Lasso identifies three types of contestants for both contest regimes, the ones with fixed and the ones with randomly matched opponents. The majority of contestants in FM treatments can be classified as *reciprocators* that attempt to cooperate by lowering own efforts when previous opponent efforts were low. In fact, contest groups that contain more *reciprocators* show lower average effort levels and thus higher group profits. For RM treatments the number of *reciprocators* appears to be significantly lower, in line with the idea that own cooperative attempts in the current round cannot be observed by ones next rounds opponents.

Our exploration on what types of contestants are out there hopefully releases a pulse on how we think about competitive situations such as conflict resolution. Once we acknowledge that choice rules of competing individuals are heterogeneous, we can develop a better understanding for the tools that are needed to mitigate conflicts and to reduce overspending in competitive situations. For example, *reciprocators* are willing to conditionally cooperate over the course of the contest, which is most

fruitful whenever the number of them in a group is high and thus cooperation can reinforce itself. We can expect that knowing the type composition of groups delivers valuable insights on what group dynamics can be anticipated, across applications.

One application is provided by Bordt and Farbmacher (2018) who use experimental data of repeated public good games to replicate the classification of behavioral types proposed by Fischbacher et al., 2001 using the C-Lasso mechanism. Their results are consistent with the re-analysis done using hierarchical clustering by Fallucchi, Luccasen, et al. (2018) and suggest a coherence of subject behavior in repeated games with the choices elicited via the strategy method. In the case of the contest typology, further research needs to be carried out to understand the main drivers of behavioral types.

We suggest two directions for future research. The first direction is to check whether different types react differently to other changes in the contest structure. For example, excessive over-expenditures observed in larger groups (Lim et al., 2014) may be driven by less cooperative activity, that has been demonstrated for different experimental settings (Nosenzo et al., 2015). The second direction concerns the use of methods that reveal latent group structures in panel data in order to shed light on different topics, such as framing e.g., Chowdhury, Mukherjee, et al., 2017 or impulsive behavior e.g., Rubinstein, 2016; Sheremeta, 2018.

Chapter 4

Clinical Foresight using Supervised Learning

4.1 Introduction

Machine learning (ML) tools are used with increasing success across industries to improve decision-making. For the pharmaceutical industry, which spends more than $160 billion annually in R&D (Malik and Lisa, 2018) but does not launch the majority of its candidate drugs due to safety, efficacy or market profitability concerns (Joseph A DiMasi, 2001), an approach that could predict the outcomes during R&D phases would be particularly valuable. We develop such an approach that relies on established supervised learning (SL) algorithms, trained and validated with the help of a large sample of an EvaluatePharma® data set which includes about 8,800 pharmaceutical projects undertaken in the United States during the last decade.

The outcome prediction of pharmaceutical projects under development has been facilitated by two streams of research. The first stream focuses on the causal relationship between successful drug development and various company or product characteristics. It has revealed that product success in the pharmaceutical industry is not only determined by

the molecular properties of the drug, but relates to surrounding characteristics such as firm specific competence (Henderson and Cockburn, 1994), firm location (Kyle, 2006), and orphan drug status (Regnstrom et al., 2010), just to name a few. The second stream of research estimates success rates across different indications and technologies using historical data on clinical trials (Joseph A DiMasi, 2001; Hay et al., 2014; Wong et al., 2019; Wong et al., 2018). These estimates, that thanks to the increasing availability of clinical trial data become more and more accurate, are often used as a first approximation for success probabilities on the project level. In this work we combine both the historical success rates of similar projects and an array of other potential success characteristics to train a selection of SL-algorithms used to predict the phase-specific outcome of pharmaceutical projects.

Artificial intelligence forms already an integral part of the healthcare industry (Topol, 2019) with ML methods being used in the early stage of the drug development process to screen candidate compounds (Burbidge et al., 2001; L. Zhang et al., 2017). It is surprising that their use has not been extended until recently to predict the outcome of more advanced clinical stages. J. DiMasi et al., 2015 present a tool that uses information on drug specific characteristics to predict regulatory marketing approval of new oncology compounds, whereas Lo et al., 2019 use data on clinical trials to analyze classification and missing data imputation techniques from phase II clinical trials to market approval. Feijoo et al., 2020, as well use clinical trial data to classify phase II and III development outcomes using random forests. We contribute to this developing field by evaluating different SL-algorithms on a mixture of product, market and company characteristics that allow us to predict project outcomes from phase I clinical trials untill market launch.

We provide a brief introduction into supervised learning in the next section. Afterwards we explain the construction of the phase-specific data sets (section 4.3) and the procedure used to train and validate the algorithms we analyze (section 4.4). In the result section 4.5 we compare the predicitve performance of all SL methods (sub section 4.5.1), and two traditionally used methods for predicting pharmaceutical project outcomes

(sub section 4.5.2). Across methods, a Bayesian additive regression tree (BART) algorithm delivers the most accurate out-of-sample predictions, which is subsequently used to predict the outcomes of the current development pipeline (sub section 4.5.3). We complete the result section by analyzing the most predictive success characteristics (sub section 4.5.4). The concluding discussion (section 4.6) puts our insights into context and addresses stakeholders in the pharmaceutical industry that could benefit from increased predictive capabilities regarding pharmaceutical innovations.

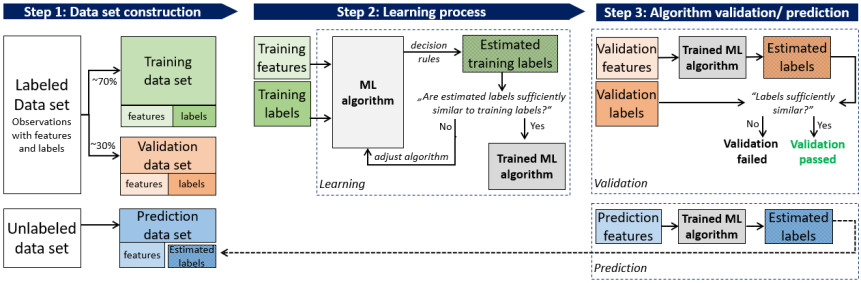
4.2 Supervised Learning in a Nutshell

Learning computer algorithms that evaluate and automatically improve their performance go back many decades. In 1952 Arthur Samuel designed one of the first computer learning programs that improved its ability to play checkers by learning from previous moves. He coined the term "machine learning" (ML), which has come to designate a computer algorithm that "learns" to better its performance on a specific task. Since the 1950s machine learning has made huge advances that have heightened its performance and broadened its appeal. Image recognition software, email filters, and personalized advertisement are just some of the applications which rely on ML technology. And thanks to the growing availability of large data sets, machine learning is making its way into healthcare, including drug discovery (Burbidge et al., 2001; L. Zhang et al., 2017), medical imaging, and health monitoring (Topol, 2019).

A simple three-step supervised learning routine is depicted in figure 4.1. Each observation of the labeled data set, in our application each pharmaceutical project, is characterized by a set of project characteristics, so called features, and a known outcome variable, its label. In our case the project outcome is categorical (either classified as success or failure) and hence the learning methods are termed classification methods whereas one would speak of regression methods whenever labels are continuous. The labeled data set is randomly split into two data sets

(step 1): a training set that contains 70% of the labeled data and a validation set (containing the other 30%). The chosen split ratio is context dependent but should be chosen so that both training and validation set contain a sufficiently high number of observations. During the learning process (step 2) the SL-algorithm establishes decision rules to assign labels to the training observations based on its features. Supervised learning algorithms, in contrast to unsupervised learning algorithms, use the true labels of the training data to evaluate their own performance (Sammut and Webb, 2017). Decision rules are repeatedly adjusted until the estimated labels become sufficiently close to the true labels. At that point, the algorithm is deemed to be trained. Note that the way decision rules are constructed and estimated labels are compared to true outcomes is algorithm specific.

Figure 4.1: Exemplary supervised machine learning routine



Once the algorithm is trained, it is validated by applying it to the validation set, which it has never been exposed to (step 3). To be successfully validated, it must approximate the labels of the validation set with an accuracy that is sufficient for its purpose. When the output data is binary (e.g. success vs. failure) the performance of the algorithm can be summarized by a confusion matrix (see template in figure 4.5). If the estimated labels of the validation set are substantially close to the true labels, it indicates that the trained algorithm has learned to classify observations. After successful validation, the trained algorithm can be used to predict

labels for similar, unlabeled observations. In our application these are projects for which the outcome is not yet known. The great advantage of SL over traditional statistical methods, such as regression analysis, is that it excels at making use of interactions and non-linear relationships between features that are embedded in the data but unknown to the researcher.

4.3 Data Set Characteristics

We train five algorithms on a database of drug development projects to predict the success or failure of the clinical research phases in which they are engaged. Each project is a combination of features and labels. The features recapitulate the attributes of each project, such as the characteristics of the molecule, intended market, company attributes, while the label indicates the status of its most advanced clinical research phase- e.g., success, failure, or on-going. For instance, a project might be *lorlatinib* to treat ALK positive, non-small cell lung cancer. The input data would describe a small molecule developed by Pfizer that was granted orphan status and expedited Food and Drug Administration (FDA) treatment. The output data would indicate: "Filed, on-going".[1] Most of the project features were obtained from a novel database developed by Evaluate Ltd., a commercial company that collects and integrates company-reported and other published pharmaceutical product and financial information, to which we were granted access. Data from PATSTAT, a database that covers information on patents worldwide, is used to complement the main source for patent-specific information.

To ensure that all projects share the same regulatory framework, we limit the sample to new molecular entities (NME) monitored by the FDA that report at least one drug with indication status in the US, the worldwide biggest geographical market for pharmaceutical products which is regulated by a single authority (Kyle, 2006). New drug applications,

[1] We define projects with expedited FDA treatment as projects that fall into at least one of the four categories defined by the FDA: priority review, breakthrough therapy, accelerated approval, and fast track. The molecule has since been approved.

generics, biosimilars, as well as over the counter products are excluded, due to differences in approval regulations. Furthermore, only projects that were initialized between 2008 and 2017 are included in the sample to reduce reporting bias.[2] Prior to 2008 companies had little incentive to publicly disclose information about their projects' development status and were required only in case of market approval by the FDA. The FDA Amendments Act (FDAAA 801) of 2007 raised the standards for regulatory approval, strengthened the FDA's post-market drug surveillance (Kaitin and Joseph A DiMasi, 2011) and since then requires companies to publicly disclose trial information on most initiated and on-going clinical trials (Zarin et al., 2016). Since then, the number of reported abandoned projects relative to the number of marketed projects has risen sharply and thus provides a more realistic project outcome distribution. The difference in the sample composition before and after 2008 is visualized in figure 4.2.

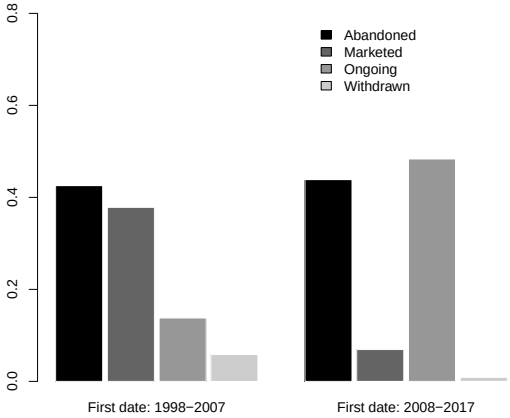
The complete data set spans 4557 NMEs associated to 1328 companies for a total of 634 indications leading to 8785 unique company-product-indication combinations, referred to as projects. Indications classify diseases or disorders using a three tier categorization, similar to EphMRA anatomical classification (ATC) system. We use the most granular third level (indlev3) in the subsequent analysis. NMEs that share the same indlev3 can be thought of as substitutes, since they target the same, or a very similar, therapeutic market. For simplicity, we follow the custom of the pharmaceutical economics literature and refer to a specific indlev3 as "market".[3]

Pharmaceutical projects regulated in the US are required to undergo subsequent clinical phases which we denote by the common FDA notation: "Phase I", "Phase II", and "Phase III". Generally speaking, the more a project advances in its development, the more its focus of investigation shifts from safety to efficacy concerns and the higher becomes the number of subjects enrolled in clinical trials. Projects that have passed

[2] Table C.5 in Appendix C.2 describes the data set construction in detail.

[3] See for example Pammolli et al., 2011, who use the ATC classification to refer to different product markets, or Regnstrom et al., 2010, who address to even broader therapeutic areas as markets.

Figure 4.2: Project status grouped by first status date before and after 2008



all clinical trial stages and are filed for approval but not yet marketed are said to be "Filed", whereas projects that were successfully launched in the US are referred to be "Marketed". Since the probability of success varies at each stage of the development and stage-specific success characteristics might change during development, we analyze success predictions depending on the development stage.[4]

Before training the algorithms, the project database was partitioned into four subsets, for projects having reached phase I, II, III, and Filed respectively. Pre-clinical projects prior to Phase I are excluded from the analysis, since companies are not required to report them. The first data set (PI) categorizes each project that has successfully passed Phase I as "success", each project that currently undergoes Phase I trials as "on-going", and each project that was abandoned or suspended during Phase I as

[4]Please refer to figure C.1 in Appendix C.2 for an overview of the average characteristics of different clinical phases in terms of success probability, years required, and costs.

"failure". For the PII data set all projects that completed PII are classified as success, all projects currently in Phase II as on-going, and all suspended or abandoned projects in Phase II as failure. Projects that have not (yet) reached phase II are not included in the PII data set. The remaining data sets PIII and Filed are constructed accordingly. Table 4.1 shows the grid used to ascertain the label of each project, and shows the success rates attributed to each phase by the projects in our sample. In each data set the successes and failures are used to train (and validate) the supervised learning algorithms. Once the SL-algorithms are trained they can be used to predict phase-specific success outcomes of on-going projects.

The success of a project depends on the interplay between treatment efficacy, safety, favorable market conditions and company specific factors. The decision rules formed by the SL-algorithms used to determine the outcome of a project are thus based on a variety of candidate features, that combine company, product and market factors which characterize the project. The set of candidate features has been developed involving a panel of industry experts during various conversations with Evaluate Ltd. Table 4.2 describes the final 36 project characteristics of the candidate feature set grouped by feature dimension. We opt for a rather extensive set of candidate features to not unnecessarily restrict the prior choice set of SL-algorithms when forming decision rules. Whether a specific feature is chosen by the algorithm to classify projects into successes and failures depends on its predictive aptitude. To avoid potential variable selection issues, we validate for each algorithm whether a ML-based variable selection step increases the algorithms out-of-sample accuracy compared to using the full set of candidate features.

Table 4.2: Description of candidate features used in SL-algorithms

Feature name	Feature description
Product marketed	Indicator. 1 if product is marketed for another indication before the phase status date
	Continued on next page

	Feature name	Feature description
PRODUCT	Product failed	Indicator. 1 if product has failed for another indication before the phase status date
	MoA validated	Categorizes whether the MoA has been validated for the same indication, a different indication or has not been validated before the phase status date
	Product experience (count)	Number of distinct indications in which the product was already active before the phase status date
	Patent duration	Duration of the patent coverage from filing date to phase status date
	Clinical trial cost	Actual or estimated trial cost of the product sourced at EvaluatePharma®
	Clinical trial results	Categorize whether clinical trial results: unavailable, partial, negative, mixed, or positive. Phase II results only used in PII, Phase III results only used in PIII
	Time in phase	Time from start of the phase until phase status date
	Patents cite	Count of the distinct patent families that refer to the main patent of the product. Sourced from PATSTAT database and merged to data set
	Patents cited	Count of the citations of distinct patent families that the main patent of the product refers to. Sourced from PATSTAT database and merged to data set
	Technology	Categorizes the technology of the product used
	Product strategy	Categorizes whether the product is developed in-house, via licensing, via company acquisition, product acquisition or joint venture
	Therapy type	Categorizes whether the product has one or more active ingredients
	Companies per product	Counts the number of companies involved in the development of the product
MARKET	Market size (companies)	Number of companies with marketed products for each indication
	Market size (products)	Number of distinct marketed products for each indication
	Market inequality	Standard deviation of product revenues (2017) for all marketed products in each indication
	Orphan drugs (count)	Current number of marketed Orphan drugs for each indication
	Indication level 1	Indication category aggregated to different therapeutic areas
	Market success ratio	The sum of marketed products over the sum of marketed withdrawn and abandoned products for each indication
	Phase time (indication)	Phase specific median development time for each indication
	Phase success rate by indication	Phase specific historic success rate by indication
COMPANY	Own similar products (count)	Number of distinct similar products [similar products are products that rely on same technology] in which the company was active before the phase status date
	Market experience (count)	Number of distinct products for the same indication, in which the company was active before the phase status date

Continued on next page

	Feature name	Feature description
COMPANY	Time in market	Number of months of a company's experience in indication level 1 before the phase status date
	Own similar markets (count)	Number of distinct similar markets [similar markets share the same indication level 1] in which the company was active before the phase status date
	R&D cost	Research and development expenses of the company in the phase status year
	Company listed	Indicator. 1 if company was publicly traded before the phase status date
	Company classification	Categorizes companies in four distinct groups: Biotechnology, Global Majors, Regional Majors and Specialty
	Region	Categorizes companies in regions based on their legal headquarter: Africa & Middle East, America ex USA, Asia & Oceania, Europe, USA
	R&D (count)	Number of active R&D products of company
	Products (count)	Number of marketed products of company
	Company success ratio	The sum of marketed products over the sum of withdrawn and abandoned products for each company
OTHER	Orphan status	Indicator. 1 if the project is assigned orphan status in the US
	Expedited status	Indicator. 1 if project is assigned expedited treatment by the FDA
	Phase success rate by MoA	Phase specific historic success rate by mechanism of action (MoA)
	Phase success rate by Technology	Phase specific historic success rate by product technology, such as biotechnology, vaccine or gene therapy

Given the set of candidate features for each stage-dependent data set, we turn in the next section to the training and testing procedure of the SL-algorithms.

4.4 Supervised Learning Approach

The section is split in two parts. In the first part we briefly comment on the SL-algorithms used, in the second part we detail the learning procedure that we perform on each algorithm.

4.4.1 Employed Supervised Learning Methods

We train five different supervised learning algorithms to detect decision rules in the feature space that discriminate between successful and

Table 4.1: Project outcome classification and number of projects for phase-specific data sets

	Data set			
Project status	PI	PII	PIII	Filed
Filed, completed	Success 441	Success 435	Success 493	Success 460
Filed, on-going	Success 45	Success 51	Success 71	On-going 75
Filed, abandoned/suspended	Success 12	Success 13	Success 15	Failure 16
Phase III, on-going	Success 336	Success 347	On-going 559	
Phase III, abandoned/suspended	Success 248	Success 147	Failure 290	
Phase II, on-going	Success 844	On-going 2372		
Phase II, abandoned/suspended	Success 735	Failure 1794		
Phase I, on-going	On-going 1858			
Phase I, abandoned/suspended	Failure 1231			
Total number of projects by phase	5750	5159	1428	551
Avg. success ratio	68.4%	35.7%	66.6%	96.6%

Table 4.1 summarizes the outcome classification of projects into four stage-dependent data sets according to their project status. Note that projects are excluded for a specific data set if the estimated time in the specific phase is erroneous.

failed projects in each of the four development phases. Four of the algorithms are tree-based classification methods, a simple decision tree (DT), boosted decision trees (C5.0) (Kuhn and Johnson, 2013), a random forest algorithm (RF) (Breiman, 2001), a Bayesian additive regression tree (BART) (Chipman et al., 2010; Kapelner and Bleich, 2013) and a linear probabilistic regression (PROBIT).[5] Tree-based classification methods are advantageous in applications at which non-linearities and interactions between features are plausible, but unknown. A classification tree can be thought of as a set of successive decision rules, called branches. At each branch a feature is selected that, according to a specified metric, best splits the set of observations to classify them as successes or failures. The classifications resulting from the successive application of decision rules are also called leaf nodes. The DT algorithm relies on only one tree while C5.0, RF and BART use an ensemble of trees but formed in different ways. The C5.0 method uses gradient boosting enabling the algorithm to learn from classification errors of prior trees, RF averages over estimates from multiple trees based on a random subset of features and projects, and lastly BART sums the contribution of multiple trees whose structure depends on Bayesian priors. In addition, we train a standard probabilistic regression PROBIT model to compare how classification accuracy changes, when outcomes are predicted by a linear combination of features without variable interactions.

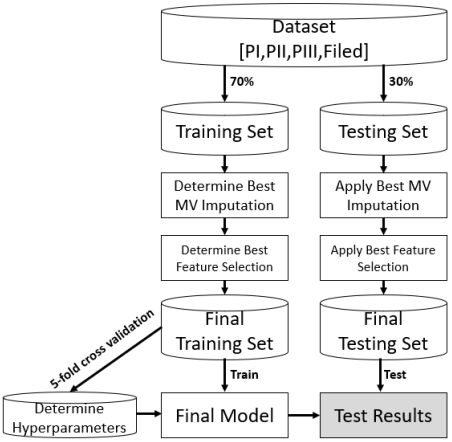
4.4.2 Learning Procedure

The known outcomes of each of the four data sets are randomly split into a validation set (30%) and a training set (70%) for which all five algorithms are trained. The validation set is kept separate to determine the classification accuracy using observations that the algorithms have not encountered during training. Figure 4.3 sketches the training and validation procedure employed in this chapter. We implement a missing

[5] A detailed description of each algorithm would go beyond the scope of this chapter. For readers who wish to learn more about the algorithms particularities, we suggest the above mentioned contributions.

value imputation step, a feature selection step, and a hyperparameter tuning step before training the final SL-algorithms.

Figure 4.3: Supervised learning procedure used for each algorithm and data set



Before running supervised learning algorithms it is important to handle missing values in the feature set, since missingness of features might not be completely at random and thus itself be indicative for label estimations. However, there exist many approaches that are appropriate to deal with missing data. So before training the algorithms, we analyze three different ways on how to cope with missing data and choose for each algorithm the procedure that maximizes its out-of-sample predictive performance. First, we perform a rudimentary complete case (CC) analysis in which features with more than 70% of missing values and all remaining observations that contain missing values are excluded. Second, we analyze a five nearest neighbor (5NN) imputation algorithm that achieved good classification results in a similar application (Lo et al., 2019). Third, for each supervised learning algorithm we use the internal imputation technique (II) that comes as default in each algorithms' *R* implementation (not available for PROBIT). Each training set is once more randomly split into 70% of training data (49% of complete data) on

which each algorithm is trained for each of the three missing value techniques. The resulting classifications are then evaluated on the remaining 30% of the training data (21% of complete data) using the area under the receiver operating characteristic curve (AUC) as evaluation criterion. That way we make sure to select the missing value technique that, across data sets, ensures the highest AUC for each algorithm without the use of the separately kept validation data (30% of complete data).

Table 4.3: AUC comparison for missing value imputation techniques across methods and data sets

		PI	PII	PIII	Filed
CC	PROBIT	**0.8318**	0.7096	0.7333	-
	DT	0.3379	0.6543	0.29	-
	RF	0.8127	0.7755	0.76	-
	BART	0.8496	0.7875	0.8	-
	C5.0	0.8293	0.7955	0.8267	-
5NN	PROBIT	0.8244	**0.8737**	**0.7664**	**0.6774**
	DT	0.6612	0.775	0.3634	**0.8104**
	RF	0.8203	0.9025	0.8226	0.9094
	BART	0.82	0.8942	0.7851	**0.9729**
	C5.0	**0.8379**	0.8945	**0.7812**	**0.7031**
II	PROBIT	-	-	-	-
	DT	**0.7349**	**0.8252**	**0.7098**	0.5
	RF	**0.8422**	**0.9299**	**0.8784**	**0.9385**
	BART	**0.9182**	**0.9607**	**0.8942**	0.9562
	C5.0	0.774	**0.8968**	0.7098	0.5

Table 4.3 shows the out-of-sample AUC for each imputation technique (CC, 5NN and II), algorithm (PROBIT, DT, RF, BART, C5.0), and data set (PI, PII, PIII, Filed). For projects in the Filed stage CC leaves a sample that consists only of successes for which no training is possible. The highest AUC of each imputation technique given the sample algorithm and the data set are highlighted in bold.

Table 4.3 shows the out-of-sample AUC for each imputation technique across data sets and SL-algorithms. It turns out, that the highest classification values in terms of AUC are achieved by 5NN imputation for PROBIT and C5.0 while the other methods are more accurate using their in-

ternal missing value approaches. Missing features are imputed without considering label information in order to prevent that imputed values reflect this information. Moreover, validation and training data are imputed separately to avoid inducing any form of relation between them.

After imputing missing values, we perform a feature selection step to rule out that excessive feature inclusion inhibits the prediction quality of the algorithms. We evaluate three feature selection methods: LASSO (Tibshirani, 1996), an often-applied method that uses the shrinkage of linear regression coefficients for feature selection, RF_SE, a successive variable elimination technique in a random forest based on the smallest prediction loss (Diaz-Uriarte and Andrés, 2005), and BART_IP that selects variables based on their inclusion proportion in a BART algorithm with a small number of trees (Bleich et al., 2014). Following the approach for missing value imputation, we use 70% of the training data to train each algorithm on each data set for each feature selection approach and compare the results based on the AUC from validating the 30% of remaining training data. Table 4.4 shows the results. For PROBIT and DT the feature selection technique based on the RF prediction loss delivers the highest AUCs across data sets. For the other three algorithms we chose to omit the feature selection step, since its effectiveness was mixed across data sets.

In a last step before the final training, we tune hyperparameters using 5-fold cross validation on the complete training data for each data set. Hyperparameters that are tuned are the number of trees and the cut-off probability in BART, the number of randomly sampled variables at each split in RF and the number of boosting iterations in C5.0.[6] After this stage, every supervised learning algorithm is trained on the full set of training data reflecting its best performing missing value technique, feature selection criterion and adjusted hyperparameters. The algorithms are evaluated using the validation data set that has been kept separated from the whole training procedure. To rule out that the test results are influenced by random selection of validation data, the final training and validation procedure is performed 10 times for each model and each data

[6]For completeness we report the hyperparameter tuning results in Appendix C.1.

Table 4.4: AUC comparison for variable selection techniques across methods and data sets

		PI	PII	PIII	Filed
LASSO	PROBIT	**0.8258**	0.8758	0.7857	-
	DT	**0.7372**	0.8423	0.7098	0.5
	RF	0.8419	0.928	0.874	0.7562
	BART	0.8961	0.9563	0.9041	0.899
	C5.0	0.8279	0.8936	0.8245	0.5
RF_SE	PROBIT	0.8247	**0.8839**	**0.8091**	**0.8562**
	DT	0.7343	**0.8363**	**0.861**	0.5
	RF	0.8345	0.9198	0.861	0.8958
	BART	0.8866	0.948	0.8923	0.8958
	C5.0	0.7924	**0.8946**	**0.8339**	0.5
BART_IP	PROBIT	0.7786	0.8758	0.8082	0.7766
	DT	0.7319	0.8444	0.7098	0.5
	RF	**0.8488**	0.9229	0.8734	**0.9469**
	BART	0.9039	0.952	**0.9086**	**0.9688**
	C5.0	0.7532	0.8765	0.8143	0.5
No feature selection	PROBIT	0.8244	0.8737	0.7664	0.6774
	DT	0.7349	0.8252	0.7098	0.5
	RF	0.8422	**0.9299**	**0.8784**	0.9385
	BART	**0.9182**	**0.9607**	0.8942	0.9562
	C5.0	**0.8379**	0.8945	0.7812	**0.7031**

Table 4.4 shows the out-of-sample AUC for each variable selection technique (LASSO, RF_SE, BART_IP and "No features selection"), algorithm (PROBIT, DT, RF, BART, C5.0), and data set (PI, PII, PIII, Filed). The highest AUC of each variable selection technique given the sample algorithm and the data set are highlighted in bold.

set and average performance measures are reported. The next section starts by comparing the average performance between SL-algorithms across the four different data sets.

4.5 Results

The result section is split in four parts. The first one (section 4.5.1) compares the performance of SL-methods with each other. The second part (section 4.5.2) contrasts the best performing SL-algorithm with commonly used classification approaches. Part three (section 4.5.3) applies the best performing algorithm to the set of on-going projects and predicts its outcomes. Lastly, part four (section 4.5.4) details which factors are most often selected to classify outcomes across algorithms.

4.5.1 Comparison of Supervised Learning Methods

The average performance of the SL-algorithms on validation sets is summarized in table 4.5.[7] For all algorithms and data sets we report the average AUC, its 95% confidence intervals computed by the method of DeLong et al., 1988, and standard deviations across validation sets. An advantage to use AUC as a performance criterion is that its value is independent of the choice of a specific classification threshold (Bradley, 1997). The higher the AUC, the better the algorithm solves the trade-off between type I and type II prediction errors. An AUC of 0.5 indicates that the maximum share of correct classifications achieved by any classification threshold is one half, meaning that random assignment would be equally predictive. An AUC of 1 indicates that there exists at least one classification threshold at which the model classifies each case correctly. Besides the AUC, the algorithms sensitivity (the number of correctly classified successes over the total number of true successes) and its specificity (the number of correctly classified failures over the total number of true

[7] Instead of reporting the average of ten validation repetitions, we report the performance of SL-algorithms using just one randomly split training and validation set in table C.6 of the supplementary material C.2. Overall, results stay the same.

failures) at a classification threshold of 0.5 are reported. We include these measures to provide two more intuitive criteria.

Table 4.5: Average validation results of five SL-algorithms across data sets

		AUC	AUC SD	AUCL	AUCL SD	AUCH	AUCH SD	SENS	SENS SD	SPEC	SPEC SD
BART	PI	**0.92**	0.01	0.91	0.01	0.94	0.01	0.90	0.02	0.75	0.03
	PII	**0.96**	0.01	0.94	0.01	0.97	0.00	0.80	0.01	0.95	0.01
	PIII	**0.94**	0.02	0.91	0.02	0.97	0.01	0.89	0.03	0.84	0.05
	Filed	**0.92**	0.06	0.76	0.10	0.99	0.01	1.00	0.00	0.00	0.00
C5.0	PI	0.86	0.01	0.83	0.02	0.89	0.01	0.91	0.02	0.58	0.06
	PII	0.90	0.01	0.87	0.01	0.92	0.01	0.70	0.02	0.92	0.02
	PIII	0.85	0.03	0.79	0.03	0.91	0.02	0.88	0.04	0.60	0.09
	Filed	0.78	0.20	0.48	0.25	0.88	0.20	0.99	0.01	0.16	0.22
RF	PI	0.87	0.02	0.84	0.02	0.89	0.02	0.95	0.02	0.51	0.05
	PII	0.91	0.01	0.89	0.01	0.94	0.01	0.78	0.03	0.90	0.02
	PIII	0.90	0.03	0.86	0.04	0.95	0.02	0.94	0.02	0.57	0.09
	Filed	0.81	0.16	0.51	0.31	1.00	0.00	1.00	0.00	0.12	0.32
PROBIT	PI	0.84	0.02	0.81	0.02	0.87	0.02	0.89	0.02	0.58	0.06
	PII	0.87	0.01	0.84	0.02	0.90	0.01	0.66	0.03	0.90	0.02
	PIII	0.85	0.01	0.79	0.02	0.92	0.00	0.87	0.03	0.66	0.01
	Filed	0.86	0.04	0.75	0.07	0.99	0.01	0.97	0.02	0.00	0.00
DT	PI	0.68	0.18	0.65	0.18	0.72	0.18	0.91	0.02	0.43	0.06
	PII	0.83	0.02	0.80	0.02	0.87	0.02	0.55	0.05	0.90	0.03
	PIII	0.79	0.03	0.72	0.03	0.86	0.02	0.87	0.06	0.52	0.08
	Filed	0.50	0.00	0.50	0.00	0.50	0.00	1.00	0.01	0.02	0.08

Table 4.5 shows the average out-of-sample AUC, its 95% confidence interval (AUCL, AUCH) the sensitivity (SENS), specificity (SPEC), and respective standard deviations (SD) of each SL-algorithm for each data set. The highest average AUC across algorithms for each data set is highlighted in bold.

Result 1. *For all data sets, the BART algorithm achieves the highest classification performance compared to the other four supervised learning methods.*

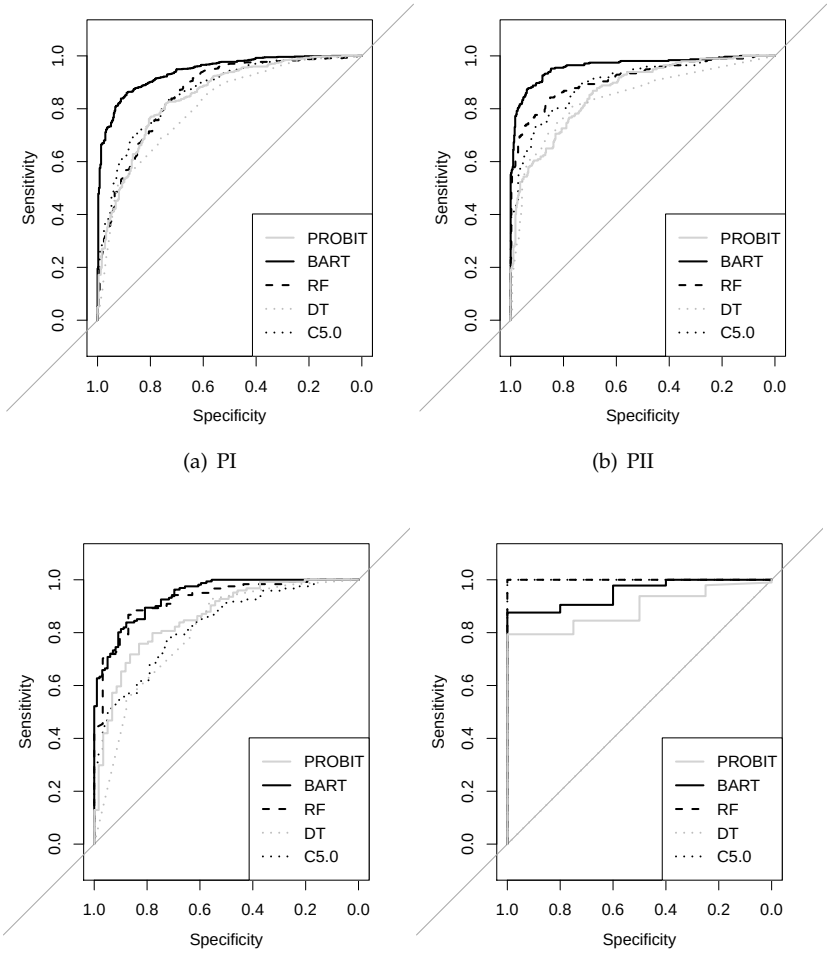
In all four data sets the BART algorithm achieves the highest classification performance with an average AUC of 0.92 in PI, 0.96 in PII, 0.94 in PIII, and 0.92 in Filed. We further refer to it as "best-in-class" algorithm. RF shows a lower performance than BART, but its AUC 95% confidence intervals overlap with the ones of BART in PII, PIII and Filed. The results of both algorithms are not influenced by additional feature selection or missing value imputation techniques, which makes both approaches powerful stand-alone tools given our application. For C5.0 we find a slightly lower average AUC than for RF. The performance of PROBIT is similar to C5.0 which shows that careful choice of missing value imputation and feature selection techniques can offset the linear restrictions imposed by the model. The worst performing method in terms of average AUC is DT, possibly since it does not rely on a corrective mechanism such as re-sampling or boosting.

Looking at the sensitivity and specificity columns, we note that the algorithms seem to distinguish successfully between successes and failures in the first three data sets. For PI and PIII the algorithms perform better during success classification, while for PII the opposite is the case. For example, the BART algorithm classifies correctly 90% of the successes and 75% of the failures in PI using the default classification threshold of 0.5. In the last data set the average specificity for all algorithms is close to zero which indicates that failures are not classified correctly. Since failures in the Filed stage are very unlikely events, it might be that not enough failures were provided to the algorithms to successfully learn a distinction between outcomes.

The out-of-sample comparison is visualized in figure 4.4, which shows the trade-off between specificity and sensitivity using so called receiver operating characteristic curves (ROC). The better an algorithm solves this trade-off, the closer the curve comes to the upper-left corner and the higher is the area under the curve, the AUC. In congruence with the previous results, the BART ROC curves lay slightly above RF and C5.0 for PI, PII and PIII, while PROBIT and DT perform worse for all data sets.[8] In

[8]Note that in the Filed data set the RF graph lays over the BART graph because the ROC curve values are derived from only one validation set and not from the average over ten

Figure 4.4: ROC curve of out-of-sample classification results



(a) PI (b) PII

(c) PIII (d) Filed

the next section the best-in-class algorithm is compared to two common evaluation methods, one based on historical data (Girotra et al., 2007), and the other on backward/forward probabilistic regression, which is a mainstay of statistical analysis for classification problems involving binary outcomes such as success and failure.

4.5.2 Supervised Learning vs. Common Estimation Approaches

The BART algorithm with internal missing value imputation and no additional feature selection delivered for each data set the best average out-of-sample performance in terms of AUC. To highlight the performance of our best-in-class algorithm we compare its performance to two rather crude methods that are easily used to estimate project outcomes, one based on historic success rates the other based on backward/ forward probabilistic regression.

The historical (HIST) method classifies projects as successful if the historic success rate for compounds targeting the same market in the same phase is greater than 50%. The backward/forward probabilistic regression (DISCR) discriminates between successes and failures by first considering the full candidate feature set and then subsequently selecting the most explanatory features using a Bayesian information criterion (BIC). BIC is used to evaluate how well a model explains the data while staying as parsimonious as possible. The better a model solves this trade-off, the higher is its BIC value. The procedure optimizes the BIC value by adding or removing features to a probabilistic regression, which means it finds the model that explains the data best without including too many parameters. The procedure stops when neither adding nor removing features contributes to the BIC of the model. For each feature of the final model a coefficient, its standard error, and its p-value are estimated. Note that feature inclusion does not necessarily imply significance of the estimated coefficient. The model coefficients are applied to the observations of the

validation sets. The ROC curves are associated to the AUC values reported in Appendix C.2 table C.6.

validation set, resulting in classification values that can be compared to the true outcomes.

Result 2. *BART shows the highest accuracy for classifications in phase PI, PII, and PIII followed by DISCR and HIST. HIST shows the highest accuracy in phase Filed.*

The comparison between true values and estimated values is reported via confusion matrices of which we provide an example in figure 4.5. Since we are comparing the performance of three methods on four data sets, we obtain twelve confusion matrices whose results appear in figure 4.6. It turns out that the SL-algorithm classifies the outcomes of clinical research phases (e.g., success or failures) with an accuracy of at least 86% (PI = 86.6%; PII = 91.1%; PIII = 85.8%). The HIST method is markedly less accurate (PI = 65.8%; PII = 69.0%; PIII = 70.6%). The DISCR is better than HIST, but is still less accurate than the best-in-class supervised learner (PI = 78.3%; PII = 82.8%; PIII = 78.5%). The AUC measure of each method confirms the conclusions drawn from evaluating the accuracy over the first three data sets. At the filed stage the BART algorithm is not able to distinguish between failures and successes which is due presumably the low number of failures at this stage. In this instance the HIST method seems to give the best classification results and at the same time it requires the least complex set-up. For all data sets we visualize the differences between the classification values of ML and HIST in figure 4.7. Across the three clinical phases, SL-classification values of successes and failures appear more representative of their actual distributions than the classifications based on historic success rates. The same is reflected by the mean classification values. Again, in the filed stage HIST captures failures more reliably.[9]

In summary we find that, among all methods considered, the share of correct outcome classifications is highest for BART across clinical stages

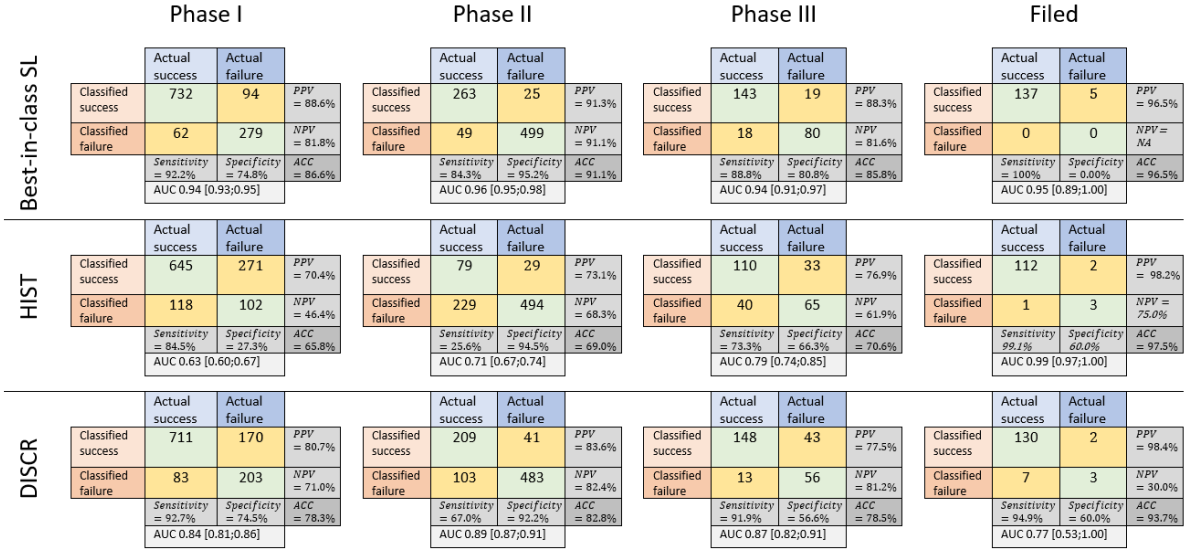
[9]We report a similar graphical comparison between the classification values of ML and DISCR in figure C.2 of the chapter's Appendix C.2.

Figure 4.5: Confusion matrix template

	Actual success	Actual failure	
Classified as success	$\sum True\ positive$	$\sum False\ positive$	$Positive\ predictive\ value\ (PPV) = \frac{\sum True\ pos.}{\sum True\ pos.. + \sum False\ pos.}$
Classified as failure	$\sum False\ negative$	$\sum True\ negative$	$Negative\ predictive\ value\ (NPV) = \frac{\sum True\ neg.}{\sum True\ neg.. + \sum False\ neg.}$
	$Sensitivity = \frac{\sum True\ pos.}{\sum True\ pos. + \sum False\ neg.}$	$Specificity = \frac{\sum True\ neg.}{\sum True\ neg. + \sum False\ pos.}$	$Accuracy\ (ACC) = \frac{\sum True\ pos. + \sum True\ neg.}{All\ outcomes}$
	Area under the receiver operating curve (AUC): The closer to one, the better the model solves the trade-off between false positives and false negatives; lower and upper 95% confidence interval in brackets.		

(PI, PII, PIII). In the next section we therefore apply it to predict the outcome of projects whose clinical research status was classified as on-going in our sample.

Figure 4.6: Confusion matrices for each clinical phase and three classification methods



4.5.3 Predicting Outcomes in the Development Pipeline

We use the trained BART algorithm to predict project outcomes of the current pipeline. Predictions are difficult to directly verify since the development of a new molecule usually takes several years. Therefore, it is crucial that the training set and the current pipeline, that acts as new input data after the training phase , share the same properties with respect to missing values and evaluated features. Since we observe that current projects contain fewer missing data points than historical completed projects on which the algorithm is trained, it could potentially induce biased predictions. To reduce the chance that the difference of missingness in the data influences prediction results, we first impute labeled and unlabeled data separately using a 5NN algorithm. Then, the algorithm is trained on the complete set of labeled data and consequently used to predict outcomes of on-going projects.

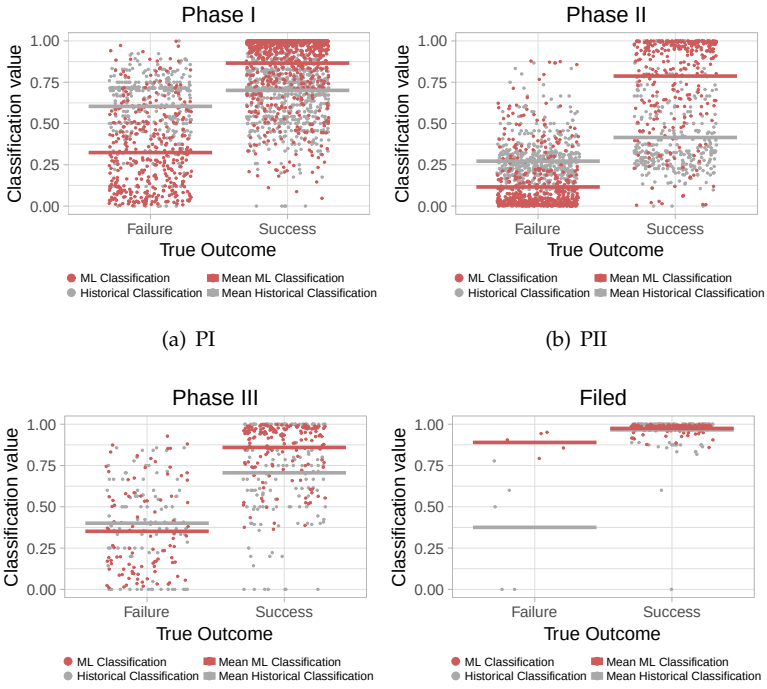
Our project database contained 4,789 projects engaged in various phases of clinical research (PI = 1,858; PII = 2,372; PIII = 559) whose success or failure in the current stage were not yet known and thus predicted.[10] Figure 4.8 shows the predicted success rates for each phase and for all phases combined. Confidence intervals are shown by the black bars, whereas orange tics show the weighted average of the historic phase success rates derived from a meta-analysis of four studies that spanned more than 43,000 total observations from 2000 until 2018 and (Wong et al., 2018; Evaluate Ltd., 2018; Hay et al., 2014; Thomas et al., 2016). The single success rates from which the reported weighted average values are formed are reported in table C.7 of Appendix C.2.

Result 3. *Estimated success rates of the current product pipeline are slightly higher than historical success rates from a meta-analysis of recent studies.*

Comparing the SL success rate predictions of current projects with the success rates from the meta-analysis, we find Phase I rates to be higher

[10] We refrain from predicting outcomes in the Filed stage because of the detected overclassification of successes.

Figure 4.7: Best-in-class SL and HIST classification values separated by phase and true outcome



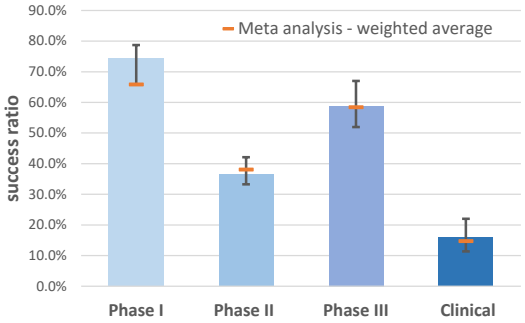
(a) PI

(b) PII

(c) PIII

(d) Filed

Figure 4.8: Predicted success ratio of current project pipeline



for SL (74.3%) vs. meta (65.8%)[11] whereas they are lower for Phase II (SL = 36.4% vs. meta = 38.1%). For phase III, they two almost match (SL = 58.9% vs. meta = 58.4%). Overall, predictions for the current pipeline are slightly more optimistic than historical success rates would suggest (SL: 15.9% vs. meta: 14.7%). Higher future success ratios likely translate into lower average development costs, since the number of investigated compounds that is needed for one successful product launch decreases. Using the R&D cost model of Paul et al., 2010, our predicted success rates would imply that capitalized costs per drug lower by $68 million compared to the historical average (SL: $1,661 million vs. meta = $1,729 million).

Despite an overall promising outlook, the persistence of low phase II success rates has long been a concern in the industry. By breaking down phase II and III success ratios, we report which project types are most likely to be successful and thus point to ways how to potentially reduce R&D portfolio risk.

For both phases we report nine different categories (orphan drug status: No, Yes; FDA expedited treatment: No, Yes; Validation of Mechanism of action (MoA): No, Yes other indication, Yes same indication; Product marketed in other indication: No, Yes) and compare the predicted success ratio of each category with the average phase success ratio (see figure 4.9). We find that projects with special regulatory status are more likely to be successful.[12] Further, projects that use a validated MoA are estimated to be less risky, especially if the MoA was validated for the same indication.[13]. Lastly, repositioning existing drugs for new indications continues to be a less risky alternative (Ashburn and Thor, 2004)[14].

[11] We cannot rule out that phase I predictions are affected by the companies' tendency to report potentially fewer risky projects. Since for phase II and phase III project reporting becomes mandatory (Zarin et al., 2016), predictions in these phases are only driven by characteristics of the current project pipeline.

[12] Success ratio PII [PIII] – no orphan drug: 0.30 [0.54]; PII [PIII] – orphan drug: 0.63 [0.66]; PII [PIII] – not expedited: 0.33 [0.56]; PII [PIII] - expedited: 0.57 [0.73].

[13] Success ratio PII [PIII] – MoA not validated: 0.31 [0.47]; PII [PIII] – MoA validated different indication: 0.36 [0.60]; PII [PIII] – MoA validated for same indication: 0.54 [0.72].

[14] Success ratio PII [PIII] – Not marketed in other indication: 0.33 [0,54]; PII [PIII] – Marketed in other indication: 0.58 [0,81].

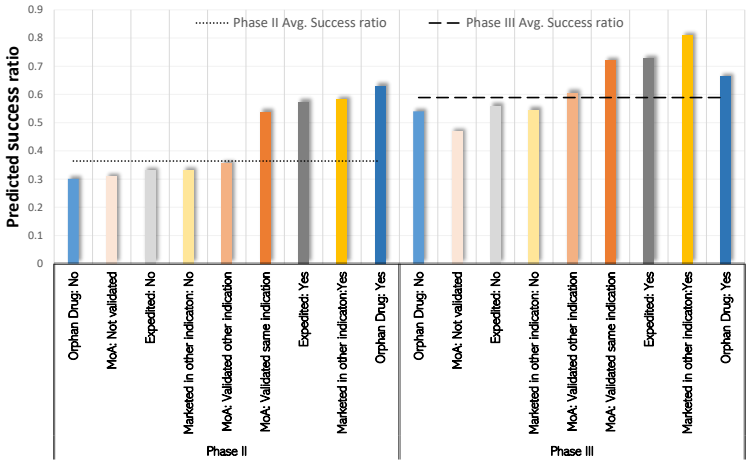
Our prediction results suggest that we will see more projects advancing from phase II to phase III that have obtained special regulatory treatment, rely on validated MoAs, or are repositioning drugs. Since these categories are also associated to above average success rates in phase III, it is likely to see more of them marketed in the future.

Result 4. *Estimated success rates in Phase II and III are higher for projects that have obtained special regulatory treatments, rely on validated MoAs or are repositioning drugs. Estimated success rates vary across indications and technological fields.*

When targeting indications of novel molecules, assessing their risk profile is vital. Table 4.6 shows how the predicted success ratios for current phase II and phase III projects vary across indication and technology categories (see also figure C.3 and C.4 in Appendix C.2 for a visualization of success ratios split only by technology or indication). The riskiest indications for phase II are "Hepatic & biliary" (success ratio: 19.0%) for small molecules and "HIV & related conditions" (28.6%) for biotechnology. For phase III, indications with the lowest predicted success ratio are "Psychiatry" (33.3%) for small molecules and "Cardiovascular" (45.5%) for biotechnology.

Estimating success rates for projects that have not yet reached the market is relevant for a multitude of stakeholders in the pharmaceutical industry. First, the methodology offers managers a tool to assess future trends and to allocate R&D resources such that the risk profile which the firm wants to pursue is achieved. Second, the results offer insights for regulatory bodies who wish to estimate the number of future approvals, and to identify which product categories will become trendy. This analysis can be done looking at historical data and then extrapolating future trends, but a forward-looking projection might complement insights in future dynamics. Third, algorithms could potentially be used to predict the project outcome on the company level, so that assessing the risk of a company's R&D portfolio could offer information prior to acquisitions or the commencement of joint ventures. Finally, predicting the outcome

Figure 4.9: Predicted success ratio of current PII and PIII pipeline by success factors



of ongoing projects could aid the decision of professional investors in which company they want to invest in. For these various stakeholders not only precise estimation of success rates is important but also knowledge about which features impact R&D outcomes. The next section describes the most important features in more detail.

4.5.4 Most Indicative Features

Most previously presented SL-results rely on the evaluation of a long list of candidate features (presented in table 4.2). A natural follow-up question to ask is which features are most indicative for project success at different development stages. We evaluate five different variable selection methods and assess which features are most frequently selected for project classification. To the three feature selection methods used in the SL training procedure (LASSO, RF_SE, BART_IP) we add the backward/forward probabilistic regression (DISCR) that selects vari-

Table 4.6: Predicted phase II and III success ratio per indication and technology

	Phase II		Phase III	
Indication	Conventional	Biotech.	Conventional	Biotech.
Hepatic & biliary	19.0% [58]	- [8]	- [6]	- [1]
Sensory organs	22.7% [44]	41.7% [36]	35.7% [14]	- [7]
Cancer	25.8% [690]	30.8% [464]	59.0% [83]	64.8% [88]
Respiratory	28.6% [35]	33.3% [24]	- [4]	- [7]
Blood	28.9% [38]	53.3% [30]	56.3% [16]	63.6% [11]
Psychiatry	36.8% [57]	- [0]	33.3% [15]	- [0]
Immunology	38.2% [34]	33.3% [51]	62.5% [16]	60.0% [20]
Skin	41.1% [73]	45.8% [24]	92.9% [14]	- [3]
Urinary tract	42.1% [19]	- [10]	- [7]	- [3]
Diabetes	43.8% [32]	75.0% [32]	- [10]	- [3]
Neurology	46.7% [120]	35.7% [28]	43.2% [44]	- [9]
Reproduction	50.0% [18]	- [1]	- [8]	- [1]
Musculoskeletal	53.1% [32]	44.7% [47]	73.3% [15]	66.7% [18]
Cardiovascular	56.8% [37]	45.2% [31]	40.0% [20]	41.7% [12]
Gastro-intestinal	63.6% [44]	37.5% [16]	- [10]	- [8]
Infections	66.2% [77]	32.4% [74]	61.8% [34]	45.5% [11]
HIV & related conditions	- [8]	28.6% [14]	- [6]	- [1]
Phase-technology average	36.30%	36.70%	57.80%	60.50%
Phase average	36.40%		58.90%	

Table 4.6 shows the predicted pipeline success ratios for PII and PIII projects across indications split by Conventional- and Biotechnology. Number of observations in brackets. The success ratio for indication/technology pairs that contain less than 10 projects is omitted. Average success ratios are weighted by the number of projects in each category.

ables based on BIC and a fast and frugal decision tree (FFT) (Phillips et al., 2017). FFT is a tree based classification approach mimicking human decision-making by establishing a small set of binary choice rules. On average, the highest number of features was selected in PII (13.6) followed by PI (11.6) PIII (7.8) and lastly Filed (4.8), suggesting that the relevant features in late stages can be narrowed down to a limited set.[15]

Result 5. *The set of indicative features change with respect to the development stage. The "indication phase success ratio" and the "phase success rate by MoA" are the most selected features across stages.*

Table 4.7 shows for each data set the features that have been selected by the majority of algorithms (three or more). We refer to these as the most indicative features, as most methods pick up an association between these features and the project outcome. Each reported feature relates to a sign (+,-,o) depending on whether the correlation between the feature and project success is positive, negative or ambiguous (for categorical features). For example, the plus sign at "indication phase success ratio" indicates that the higher the historical success ratio of the same indication in the same phase, the higher is the probability of success. The "indication phase success ratio" together with the "phase success rate by MoA" are two features that remain indicative for project success throughout the development process. Other success ratios based on historical data are also indicative such as the "market success ratio" for early stages and the "company success ratio" for phase II and III. Clinical trial costs and orphan drug status relate positively to stage-specific success in Phase I and II, possibly because it signals a companies commitment to pursue project development. On the other hand, company experience in terms of market and product ("Own similar products (count)" and "R&D (count)") have a negative effect on the stage-specific success in early phases presumably due to the increased opportunity costs of

[15]Number of selected features by data set and method. PI: Lasso 25, RF_SE 19,BART_IP 8, DISCR 12, and FFT 4; PII: Lasso 16, RF_SE 14,BART_IP 10, DISCR 15, and FFT 3; PIII: Lasso 10, RF_SE 6,BART_IP 7, DISCR 12, and FFT 4; Filed: Lasso 2, RF_SE 3,BART_IP 4, DISCR 13, and FFT 2).

capital a company faces. Projects are evaluated against other projects in a company's pipeline and potentially replaced if their relative value results questionable. Overall, we find that for all clinical phases the most indicative features relate to a broad variety of feature dimensions, undermining the importance of a diversified candidate feature set to predict the outcome of pharmaceutical projects.

In this section we show which features are mostly selected across variable selection methods to classify outcomes of pharmaceutical projects. It might be tempting to think about these features as causally contributing to the probability of a project's success. Yet, a proper claim that regards any kind of causal relationship requires a set-up that would exceed our analysis. The above-mentioned associations hopefully stimulate further research in this direction.

Table 4.7: Most indicative features across methods

Feature category	Phase I	Phase II	Phase III	Filed
Product	Clinical trial cost [5,+]	Clinical trial cost [5,+]		
Product	Therapy type [5,o]		Clinical trial results [5,+]	
Market	Indication phase success ratio [4.+]	Indication phase success ratio [5,+]	Indication phase success ratio [5,+]	Indication phase success ratio [3,+]
Market	Market success ratio [5.+]	Market success ratio [5,+]		
Market	Indication level [3,o]	Market inequality [3,-]	Indication level [3,o]	
Company	Own similar products (count) [3,-]	Company success ratio [4,+]	Company success ratio [4,+]	
Company	Market experience (count) [3,-]	R&D (count) [3,-]		
Company		Market experience (count) [3,-]		
Other	Phase success rate by MoA [4,+]	Phase success rate by MoA [5,+]	Phase success rate by MoA [5,+]	Phase success rate by MoA [5,+]
Other	Orphan Status [3,+]	Orphan Status [3,+]	Expedited status [3,-]	

Table 4.7 shows the features that were selected by the majority of selection methods. The number in brackets shows the number of methods which selected the feature. Positive (+), Negative (-) or ambiguous (o) relationship between feature and likelihood of phase success.

4.6 Final Discussion and Concluding Remarks

We have evaluated the performance of various supervised learning algorithms to classify the clinical success of pharmaceutical projects as they progress through the phases of development. Our best-in-class algorithm is substantially more accurate than traditional methods based on historic success rates or regression analysis. Given the limitations of traditional approaches, we predict the success ratio for the current project pipeline using the best-in-class method. On aggregate, our predictions are more optimistic than historical benchmarks, even though low phase II success ratios continue to persist.

Evaluating current projects via supervised learning provides actionable insights for companies who wish to improve their R&D portfolio in two ways. First, by analyzing the risk profile of different project types. For example, repositioning an existing drug is suggested to lower the risk of the R&D portfolio leaving room for developing new molecular entities. Second, by directly assessing the success prediction of a single project under development, companies can take specific measures. Through a detailed description of our algorithm training procedure and features which we find most indicative for outcome classification, our work hopefully inspires similar research that aim at improving the risk assessment of pharmaceutical projects under development.

Even though machine-learning algorithms are far from perfect at predicting outcomes in a complex industry, they provide a useful tool to help pharmaceutical companies to adjust the profile of their projects and to potentially raise their odds of clinical success (Feijoo et al., 2020). Other important applications that require improved risk-assessment of developmental drugs include to derive more accurate valuations for companies seeking to raise capital, for Initial Public Offerings, or for R&D assets being transacted. Supervised learning tools could also be used to help assess the risk inherent in a company's pipeline and thus provide a check on management claims about its prospects. At the macro level, they could aid assessing the amount of innovation embedded in the industry pipeline.

Chapter 5

Stock Market Reactions to Product Innovation: Evidence from the Pharmaceutical Industry

5.1 Introduction

On March 21st 2019, the pharmaceutical company Biogen publicly declared to discontinue its advanced clinical trials on *aducanumab*, its top candidate drug that aimed at a breakthrough to treat patients diagnosed with Alzheimer's disease. Consequently, the shares of Biogen fell by 29.2%, reflecting a revision of shareholders' expectations on future revenue streams generated by *aducanumab*.

The stock price correction that followed Biogen's announcement is not uncommon for industries in which product innovation is not only imperative for long-term success but also associated to costly, lengthy, and risky development processes (Mazzucato and Tancioni, 2008). Stock market participants are aware of the importance of product innovation for future cash flows and thus closely monitor news on companies' development pipelines. In case innovation announcements are at odds with

prior investor sentiment, severe reactions on the company stock price can be the consequence. These, potentially disruptive, price jumps foster the need of corporate innovators and financial institutions to better understand the underlying patterns that guide them.

In this work we explain company stock price reactions that follow from product innovation announcements by two general effects: a probability effect, that reflects the sudden change of the perceived probability that a product will be launched for a specific market, and a portfolio effect, that reflects the share of discounted cash flows expected from the product once it is marketed relative to the total expected cash flows of the company. Our findings confirm the common intuition that products whose success probability (portfolio importance) is high, show lower (higher) market reactions after product-related innovation announcements. Yet, establishing a sensible measure for both effects that can be used in empirical settings is not trivial.

We approximate both effects using a detailed database on pharmaceutical products, their characteristics, product-related company announcements, and subsequent stock price reactions. To estimate the probability effect, we gauge market sentiment on the product success probability in a given therapeutic market via different supervised learning (SL) methods. The measure on the portfolio effect is obtained directly from date-, product-, and company-specific net present value (NPV) data, that is difficult to calculate directly for large-scale empirical applications. We find that both measures contribute to better understand market reactions that follow product innovation announcements.

The bio-pharmaceutical industry is known to be one of the most innovative industries (H. Grabowski and Vernon, 1994). For companies operating in such industries it is vital to keep information on product innovation confidential which makes the relationship between innovation announcements and financial market reactions difficult to assess. In contrast, regulatory bodies of the pharmaceutical industry (e.g., the Food and Drug Administration (FDA) in the United states and the European Medicine Agency (EMA) in the EU) enforce companies to disclose

information on drugs under development. Since 2007 and 2015, respectively, FDA and EMA require that the outcome of clinical trials which have passed "Phase I" must be disclosed irrespective of their content (see Zarin et al., 2016; European Medicines Agency, 2014; US Food and Drug Agency, 2007). Besides being legally obliged, pharmaceutical companies are found to voluntarily disclose important news on their project pipeline, positive and negative ones, possibly to signal innovation transparency to investors (Enache et al., 2018).

Another characteristic that makes the pharmaceutical industry an interesting candidate for investigating the relationship between innovation announcements and stock market reactions is that the innovation process is standardized and thus announcements can be categorized by their advancement within the process. The US development process requires of candidate drugs the completion of multiple sequential stages: after a pre-clinical stage, that involves the identification of a candidate drug and the assessment of its toxicity, an investigational new drug (IND) application needs to be approved by the FDA. After the IND approval, clinical research on human subjects commences and commonly undergoes three distinct phases. In "Phase I" the candidate drug is tested on safety and dosage using small-scale clinical trials. Once the candidate drug has passed "Phase I", efficacy of the treatment is evaluated in "Phase II" clinical trials and reassessed in "Phase III" using a larger sample of patients. If the endpoints of clinical trials are met and the drug fulfills its safety requirements, companies file a new drug application (NDA) to the FDA - and possibly to other regulatory authorities of different geographical markets - which evaluate the complete history of clinical trials and either approve or refuse its market launch. We refer to this stage as "Filed". In case of a negative FDA response in the Filed stage, companies may run additional clinical trials to support the safety and efficacy of the candidate drug and refile afterwards. In case of a positive FDA response, the company is entitled to market the product, however drug safety monitoring continues also in the post clinical stage.[1] In the pharmaceutical

[1]Figure C.1 in the supplementary material shows the pharmaceutical development process. The process regards the United states but other western countries follow similar pro-

industry that spends an average $19 million on each clinical trial (Moore et al., 2018), but faces long development times (Joseph A DiMasi and H. G. Grabowski, 2007) and high failure rates (Wong et al., 2018), innovation related announcements should have a substantial impact on stock price reactions. If these price reactions can be explained by some general effects that plausibly hold up across business environments, then we should carefully analyze them to better understand the forces that drive market reactions around product innovation announcements.

The chapter proceeds as follows: in the next section we review studies that employ stock market data of pharmaceutical companies to study the impact of firm announcements. Motivated by previous empirical findings, we set up a simple company stock reaction model in section 5.3 from which we derive our hypotheses that market reactions on innovation announcements can be described by a "portfolio" and a "probability" effect. In section 5.4 we provide details on how to measure both effects and how we relate them to market reactions using an event study methodology. Section 5.5 describes our data set and highlights some descriptive statistics and preliminary evidence. In section 5.6 we present our results in which we relate market reactions to portfolio and probability effects using different regression specifications. We discuss our findings and potential implications for innovative companies in the conclusive section 5.7.

5.2 Literature Review

The study of the relationship between company announcements and subsequent market reactions enjoys a long tradition across economic fields from entrepreneurial- and management research to applications in accounting and finance. A common property of such event studies is that the type and precise date of the announcement is retrieved ex-post so that a direct effect on stock market valuation can be estimated. Because

cesses. Differences across these countries are mainly found in the post-clinical stage due to different pricing and reimbursment regulations.

of the vast literature on event studies across sectors, we chose to focus on articles that use with the bio-pharmaceutical industry a comparable empirical setting.

Over the last two decades, the bio-pharmaceutical industry has increased its transparency concerning R&D related information. Consequently, researchers have started to investigate, whether the disclosure of R&D related information is i) value-relevant for the company and ii) improving the decision-making of investors. Dedman et al., 2008 show that positive R&D announcements in the UK biotechnology sector are associated to higher market reactions than earning announcements. Similarly, Shortridge, 2004 finds that reported R&D expenses are associated to higher company valuations for companies with an above-average number of new drug applications. Both studies suggest that R&D disclosure is considered by investors, while Hao et al., 2017 add that information on R&D also improves their decisions. They show that the forecast accuracy of financial analysts improves for companies that disclose more details about their development pipeline.

Disclosing the composition of the development pipeline may aid financial investors to assess the value of a company (Ely et al., 2003), via the use of a real-option valuation approach (Kellogg and Charnes, 2000) or by calculation of the probability adjusted discounted cash flows (Girotra et al., 2007). Both valuation approaches are frequently used among financial practitioners and share the underlying assumption of our contribution, that shifts in stock prices can be explained by changes in investors' expectations. Girotra et al., 2007 proxy investor expectations by market specific historical success rates and find that the drop in a company's valuation after a negative event is associated to a drop in the company's probability to obtain at least one drug launch in a specific therapeutic market. Kellogg and Charnes, 2000 use a real-option valuation approach to estimate investor expectations on the value of a single biotechnology company. They find that historic success probabilities are sufficient to explain market valuations before Phase II, yet investor expectations are not reflected by average success rates for later development stages. We di-

rectly address this obstacle by gauging success probabilities on product level using a supervised learning procedure.

A substantial body of research uses event studies on pharmaceutical R&D announcements to address research questions like: "what effect does variable of interest (VoI) have on companies' stock market performance pre-, at, or post-announcement?".[2] To provide a detailed overview on previously conducted research of this form, table 5.1 summarizes studies according to sample composition, sample period, VoI, event interval, and main findings. The entries are ordered by sample size. The last row shows our contribution to simplify a comparison with the related literature.

Many of our reviewed studies report that negative events cause higher absolute market reactions around the announcement date than positive ones (see A. Sharma and Lacey, 2004; Sarkar and Jong, 2006; Rothenstein et al., 2011; Hwang, 2013). Different reasons for this market reaction asymmetry have been proposed, for example, overconfidence in future successes and financial certainty of negative news versus financial ambiguity of positive news (A. Sharma and Lacey, 2004). Another often explored topic is how the stage of product development at which an announcement occurs relates to stock market reactions. Dedman et al., 2008 find significant market reactions for late stage positive announcements in UK biotech firms, whereas Ely et al., 2003 find significant announcement effects for the US market only in Phase II. Focusing on negative events, Urbig et al., 2013 report higher negative reactions for late phase failures, which is amplified for companies with higher R&D costs per employee. Conversely, Sarkar and Jong, 2006 find that news induced market reactions diminish over the approval process.

The difference in these findings could be due to small sample sizes that are frequently restricted to certain categories and thus might not be comparable, e.g. large bio-pharmaceutical companies (Hwang, 2013) or cancer drugs (Rothenstein et al., 2011). Another potential issue is that

[2]For the classification of the reviewed literature, we define "at-announcement" as -5/+5 days around the news. Studied market reactions before and after this interval are classified as "pre-announcement" and "post-announcement", respectively.

Table 5.1: Selected literature using pharmaceutical R&D events ordered by sample size

Author	Sample composition	Period	VoI	Event interval	Main Finding
Xu, 2006	3420 (only positive)	1980-2003	R&D progress	post event	Post announcement drift and volatility decrease with R&D progress.
A. Sharma and Lacey, 2004	447 (Filed)	-	FDA announcements	at event	Return asymmetry between positive and negative news.
Dedman et al., 2008	234	1990-1998	Earning announcements, R&D announcements	at event	R&D announcements have a higher impact than earning announcements.
Ely et al., 2003	193	1988-1998	Development stage	at event	The relationship between R&D expenses and market value is positive and significant for companies with high drug portfolio potential.
Sarkar and Jong, 2006	189 (Filed)	1990-2001	FDA announcements	at event	Market reactions decrease over the approval process.
Urbig et al., 2013	148 (only negative)	1994-2008	companies R&D focus	at event	High debt to equity ratio, industry experience, and R&D focus are associated to higher negative price shocks for late stage failures.
Girotra et al., 2007	132 (Phase III)	1994-2004	Phase market success rate	at event	Market reaction lower for companies with higher number of products in development for the same therapeutic market.
Rothenstein et al., 2011	109	2000-2009	Positive/ negative news	pre-, post event	Increase (decrease) in stock price before positive (negative) clinical trial. announcement.
De Carolis et al., 2009	105 (only negative)	1992-2003	organizational characteristics	at event	Stronger market reaction for companies with lower technological capabilities.
Donovan, 2018	100 (only positive Filed)	2013-2017	firm and drug attributes	pre-, at, post event	Positive (negative) association between price shock and novelty of drug (company market capitalization).
Pérez-Rodriguez and López-Valcárcel, 2012	46 (one company)	1994-2008	Scientific and regulatory news	at event	Higher market reaction for regulatory news than for scientific news. No spillover effects on competitors.
Panattoni, 2011	37	1993-2007	generic entry	at event	Negative (positive) market impact for brand (generic) company upon generic entry approval before end of exclusivity period.
Hwang, 2013	24	2011-2013	clinical trial announcements	at event	Return asymmetry between positive and negative news
This study	703	2017-2018	Product portfolio importance and success probability	at event	Market reactions can be described by a portfolio effect and probability effect.

samples that start in years for which companies were not required to announce failures and successes alike, might suffer from publication bias (Joos, 2003). We hope to overcome these limitations by using a large sample of more than 700 innovation announcements that occurred between January 2017 and December 2018, when mandatory disclosure policies were already in place. It contains both positive and negative outcomes in different research stages for a heterogeneous set of companies, products, and therapeutic areas.

The heterogeneous findings reported in the literature on how markets react to product related announcements motivate us to formalize a general company stock reaction model. It can serve to put prior empirical findings into context and we use it to derive the hypotheses tested throughout this chapter.

5.3 Company Stock Reaction Model and Hypothesis Derivation

The market value of a company is the price at which financial investors agree to trade the companies stock. Upon the announcement of positive news, investors perceive that a company's value has increased, hence increase their demand and thus its stock price rises. Contrary, negative news lead investors to perceive the company as less valuable which consequentially results in a decrease of the company's market valuation. Since investors try not to miss a favorable trading opportunity, we assume that market reactions after announcements usually happen fast. This is a general assumption in event studies which implies that announcements can be related to subsequent changes in the market value. One traditional approach to numerically derive a company's market value is the discounted cash flow (DCF) model (DeFusco et al., 2015). Our company stock reaction model is constructed in this spirit, assuming that the market value of a company MV_c is proportional to the expectation in future DCF that result from all projects p, where $|p|$ is the total number

of projects of which company c can expect future revenues (see equation 5.1). The expected DCF for projects under development is weighted with their probability of success $Prob(S_p)$. For already completed projects such as launched products $Prob(S_p)$ is equal to one. In our target setting, a project can be considered a pharmaceutical product for a specific therapeutic market (e.g. aducanumab for Alzheimers disease).

$$MV_c \propto E(DCF_c) = \sum_{p=1}^{|p|} E(DCF)_p = \sum_{p=1}^{|p|} Prob(S_p) \cdot DCF_p|S_p \tag{5.1}$$

Since each project has to pass sequentially through the development stages ($k = \{1 : \text{Phase I}, \quad 2 : \text{Phase II}, \quad 3 : \text{Phase III}, \quad 4 : \text{Filed}\}$), $Prob(S_p)$ in equation 5.2 is the product of the success probabilities at each stage k which it still has to pass.

$$Prob(S_p) = \prod_{i=k}^{|k|} Prob(S_{pi}) \tag{5.2}$$

We assume that in case of a negative event related to project p^* the success probability decreases ($Prob(S_{p^*}|event = neg) < Prob(S_{p^*})$), while in the case of a positive event, e.g. the project advances one stage, the probability of a success increases ($Prob(S_{p^*}|event = pos) > Prob(S_{p^*})$).[3] The relative change in the market value of a company, that results from news referring to a specific project, can thus be written as equation 5.3.

[3]For simplicity, we assume no spillover effects of the probability of success across projects in the same company (i.e. $Prob(S_p|\Delta Prob(S_{p^*})) = Prob(S_p) \quad \forall p \neq p^*$). On the other hand, the model does not restrict events to announcements of the focal company itself but can account for all sorts of news which indirectly affect the success probability of a product, such as news of substitute or complementary products of third parties. Given the nature of our event data, we focus in our empirical application on announcements directly associated to the product under development.

$$\Delta MV_c \propto \frac{[Prob(S_{p^*}|event) - Prob(S_{p^*})] \cdot DCF_p^*|S_{p^*}}{\sum_{p=1}^{|p|} Prob(S_p) \cdot DCF_p|S_p} = \\ = \Delta Prob(S_{p^*}) \times \frac{E(DCF)_{p^*}}{E(DCF_c)} \tag{5.3}$$

The change in market value thus depends on two main effects. First, the more an event changes the probability that a specific project will generate cash flows $\Delta Prob(S_{p^*})$, the higher should be the market reaction. We call it the "probability effect". Second, the higher the relative importance of a project in a companies portfolio in terms of future discounted cash flows $\frac{E(DCF)_{p^*}}{E(DCF_c)}$, the higher should be the market reaction. We refer to this effect as "portfolio effect". For the sake of deriving testable hypotheses on both effects, we treat them as separable measures, assuming that the relative product portfolio importance is conditional on product success. We address the relationship of both measures in the result section 5.6.

Exact quantification of the probability effect would require measuring the projects' perceived success probability right before and right after the event, which is, given common data sources, not feasible. We thus can only provide an identification attempt of the probability effect by estimating the probability of success for project p^*. In case the success probability of a project is already high, we expect that the average event will not drastically change the success probability. As the project advances through the innovation process and increases its success probability, uncertainty about the project is resolved (Sarkar and Jong, 2006) which should be reflected in lower market reactions. For example, after a positive event that affects a project whose success probability is already high, investors will only marginally update their beliefs about its success and thus the market reaction reads low. Yet, the opposite could be argued for negative events. Negative events which occur late in development (i.e., that affect pharmaceutical products with a relative high probability of success) might be associated to higher market reactions since investors have already built up expectations (Urbig et al., 2013). But if the severity of reported negative events changes during the development process,

then one should not make the mistake to equate negative project announcements with the termination of product development.[4] In our data set, we find that the majority of projects associated to negative late-stage events do not terminate but rather imply a prolonged continuation and thus carry with them some residual value. We therefore assume that the announcement induced probability revision, and thus the related market reactions, decrease in the probability of the product being launched. We formulate *Hypothesis 1* accordingly:

Hypothesis 1: The higher the success probability of a project, the lower are market reactions that follow its innovation announcements.

The probability effect also provides an explanation to the often reported return asymmetry between positive and negative events. Assume that positive events induce, on average, a lower change in the probability of success than negative events. Then, average returns of positive events should be lower, as well. In the pharmaceutical context this assumption seems reasonable since positive events are mostly related to a product passing from one stage to the next, implying an incremental change in the success probability. Negative news, on the other hand, are not restricted to an incremental change in the success probability and thus average market reactions tend to be higher.

Besides the probability effect, market reactions should depend on a portfolio effect. As previously suggested, the more important is the project for the project portfolio of a company, the higher is the assumed market reaction after project specific news, independent whether news are positive or negative. We thus formulate *Hypothesis 2* as:

Hypothesis 2: The higher the relative importance of a project in the company's product portfolio, the higher are market reactions that follow its

[4]Mathematically expressed, we cannot generalize that $Prob(S_{p^*}|event = neg) = 0$ assuming heterogeneity in negative events. We only hypothesize that $\Delta Prob(S_{p^*}) > \Delta Prob(S_{p^\circ})$ for $Prob(S_{p^*}) < Prob(S_{p^\circ})$.

innovation announcements.

The company stock reaction model is kept general to be applicable across corporate environments that relate to innovation announcements on the project level. Projects relate in our empirical application to pharmaceutical products under development for a given therapeutic market but can more generally be thought of separable company ventures whose value is uncertain. Similarly, the events considered by the model are not limited to official communications of companies associated to the project but might contain any public announcement that affects the project's perceived probability of success or its expected portfolio importance, for example, news on rival companies, opinions of expert panels, or announcements about R&D collaborations. The generality of the model comes at a cost, namely to find suitable approximations for both the probability and the portfolio effect. The three parts of the next section describe how we approximate success probabilities, measure portfolio importance, and quantify the market reactions that relate to both effects.

5.4 Empirical Strategy

In our main analysis we associate market reactions after project-specific innovation announcements to the success probability and the portfolio importance of these projects. In this section we discuss our approaches to estimate market reactions (section 5.4.1), success probabilities (section 5.4.2), and portfolio importance (section 5.4.3) combining financial data with the information obtained from a large bio-pharmaceutical data base.

5.4.1 Measurement of Market Reactions

We are interested in how innovation announcements affect the stock price of companies. Since financial markets adjust fast to new information, we collect product related innovation announcements for which we know

the exact date.[5] For all announcements we calculate daily logarithmic stock returns using adjusted closing price data from *Yahoo! Finance*. In case a company's stock trades at multiple exchanges, we choose the market with the highest average trading volume to avoid return anomalies due to market frictions. To extract the part of the stock price adjustment that is due to the respective announcement, we use a traditional event study approach (MacKinlay, 1997). The daily return of company i at event date τ, denoted by $R_{i\tau}$, is assumed to contain an expected component $E[R_{i\tau}]$ and an event component called abnormal return $AR_{i\tau}$. To obtain the abnormal return one subtracts the expected return of company i at time τ (see equation 5.4).

$$AR_{i\tau} = R_{i\tau} - E[R_{i\tau}] = R_{i\tau} - \hat{\alpha}_i - \hat{\beta}_i R_{m\tau} \tag{5.4}$$

The expected return is given by the market return $R_{m\tau}$ of market m at event date τ weighted by the estimated coefficient that relates market returns to company returns $\hat{\beta}_i$ and added to the idiosyncratic company return component $\hat{\alpha}_i$. Both parameters are estimated in an ordinary least squares regression over a 100-day estimation window around the event date $[\tau - 51, \tau - 2] \cup [\tau + 2, \tau + 51]$ using company and market returns.[6] To account for the specific characteristics of the bio-pharmaceutical industry, we choose the *ARCA Pharmaceutical Index* as reference market. Since the choice of the reference market directly affects the estimated abnormal returns and thus might alter our findings, we calculate two additional AR measures based on market indices that are more disperse across sectors and geographical markets (*Dow Jones Industrial index* and *Ishares MSCI World*).

Our event data reports the date of the market impact alongside with the date of the company announcement, which mostly coincide. Yet, in some

[5] While there exists some evidence that events in the industry are subject to insider trading (Rothenstein et al., 2011), the market reaction at the announcement date is usually substantial so that we infer that the general public is kept unaware of innovation results and thus changes in company value happen with the announcement.

[6] There exist various ways in the literature to estimate $E[R_{i\tau}]$ including the capital asset pricing model (Black et al., 1972) that assumes no idiosyncratic effects ($\alpha = 0$) or factor models that include additional company specific factors (Fama and French, 1993). We use a linear regression with constant to capture company-specific effects and, at the same time, stay as parsimonious as possible in the estimation.

cases the announcement date does not fall on a trading date or the market impact spans two days. To consider this timing effect, we calculate the cumulated abnormal return (CAR) of company i around event date τ, according to MacKinlay, 1997, as:

$$CAR_{i\tau} = \sum_{t=\tau-1}^{\tau+1} AR_{it} \tag{5.5}$$

The range of dates over which the abnormal returns are summed is called the event window ($[\tau - 1, \tau + 1]$). The event window is commonly symmetrically chosen around the event date such that CAR includes the return of the event but the impact of returns not associated to the event stays limited. The average cumulated return ($\overline{CAR}$) over a set of N events is then simply given by averaging over CARs of individual events (see equation 5.6).

$$\overline{CAR} = \frac{1}{N} \sum_{n=1}^{N} CAR_n \tag{5.6}$$

The $\overline{CAR}$ for different groups, for example projects with high estimated success probability or high portfolio importance, can then be compared using parametric and non-parametric test statistics. To test our hypotheses, we relate the CAR around each innovation announcement to the success probability and the portfolio importance of its associated product using different regression specifications. Since the CAR measure is influenced by the chosen specifications during its calculation, we additionally use the discrete daily stock return as a more simplistic measure within each regression specification.

5.4.2 Measurement of Product Success Probability via Supervised Learning

To test *Hypothesis 1*, we need a qualified measure for the project-specific success probability which reflects the sentiment of market participants. Since such a measure is difficult to obtain, it is, within the pharmaceutical context, commonly approximated via the historical success rate of

similar products in the same stage of development (Girotra et al., 2007; Kellogg and Charnes, 2000). All products that share the same stage of development and therapeutic market would then be associated to the same success probability, neglecting individual differences between products of which investors are most likely aware. Another implicit assumption is that the perceived success probability of market participants is accurately reflected by historical success ratios.

We aim to loosen these restrictions by proposing a project specific measure of success probability. In chapter 4 we have successfully used supervised learning (SL) methods to classify the outcome of pharmaceutical projects. Here, we build on our insights and approximate investors' sentiment by estimating project success probabilities via an optimized combination of SL methods. These consider multiple product characteristics that are used to determine individual success probabilities. Each used SL algorithm employs a different degree of complexity that relates to the heterogeneous approaches chosen by investors to form probability estimates. All of the SL probability estimates are then weighted according to an optimization function that minimizes the forecast error. This way our probability estimates reflect product-specific characteristics and moreover do not only depend on a single estimation method. The rest of this subsection will detail the SL procedure used to derive the product success probability estimates.

In recent years, supervised learning algorithms have become increasingly popular to classify observations into different known categories. In our context, observations are pharmaceutical products under development for a specific therapeutic market (referred to as projects) which are classified either as successful or unsuccessful by estimating the project's probability of success. Comparable to a human decision-maker, an SL algorithm forms decision rules on how to best distinguish successful from unsuccessful projects based on their characteristics (referred to as features). Features are chosen such that most investors find them easily accessible and decision-relevant based on previous evidence.[7] The feature set contains information on the novelty of the product that relates to

[7] We list the input features in table D.1 of appendix D.1.

its regulatory approval requirements and consequently to its probability of being approved (Cohen, 2005). It further indicates the product's expedited treatment status and its orphan drug status granted by the FDA which is linked to the probability of drug approval (Regnstrom et al., 2010; Joppi, Garattini, et al., 2013). Lastly, it contains the success ratio of products that use similar technology and the success ratio of products in the same therapeutic market; both ratios being practical approximations to estimate a product's probability of success (Joseph A DiMasi and H. G. Grabowski, 2007; Wong et al., 2018). One can think of extending the feature set in many ways which would likely raise the predictive performance of considered SL methods, but would, at the same time, feed the doubt whether the majority of investors i) enjoys access to such a large information base and ii) holistically uses these characteristics to form beliefs about project outcomes. Since the challenge here is to gauge perceived market success probabilities, that are ultimately driven by human decision-makers, we opt for keeping the feature set rather parsimonious.[8]

Expressed mathematically, each project and phase specific success probability conditional on the array of product features $Prob(S_{pk})|\mathbf{X})$ can be estimated by a decision function f that combines product features $\mathbf{X}$ and their importance weighting β. Equation 5.7 shows this general relationship where Φ is a probability mapping function $\Phi : \mathbb{R} \rightarrow [0, 1]$.

$$Prob(S_{pk}|\mathbf{X}) = \Phi[f(\mathbf{X}, \beta)] \tag{5.7}$$

If investors assess the success propensity of a specific project and phase by weighting various success characteristics, then estimating the decision function in equation 5.7 should result in probability estimates that proxy the direction of market sentiment. The functional form of the decision function reflects the way market participants combine single success indicators to derive their success probability estimates. Since we are agnostic about how sophisticated market participants are at forming their decisions, we derive probability estimates using four different models

[8] In fact, if the majority of investors were able to predict outcomes accurately, we would not observe succinct market reactions after innovation announcements in the first place.

that differ in terms of their functional complexity, as summarized in table 5.2.

Table 5.2: Summary of analyzed project success probability estimation methods

Name	**Decision function complexity**	**Average weight**	**Functional form of** $\Phi[f(\mathbf{X}, \beta)]$
HIST	Low	2%	$Prob(S_{mk})$
FFT	Medium	14%	$RF[\mathcal{T}^{\mathcal{M}}(\mathbf{X})]$
LOGIT	High	15%	$(1 + e^{-\mathbf{X}'\hat{\beta}})^{-1}$
BART	Very High	69%	$\Phi[\mathcal{T}_1^{\mathcal{M}}(\mathbf{X}) + \mathcal{T}_2^{\mathcal{M}}(\mathbf{X}) + \ldots + \mathcal{T}_m^{\mathcal{M}}(\mathbf{X})]$

Table 5.2 compares the four project success propensity estimation methods HIST, FFT, LOGIT, and BART with respect to their complexity of employed decision function, the average weight over all stages assigned to each method, and the functional form described in the text.

As mentioned above, the first model uses a simple historical heuristic (HIST) to classify projects. It assumes that the success probability of a project at a given stage can be approximated by the historical success rate of products in the same stage and same therapeutic market m ($Prob(S_{pk}) = Prob(S_{mk})$).

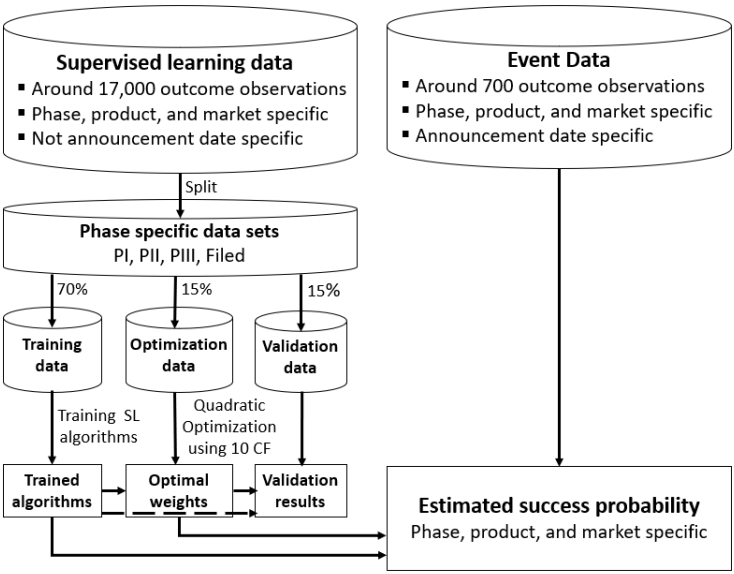
The second model, a fast and frugal tree (FFT, Phillips et al., 2017), considers all potential success factors. The algorithm is inspired by human heuristic decision-making in that it consecutively evaluates only a handful of decision rules. At each decision rule, the algorithm classifies a subset of observations and keeps the rest to be evaluated using the next decision rule. The last rule classifies the remaining observations. The probability $Prob(S_{pk})$ can be described by $RF[\mathcal{T}^{\mathcal{M}}(\mathbf{X})]$, where $\mathcal{T}$ is the tree structure or sequence of rules, $\mathcal{M}$ the decision rules that relate to the

features $\mathbf{X}$, and $RF(.)$ the relative frequency of success at the classifying decision rule.[9]

The third model is a common logistic regression (LOGIT) which linearly weights the set of product features with $\hat{\beta}$, chosen such that the prediction error is minimized. The product success probability is then calculated via $(1 + e^{-\mathbf{X}'\hat{\beta}})^{-1}$.

Lastly, we use a Bayesian additive classification tree (BART, Kapelner and Bleich, 2013), a powerful classifier, that combines Bayesian probability updating with ensemble tree models. The success probability $Prob(S_{pk})$ becomes a function of m trees with different tree structure and decision rules, whose contributions are added and then scaled using a cumulative density function of a standard normal distribution Φ.

Figure 5.1: SL procedure to estimate success probabilities



[9]For a graphical visualization of the FFT decision rules to estimate Phase I probabilities see Appendix D.1 figure D.1.

For around 17,000 product-market observations obtained from the EvaluatePharma® data base, the phase specific outcome is known, yet without exact information regarding its announcement date. Even though these observations cannot be directly linked to financial market reactions, they are valuable to rigorously train and validate the SL procedure that is used for estimating phase specific success probabilities (the estimation procedure is sketched in figure 5.1). To do so, we assign all project observations, that is the product features and their phase specific outcome, to four phase specific data sets: PI, PII, PIII and Filed. The data set PI contains all projects that either successfully moved on to Phase II or beyond (codified as success) or were abandoned during Phase I (codified as failure). The remaining data sets are constructed alike. That way SL algorithms can assign for each project the probability it will advance to the next stage. We split each phase-specific data set into three parts: 70% of the data form the training set on which each algorithm learns to form its decision function; 15% are used to derive the optimal weights between algorithms at each stage using a quadratic optimization procedure and 10-fold cross validation; 15% are used after the training steps to validate the out of sample performance.[10]

To test *Hypothesis 1*, we estimate for the event data set, that contains exact announcement dates related to project news and has so far been kept separated, the phase-, product-, and therapeutic market-dependent success probability using the previously trained methods. For each method, we then multiply stage-specific probability estimations to obtain a measure of the project probability to pass all stages (see equation 5.2). The success probability estimates from each method are weighted with the previously calculated optimal stage-specific weights to derive an aggregate probability estimate of success reflecting all methods. The probability estimates derived from the SL procedure can then be related to market reactions that follow innovation announcements. As a robustness check we also relate the historical success ratio of similar products to market

[10]We provide more details on the SL methods used, the estimation process, the out-of-sample performance, and the derivation of algorithm weightings in Appendix D.1.

reactions. Next, we turn to the estimate of product portfolio importance that will be related to market reactions, as well.

5.4.3 Measurement of Relative Portfolio Importance

According to our model, the relative importance of a company's product depends on the expected cash flow assigned to each of the company's products discounted to the event date. Finding suitable data that appropriately reflects this specific information is complicated. We try to come as close as possible to the ideal measurement by assessing the relative importance of a product in a company's portfolio by its NPV. That is we take the product-specific NPV at the event date and divide it by the NPV of all products of the company at that date. Usually product-specific NPV estimates are cumbersome to calculate and, for historical data, difficult to obtain because of their time dependency. For our research purpose, we sourced this exclusive information via the EvaluatePharma® platform which contains discounted cash flow information for over 5,000 products (Evaluate Ltd., 2019).[11] A product NPV is the expected free cash flow that is forecast over the lifetime of the product and then discounted to the value at the event date. For our data, expected free cash flows are calculated by subtracting estimated product costs and taxes from expert consensus forecast of product revenues. These net projections are afterwards discounted to the event date using their cost of capital. The revenue extrapolation considers the estimated lifetime of the product depending on its estimated patent expiration. If we assume that the provided NPV estimates reflect market expectations, we should observe that products with a high NPV share relative to the total company's NPV, namely products with high portfolio importance, show higher market reactions after innovation announcements.

[11] The real option approach is another prominent method to derive the value of projects under development. However, data on the time-dependent option value of projects across a sample of companies is usually infeasible to obtain (see, for example, Kellogg and Charnes, 2000; León, Piñeiro, et al., 2004 who base their valuation analysis on a single company). Since we are mainly interested in a measure of the relative project value within a company's portfolio that can be applied across firms, we opt to use the NPV framework for which data is available to us.

Since the NPV measure relies on future projections that are only possible to verify with hindsight, we provide an alternative measure for portfolio importance that we use as a robustness check. We construct a simple development index that weights the number of company products by their success probability according to their R&D stage (Hao et al., 2017). The companies development index (DI) is obtained by equation 5.8:

$$DI = \frac{\sum_{i=1}^{|p|} p_{ki} \cdot \omega_k}{100} \tag{5.8}$$

where ω_k is the average success probability of a product at stage k to be launched according to Paul et al., 2010. The measure is scaled by 100 to better visualize its impact in the subsequent regression specifications. A high DI implies that the expected number of products that generates future revenues for the associated company is high. Thus, companies with high DI should show lower market reactions after product specific news. In the next section we describe the data set used in our analysis and present descriptive statics of the measures introduced in this section.

5.5 Data and Descriptive Statistics

We originally sourced 1,340 announcements from the *EvaluatePharma*® platform including the date, a short announcement description, and further the company, the product and the therapeutic market the announcement refers to. From the original data set we need to exclude 636 events mainly because they were not related to clinical phase innovation announcements.[12] The remaining 703 events, on which our main analysis

[12]Breakdown of excluded events: for 57 events data on historical company stock prices was not available; for 154 events a conclusive description of whether the nature of the event was positive or negative was missing (e.g. "company reports data of its clinical trial xyz regarding product p for market m."); 273 events could not be mapped specifically to one of the clinical research stages Phase I, Phase II, Phase III, Filed and therefore not assigned to an estimated success probability (e.g. announcements regarding priority review or opinions of the EU advisory committee are not phase-specific); for 100 events the event date was within +/- 3 days from another event of the same stock and thus might have caused overlapping market reactions. 52 events occurred before 2017 and were dropped to ensure

relies, took place between 04.01.2017 and 31.12.2018. They are associated to 289 distinct companies, 444 products and 209 therapeutic markets.[13] Each event can be associated to the development stage at which it occurred: Phase I (79), Phase II (167), Phase III (285), and Filed (172); and was grouped after manual inspection of the event description as positive (490) or negative (213). The event distribution according to stage and outcome is summarized in table 5.3.

Table 5.3: Data distribution of development phases and outcomes

	Phase I	Phase II	Phase III	Filed	Sum
Positive	60	101	191	138	490
Negative	19	66	94	34	213
Sum	79	167	285	172	**703**

Using the final sample we relate market reactions, which constitute the dependent variables measured by CAR and $Return$, to our variables of interest namely the product portfolio importance $Port$ and the product success probability $Prob$ and interact them with the outcome of the event ($Positive$). We report the descriptive statistics of all variables including their counterparts used in robustness checks (DI and $HIST$) in table 5.4. We witness that both return measures show pronounced outlier reactions, which are addressed in the following regression specifications. Most measures of the variables of interest are available for all observations. The only exception is the $Port$ measure for which it was not possible to obtain product-specific NPV measures for every product. This potentially induces some sort of selection bias, that affects the composition of sampled market reactions. Yet, the observations for which the $Port$ variable is available do not differ significantly in terms of the CAR

that mandatory disclosure policies apply for the whole sample. One event was associated to a product for which the data reported an NPV share greater than 1.

[13] About 60% of the companies are associated to just one event. More than 90% are associated to less than five events.

and $Return$ measures from the observations for which $Port$ is unavailable.[14]

Table 5.4: Summary statistics of dependent variable and variables of interest

	Obs	Min	1. Quartile	Median	3. Quartile	Max	Mean	Variance
Dependent Variables								
CAR	703	-2.68	-0.046	0.007	0.078	1.449	-0.033	0.132
Return	703	-0.922	-0.055	0.012	0.105	4.476	0.031	0.151
Variables of interest								
Port	510	0.000	0.038	0.159	0.770	1	0.384	0.148
(DI)	703	0.000	0.014	0.056	1.454	12.023	1.182	4.665
Prob	703	0.037	0.349	0.729	0.911	0.987	0.626	0.098
(HIST)	703	0	0.137	0.500	0.860	1	0.512	0.130
Positive	703	0	0	1	1	1	0.697	0.211

Table 5.4 summarizes the dependent variables CAR and $Return$ together with the variables of interest $Port$,DI, $Prob$, $HIST$, and $Positive$. The table contains information on the number of observations, minimum value, 25% percentile, median, 75% percentile, maximum value, mean, and variance.

Besides our variables of interest, we consider in the regression analysis a set of control variables that accounts for company, market, product, and time specific information available to us (see table 5.5). The variable *In-house* categorizes whether the product associated to the event was developed in-house, was in-licensed, or acquired from another company. Sharing the risks and rewards of product development with another company via licensing agreements could possibly reduce the impact of product related innovation announcements. The variable $Phase$ indicates the development stage of the product at time of the event and, as suggested by the literature, might have an effect on market reactions.

[14] We test for mean differences in CAR between observations for which the $Port$ measure is available and observations for which it is not available using Wilcoxon rank sum tests (p-value: 0.458) and Welch's t-test (p-value: 0.365). We perform the same tests using the $Return$ measure leading to a p-value of 0.650 and 0.265, respectively.

The therapeutic area for which the product is aiming to get approval is controlled by $Market$. Therapeutic markets differ in their size, number of competitors, embedded development risk, and thus potentially also in average market reactions. We also include an indicator whether the product seeks approval for multiple therapeutic markets $Multiple$, since we expect that news for products that have multiple market entry options are, on average, less severe. On the company level we account for the size of the company and its breadth of operations by including the variable $Class$. Product specific news might affect big, globally operating companies differently than smaller, specialized firms. We also account for the $Region$ the company stock is listed in, since market reactions after R&D announcements might be different across geographical locations. Lastly, we include with the variable $Quarter$ time fixed effects to pick up potential seasonal trends.[15]

We provide a graphical illustration in figure 5.2 on the relationship between positive and negative market reactions and the presumed probability and portfolio effect. Sub figure (a) shows the differences in average cumulated abnormal returns before, at, and after the event date for positive and negative events whose product success probability is above (high) or below (low) the median success probability. Confirming previous evidence of the literature on company announcements, we observe larger market reactions after negative news (red lines) than after positive news (green lines). Positive events which relate to products with a low probability of success, show higher reactions (green dotted line) than products with an estimated high success probability (green solid line). The same holds for negative news, but for a lesser extent. Shifting our focus to the portfolio effect in sub figure (b), we observe that for products whose relative NPV share in the companies portfolio is high (solid lines), the absolute return jump at the event date is amplified, even more so in case the event is negative. We quantify the $\overline{CAR}$ for each of the reported groups of events in table D.4 and D.5 in Appendix D.2 reporting in addi-

[15]Since most companies are associated to only one event, we do not include fixed effects on the company level.

Table 5.5: Summary of control variables - number of observations per group

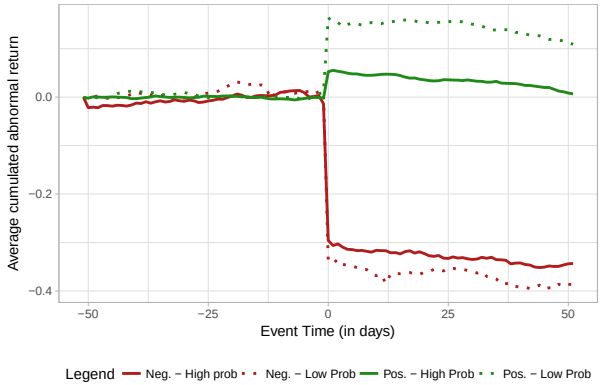
	Organic	Phase	Market	Multiple	Class	Region	Quarter
1	Organic (346)	Phase I (79)	Cancer (227)	One Market (563)	Bio- technology (351)	North America (463)	Q1 '17 (68)
2	In-licensed (181)	Phase II (167)	Blood (35)	Many Markets (140)	Global Majors (192)	Europe (222)	Q2 '17 (110)
3	Acquisition (157)	Phase III (285)	Cardio-vascular (29)	-	Regional Major (35)	Other (18)	Q3 '17 (93)
4	Other (19)	Filed (172)	Immunology (37)	-	Speciality (123)	-	Q4 '17 (80)
5	-	-	Infections (32)		Other (2)	-	Q1 '18 (70)
6	-	-	Miscellaneous (176)	-	-	-	Q2 '18 (91)
7	-	-	Musculo-skeletal (41)	-	-	-	Q3 '18 (85)
8	-	-	Neurology (86)	-	-	-	Q4 '18 (106)
9	-	-	Respiratory (40)	-	-	-	-

Table 5.5 summarizes the set of control variables used in regression specifications 2-4. Number of observations in parenthesis. Groups in the first row are used as reference group in the regressions.

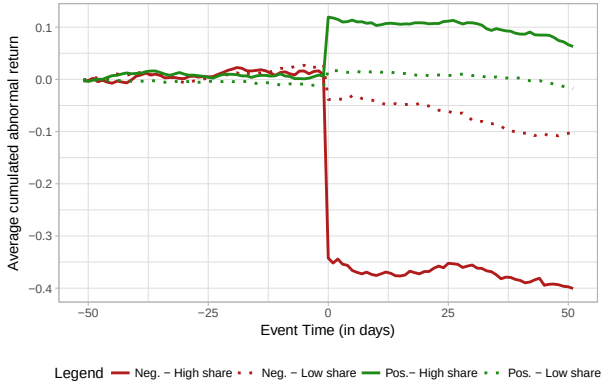
tion parametric and non-parametric test statistics for each group.[16] Yet, by looking at average market reactions in isolation we can only provide preliminary evidence on our hypotheses. In the following result section we present regression specifications that control for confounding factors and are thus suited to test our hypotheses on the probability and portfolio effect appropriately.

[16]The employed test statistics are detailed in Appendix D.3.

Figure 5.2: Cumulative abnormal return for positive and negative product announcements



(a) Below and above median product success probability



(b) Below and above median product portfolio importance

Figure 5.2 shows the average cumulated return for different event types between event time -50 and + 50 (in days). Green lines represent returns from positive events, red lines returns from negative events. Figure (a) shows the average return of events associated to a success probability above median (high) with solid lines, and below median (low) with dotted lines. Figure (b) shows the average return of events associated to a NPV share above median (high) with solid lines and below median (low) with dotted lines.

5.6 Results

We present the results from our regression analysis in subsection 5.6.1. Robustness checks using different measures are then discussed in subsection 5.6.2.

5.6.1 Regression Analysis

We implement three regression models that focus on the portfolio effect (1), the probability effect (2), and both effects combined (3). Each model relates the dependent variable y_i, which is either the CAR or the discrete one-day return ($Return$) of event i, to a linear combination of variables of interest and a set of control variables.

$$y_i = \alpha + \beta_{pos} Pos_i + \beta_{port} Port_i + \beta_{pos \times port} Pos_i \times Port_i + \gamma \Gamma_i + \epsilon_i \quad (1)$$

$$y_i = \alpha + \beta_{pos} Pos_i + \beta_{port} Prob_i + \beta_{pos \times prob} Pos_i \times Prob_i + \gamma \Gamma_i + \epsilon_i \quad (2)$$

$$y_i = \alpha + \beta_{pos} Pos_i + \beta_{port} Port_i + \beta_{pos \times port} Pos_i \times Port_i + \\ + \beta_{port} Prob_i + \beta_{pos \times prob} Posi_i \times Prob_i + \gamma \Gamma_i + \epsilon_i \quad (3)$$

The variable of interest in the first model is the portfolio importance $Port_i$ of the product related to event i that is measured by the share of the company NPV attributed to the product at the time of the event. The indicator variable $Positive_i$ is equal to one for positive events and zero for negative news. We interact $Positive_i$ with $Port_i$ to observe whether the portfolio importance has a different effect for positive and negative announcements. The matrix of control variables is denoted by Γ_i. The second model relates the product success probability $Prob_i$, estimated via a weighted average across supervised learning predictions, to the return measure at event i. Again we pick up differences between positive and negative events by interacting $Positive_i$ with $Prob_i$. In the third model we examine the mutual relationship of portfolio and probability effect on abnormal returns and thus include information on $Port$, $Prob$, and $Positive$. It is theoretically possible that the probability effect interacts

somehow with the portfolio effect, yet its direction is ambiguous. Pharmaceutical products that carry more risk are found to be associated with higher potential rewards (Pammolli et al., 2011), that could positively impact a product's relative importance in the company portfolio. But product innovations that are vital in a companies portfolio presumably enjoy favorable conditions during their development, which in turn may reduce the risk of failure in development. From an empirical standpoint, we check whether both measures are correlated across clinical phases and thus provide a first hint on their interrelation. The estimated correlation coefficients of $Prob_i$ and $Port_i$ are -0.231 (p-value 0.123) for Phase I, 0.011 (p-value 0.919) for Phase II, -0.064 (p-value 0.335) for Phase III, -0.049 (p-value 0.559) for Filed projects. Since one cannot reject the null hypothesis of (linear) independence at none of the development stages, we decide to omit the inclusion of interaction terms between these two variables.[17]

We report four different specifications for each of the three models. Specification one uses an ordinary least squares (OLS) regression and includes only the variables of interest, the interactions and a constant. In specification two we perform the same regression but include all control variables, as specified above, and additionally cluster standard errors on company level to allow for dependence between events that affect the same company. Since some market reactions are so extreme that they could potentially affect mean-based point estimates, we add two regression specifications that are robust to outliers. In specification three we perform a quantile regression on the median including all control variables and robust standard errors. These estimates are not influenced by single market reactions at the tails of the distribution. In specification four we derive robust regression estimates based on weighted least squares with Huber and biweight iterations which has been proposed in a similar setting (Urbig et al., 2013). Robust regression estimates also deal with not normally distributed residuals, which are common for event

[17] From a parsimony perspective, Bayesian information criterion comparisons suggest that interaction effects between $Port$ and $Prob$ over-specify model (3); see table D.6 in Appendix D.2.

studies under OLS and could lead to wrong statistical inference (Henderson Jr, 1990). We run each of the four regression specifications with CAR and with $Return$ as dependent variables.

Table 5.6: Model 1 - Effect of product portfolio importance $Port$ on $Return$ and CAR

	Dependent variable: $Return$				Dependent variable: CAR			
	Spec. 1	Spec. 2	Spec. 3	Spec. 4	Spec. 1	Spec. 2	Spec. 3	Spec. 4
$Port$	-0.27***	-0.32***	-0.32***	-0.34***	-0.44***	-0.47***	-0.35***	-0.29***
	(0.05)	(0.06)	(0.05)	(0.03)	(0.05)	(0.10)	(0.10)	(0.04)
$Positive$	0.09***	0.10***	0.03**	0.07***	0.07*	0.06*	0.03***	0.09***
	(0.03)	(0.03)	(0.02)	(0.02)	(0.04)	(0.03)	(0.01)	(0.02)
$Port \times Positive$	0.47***	0.46***	0.41***	0.45***	0.54***	0.55***	0.42***	0.38***
	(0.05)	(0.07)	(0.06)	(0.03)	(0.06)	(0.10)	(0.10)	(0.04)
Constant	-0.06**	-0.04	0.05**	0.00	-0.04	0.01	0.03	-0.05
	(0.03)	(0.05)	(0.02)	(0.03)	(0.03)	(0.06)	(0.03)	(0.04)
Observations	510	510	510	510	510	510	510	510
Adjusted R^2	0.359	0.398	-	0.601	0.353	0.407	-	0.475
Control variables	No	Yes	Yes	Yes	No	Yes	Yes	Yes
Outlier correction	No	No	Yes	Yes	No	No	Yes	Yes

Table 5.6 contains the regression results for model 1 concerning the effect of product portfolio importance $Port$ on $Return$ (column 2-5) and on CAR (column 6-9). Standard errors in parenthesis. P-value $\leq 0.1 :^{*}$ $\leq 0.05 :^{**}$ $\leq 0.01 :^{***}$.

Table 5.6 shows the effect of product portfolio importance on market reactions that follow innovation announcements. The variable $Port \times Positive$ is significant and positive for all specifications implying that, in case of a positive event, products of higher portfolio importance show higher returns. The variable $Port$ captures the effect for negative events: products of higher portfolio importance are associated to lower returns for all specifications. Overall, as assumed in *Hypothesis 2*, products with higher portfolio importance are associated to higher absolute returns at

innovation announcements. The variable $Positive$ captures the difference between positive and negative outcomes for products whose relative NPV share is zero. The effect is positive and significant. Most control variables seem not to affect market reactions across specifications. The only exception are products that are in development for more than one market or that are in the filed stage, for which we observe a significant negative shift in returns with respect to their base category (*One market* and *Phase I*, respectively) in four out of six specifications. For reasons of clarity we report the full regression table including control variables in Appendix D.2 , table D.7.

Result 1. *The higher the importance of a product in a companies portfolio, the higher are market reactions that follow innovation announcements.*

For model two the results are reported in table 5.7. The variable $Prob \times Positive$ shows that, for positive news, projects that are estimated to be likely successful are associated significantly lower returns than projects that were estimated to be more risky. The variable $Prob$ captures the relationship between stock returns and success probability for negative news. The effect is not significant in the first specification that does not control for confounding factors. When adding control variables, such as the stage of development, we find in five out of six specifications that returns are significantly higher (so for negative returns lower in magnitude) if they are associated to products with a high success probability. Yet, for negative news the relationship between product success probability and market reaction is attenuated, possibly because the severity of negative events visibly varies more than the one of positive events and thus variability in the tails of the return distribution is left unexplained.[18] Consequently regression specification three and four, that mitigate the effect of outlier events, capture the effect better. The variable $Positive$

[18] Performing a content analysis (Bell et al., 2018) on the event descriptions in our sample, we observe that negative events range from delays in product launch to abrupt termination of development whereas positive announcements commonly report that products have reached the next stage in development. Figure D.2 in Appendix D.2 shows that the return distribution of negative announcements is more disperse.

Table 5.7: Model 2 - Effect of product success probability $Prob$ on $Return$ and CAR

	Dependent variable: $Return$				Dependent variable: CAR			
	Spec. 1	Spec. 2	Spec. 3	Spec. 4	Spec. 1	Spec. 2	Spec. 3	Spec. 4
$Prob$	0.10	0.12*	0.23***	0.12**	0.06	0.07	0.22**	0.12*
	(0.07)	(0.07)	(0.08)	(0.05)	(0.07)	(0.10)	(0.11)	(0.06)
$Positive$	0.66***	0.65***	0.47***	0.53***	0.58***	0.56***	0.41***	0.52***
	(0.06)	(0.08)	(0.05)	(0.04)	(0.05)	(0.07)	(0.09)	(0.04)
$Prob \times Positive$	-0.40***	-0.35***	-0.31***	-0.27***	-0.23***	-0.21**	-0.28***	-0.24***
	(0.09)	(0.10)	(0.08)	(0.05)	(0.08)	(0.10)	(0.11)	(0.06)
Constant	-0.31***	-0.35***	-0.31***	-0.32***	-0.37***	-0.36***	-0.28***	-0.34***
	(0.05)	(0.06)	(0.05)	(0.04)	(0.04)	(0.07)	(0.09)	(0.05)
Observations	703	703	703	703	703	703	703	703
Adjusted R^2	0.268	0.282	-	0.453	0.314	0.336	-	0.374
Control variables	No	Yes	Yes	Yes	No	Yes	Yes	Yes
Outlier correction	No	No	Yes	Yes	No	No	Yes	Yes

Table 5.7 contains the regression results for model 2 concerning the effect of product success probability $Prob$ on $Return$ (column 2-5) and on CAR (column 6-9). Standard errors in parenthesis. P-value $\leq 0.1 :^{*} \quad \leq 0.05 :^{**} \quad \leq 0.01 :^{***}$.

is significant and positive while the $Constant$ is significant and negative indicating that if the product success probability is estimated to be zero then negative events differ substantially from positive ones. Again the majority of control variables is not significantly associated to shifts in market reactions across regression specifications, except for two groups. We observe negative return shifts for large companies ("global majors") and products that are candidate for multiple therapeutic markets.[19]

Result 2. *The higher the success probability of a product, the lower are market reactions that follow innovation announcements. The effect is moderated for*

[19]The complete regression table (table D.8) that includes all control variables is reported in the Appendix D.2.

negative announcements.

Since usually probability and portfolio effects cannot be observed in isolation, model three includes $Prob$ and $Port$ to check how their mutual variation affects abnormal returns around innovation announcements. Table 5.8 shows that sign and magnitude of the portfolio effect stay practically unchanged. All specifications suggest that events related to products with high portfolio value are subject to more extreme market reactions. As in model two, the probability effect is significant for most specifications independent of the event being positive or negative. Yet the size of the probability effect, in particular for positive news, seems lower. Specifications that show no significant relationship do either not include control variables or, in case of CAR as dependent variable, do not deal with outlier returns. The effect of control variables on market reactions is not significant across specifications apart from the already mentioned negative shift in returns for multi-indication products.[20]

Moreover, the adjusted R^2 of model three, a measure of explained return variation, is slightly higher than in model one for all specifications suggesting that the inclusion of information on the product success probability increases the explained variation in market reactions beyond the information that is contained in the $Port$ measure alone. Since the probability measure is static for each product, $Port$ might better capture time-dependent information right before the event. We assume that product-specific probability measures that reflect all information available to investors before the event would capture even more return variation. Yet, with the measures at hand, the portfolio effect explains return variation around innovation announcements better than the probability effect.

To visualize the mutual effect of probability and portfolio measures on daily stock price returns, we report the marginal effects of $Prob$ and $Port$ of model three- specification four in figure 5.3 (a) for positive events and in figure 5.3 (b) for negative events. The estimated market reaction of positive events is highest for products of high portfolio importance

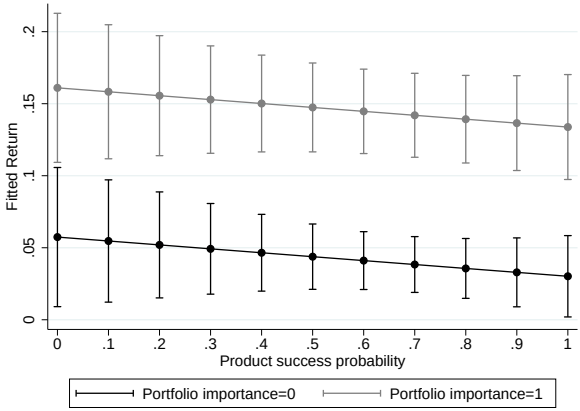
[20]The complete regression table (table D.9) that includes all control variables is reported in the Appendix D.2.

Table 5.8: Model 3 - Effect of product success probability $Prob$ and product portfolio importance $Port$ on $Return$ and CAR

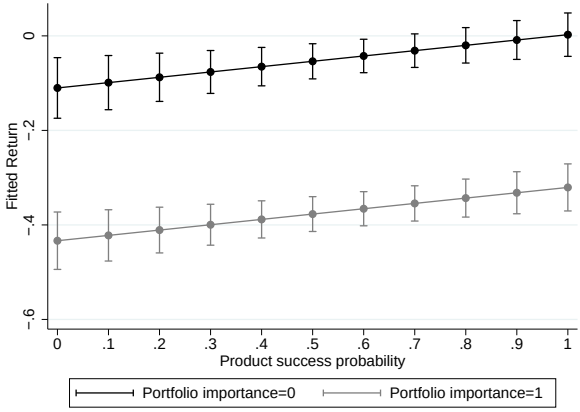
	Dependent variable: $Return$				Dependent variable: CAR			
	Spec. 1	Spec. 2	Spec. 3	Spec. 4	Spec. 1	Spec. 2	Spec. 3	Spec. 4
$Positive$	0.23***	0.23***	0.17***	0.17***	0.08	0.07	0.11***	0.18***
	(0.06)	(0.06)	(0.03)	(0.04)	(0.07)	(0.08)	(0.03)	(0.04)
$Port$	-0.26***	-0.30***	-0.25***	-0.32***	-0.45***	-0.47***	-0.32***	-0.27***
	(0.05)	(0.06)	(0.05)	(0.03)	(0.05)	(0.10)	(0.07)	(0.04)
$Port \times Positive$	0.43***	0.43***	0.34***	0.43***	0.53***	0.55***	0.38***	0.36***
	(0.05)	(0.07)	(0.05)	(0.03)	(0.06)	(0.10)	(0.07)	(0.04)
$Prob$	0.07	0.15**	0.17***	0.11***	-0.10	-0.02	0.10**	0.09*
	(0.06)	(0.07)	(0.05)	(0.04)	(0.07)	(0.10)	(0.05)	(0.05)
$Prob \times Positive$	-0.19***	-0.20**	-0.18***	-0.14***	0.00	-0.00	-0.11**	-0.13**
	(0.07)	(0.08)	(0.05)	(0.04)	(0.08)	(0.11)	(0.05)	(0.05)
Constant	-0.11**	-0.13**	-0.08**	-0.07*	0.02	0.02	-0.03	-0.11**
	(0.05)	(0.06)	(0.03)	(0.04)	(0.06)	(0.09)	(0.04)	(0.05)
Observations	510	510	510	510	510	510	510	510
Adjusted R^2	0.372	0.404	-	0.609	0.359	0.405	-	0.480
Control variables	No	Yes	Yes	Yes	No	Yes	Yes	Yes
Outlier correction	No	No	Yes	Yes	No	No	Yes	Yes

Table 5.8 contains four specifications for model 3 concerning the effect of product success probability $Prob$ and product portfolio importance $Port$ on $Return$ (column 2-5) and on CAR (column 6-9). Standard errors in parenthesis. P-value $\leq 0.1 :^{*} \quad \leq 0.05 :^{**} \quad \leq 0.01 :^{***}$.

Figure 5.3: Estimated portfolio and probability effect for positive events and negative events



(a) Positive events



(b) Negative events

Figure 5.3 shows the daily discrete return estimate and its 95% confidence interval for 11 levels of product success probability (x axis) and two levels of product portfolio importance (gray and black scaled lines). Positive events are depicted in sub figure (a) and negative events in sub figure (b).

whose success sentiment was estimated to be low (16.1%). For events associated to a product with low importance but high success probability, the estimated market reaction is 3.0%. In case of negative news, predicted returns are lower for products with high portfolio importance and low product success probability (-43.3%) and higher for products with low portfolio importance and high product success sentiment (0.0%). For both good and bad news, a change in the product importance is estimated to result in a more drastic market reaction than a change in the success probability. In line with the return asymmetry that has been frequently observed in related settings, the magnitude of estimated market returns after positive announcements is, on average, lower than the ones of negative announcements. The next section presents robustness checks which confirm these results.

5.6.2 Robustness Checks

Since the portfolio and probability effect might depend on our specific choice of variables, we replace both original measures with simpler approximations. To measure the portfolio effect, we construct a development index (DI) for each company that weights the number of products in a company's portfolio by its current development stage (see section 5.4.3). According to *Hypothesis 2*, companies with a high DI should show, on average, low market reactions after product innovation announcements. As a measure of a product's success probability we use the historical success rate of products in the same stage and therapeutic market $HIST$. The measure of HIST has the advantage that it is easily accessible and already established to approximate success rates.

Table 5.9 and 5.10 show the regression results for the two effects using the same four specifications and two return types as for the original measure. The value of the development index is negatively associated to returns for positive events ($DI \times Positive$) and positively associated to returns for negative events (DI) across specifications. In other words, market reactions that follow product-specific news, independent

whether positive or negative, affect companies with a large development portfolio less than companies whose revenue projections rely on a few products.

Table 5.10 reports results that are consistent with the original findings presented in table 5.7. The market reactions that follow positive announcements are estimated smaller for products with high success probability ($HIST \times Positive$ is negative). Market reactions after negative events are also decreasing in the product success probability ($HIST$) but the effect is weaker. We compare the adjusted R^2 of the original measures with the ones of the measures used in the robustness check. As expected, we find that the product portfolio importance $Port$ explains observed market reactions better than the more generic company portfolio index DI across all specifications. Concerning the success probabilities, recall that the historical success rate contributes only, on average across phases, 2% to the weighted $Prob$ estimate (see table 5.2). Changing this weighting to 100%, as in the $HIST$ specification, does not create substantial differences in terms of explained variation suggesting that changing the weighting of probability estimates in the SL procedure delivers robust effects on market reactions.[21]

Another robustness check concerns the dependent variable CAR, which is supposed to capture the stock price reactions that are specific to an event and not any other, market-driven, price changes. To test whether the results we obtain are due to a specific choice of market index, we estimate the abnormal return (see equation 5.4) using two different market indices, *Ishares MSCI World index* and *Dow Jones Industrial index*, that are broadly constructed and thus might capture market variation differently. Yet, the conclusions we can draw from the regression results, provided in table D.11-D.16 of Appendix D.2, remain virtually unchanged.

[21] Using the alternative measures jointly in model three, we obtain similar results than when using the original measures. Namely, we find strong support for the portfolio effect for both positive and negative announcements, strong support for the probability effect for positive announcements and partial support for the probability effect for negative announcements. We report model 3 results in table D.10 of Appendix D.2.

Table 5.9: Robustness check 1- Effect of product portfolio importance DI on $Return$ and CAR

	Dependent variable: $Return$				Dependent variable: CAR			
	Spec. 1	Spec. 2	Spec. 3	Spec. 4	Spec. 1	Spec. 2	Spec. 3	Spec. 4
DI	0.05***	0.05***	0.06**	0.06***	0.07***	0.06***	0.05***	0.06***
	(0.01)	(0.02)	(0.03)	(0.01)	(0.01)	(0.02)	(0.01)	(0.01)
$Positive$	0.49***	0.51***	0.33***	0.45***	0.52***	0.52***	0.31***	0.45***
	(0.03)	(0.04)	(0.03)	(0.02)	(0.03)	(0.04)	(0.04)	(0.02)
$DI \times Positive$	-0.08***	-0.08***	-0.06**	-0.08***	-0.09***	-0.09***	-0.06***	-0.08***
	(0.01)	(0.02)	(0.03)	(0.01)	(0.01)	(0.02)	(0.01)	(0.01)
Constant	-0.29***	-0.29***	-0.20***	-0.28***	-0.39***	-0.35***	-0.20***	-0.31***
	(0.02)	(0.06)	(0.03)	(0.04)	(0.02)	(0.05)	(0.04)	(0.04)
Observations	703	703	703	703	703	703	703	703
Adjusted R^2	0.275	0.299	-	0.491	0.352	0.374	-	0.417
Control variables	No	Yes	Yes	Yes	No	Yes	Yes	Yes
Outlier correction	No	No	Yes	Yes	No	No	Yes	Yes

Table 5.9 contains the regression results for robustness check 1 concerning the effect of product portfolio importance DI on $Return$ (column 2-5) and on CAR (column 6-9). Standard errors in parenthesis. P-value $\leq 0.1 :^{*} \quad \leq 0.05 :^{**} \quad \leq 0.01 :^{***}$.

5.7 Discussion and Concluding Remarks

Innovation is of vital importance for corporations and its announcements of high interest to customers, competitors, and investors who rapidly capitalize the news in the value of the companies. The resulting market reactions can be severe, and in case of negative news, put further innovation attempts at risk. Characterizing elementary factors that drive these market reactions is thus a substantial requirement to better understand future market reactions and form appropriate preventive actions. Using a general framework, we propose that market reactions which follow project related innovation announcements can be described by two main effects: a probability effect, that relates to investors' sentiment re-

Table 5.10: Robustness check 2- Effect of product success probability $HIST$ on $Return$ and CAR

	Dependent variable: $Return$				Dependent variable: CAR			
	Spec. 1	Spec. 2	Spec. 3	Spec. 4	Spec. 1	Spec. 2	Spec. 3	Spec. 4
$HIST$	0.11	0.14**	0.20***	0.14***	0.15**	0.17*	0.13*	0.12*
	(0.07)	(0.07)	(0.06)	(0.05)	(0.06)	(0.09)	(0.08)	(0.06)
$Positive$	0.58***	0.57***	0.41***	0.48***	0.58***	0.56***	0.34***	0.48***
	(0.05)	(0.06)	(0.04)	(0.03)	(0.04)	(0.06)	(0.06)	(0.03)
$HIST \times Positive$	-0.34***	-0.29***	-0.24***	-0.24***	-0.30***	-0.25***	-0.19**	-0.22***
	(0.08)	(0.07)	(0.06)	(0.05)	(0.07)	(0.08)	(0.08)	(0.06)
Constant	-0.30***	-0.32***	-0.26***	-0.31***	-0.40***	-0.38***	-0.22***	-0.33***
	(0.04)	(0.06)	(0.04)	(0.04)	(0.03)	(0.06)	(0.06)	(0.05)
Observations	703	703	703	703	703	703	703	703
Adjusted R^2	0.261	0.278	-	0.451	0.319	0.340	-	0.373
Control variables	No	Yes	Yes	Yes	No	Yes	Yes	Yes
Outlier correction	No	No	Yes	Yes	No	No	Yes	Yes

Table 5.10 contains the regression results for robustness check 2 concerning the effect of product success probability $HIST$ on $Return$ (column 2-5) and on CAR (column 6-9). Standard errors in parenthesis. P-value $\leq 0.1 :^{*}$ $\leq 0.05 :^{**}$ $\leq 0.01 :^{***}$.

garding the innovation, and a portfolio effect, that relates to the relative importance of the innovation in terms of future expected cash flows.

Data on the bio-pharmaceutical industry is particularly well-suited to empirically test our theoretical considerations, since product related R&D announcements are mandatory, frequent, and market reactions after announcements mostly substantial. With the help of a recent and sizable data set, we provide two novel measures to approximate our variables of interest. First, we extend the use of supervised learning methods to approximate investor project success sentiment, which goes beyond prediction tasks such as predicting equity premia (Gu et al., 2018) or classifying pharmaceutical project outcomes (e.g., Feijoo et al., 2020). Second,

we directly apply date-specific NPVs at the product and company level to construct a measure that closely reflects the relative importance of a product with respect to its expected discounted cash flows.

The predicted relationships between product characteristics and market reactions are confirmed. The higher the importance of a product in a company's portfolio, the higher is the abnormal market return that follows product related innovation announcements, independent of whether the announcement is positive or negative. Further, the higher the success probability of a product for a specific therapeutic market, the lower are the market reactions that follow the announcements. Yet, in the case of negative news, this relationship seems to be attenuated. Two lines of reason might jointly be responsible for this. First, our assumption, that the project's success probability is inversely related to the change of probability induced by the announcement, might hold on average but for a few late-stage, negative events we can observe extreme market reactions (e.g., after announcing the discontinuation of a project) that mitigate the overall effect. In fact, in the regression specifications that are robust to outlier market reactions, we observe the hypothesized relationship between project success probability and market reactions after negative announcements to be significant. Second, it has been discussed that people have difficulties to objectively assess success probabilities (e.g, Barberis, 2013). Even if the derived success probabilities reflect reasonable estimations, we cannot be certain that investors perceive them as such and thus react in some situations different from our expectations.

In addition, we confirm previously noted return asymmetries between positive and negative events. Possibly, positive events, that mostly refer to an incremental advance in the development process, are associated to lower product success probability revisions while negative events can range from product launch delays up to the abandonment of a research line and hence reflecting higher average success probability revisions. In fact, we observe that the variation of market reactions that follow negative news is higher than the one after positive ones, but this explanation remains tentative since we do not measure the success probability revision itself. Future work in which the aggregate investor perception of

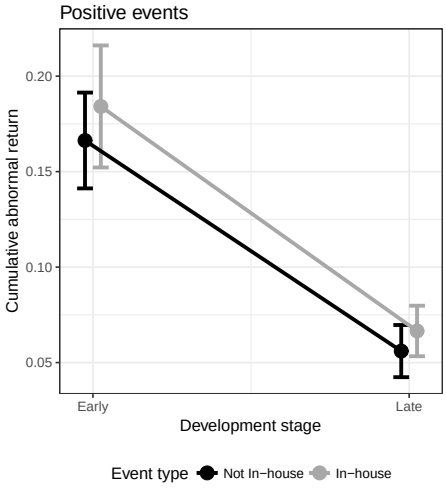
success probabilities before and after an event can be directly measured would be a silver bullet to further characterize the probability effect and to dissolve the above mentioned concerns.

Independent of their positive or negative nature, innovation announcements that relate to products of high importance and high risk show, on average, the highest market reactions. Especially managers who profit from increasing stock prices (e.g., due to share based renumeration packages Dawid et al., 2019) but do not suffer financial consequences from negative market reactions can be expected to promote such "high risk-high gain" products. Yet, company stakeholders with contrary incentives might be interested in mitigating the risk of market reactions that follow from adverse innovation announcements.

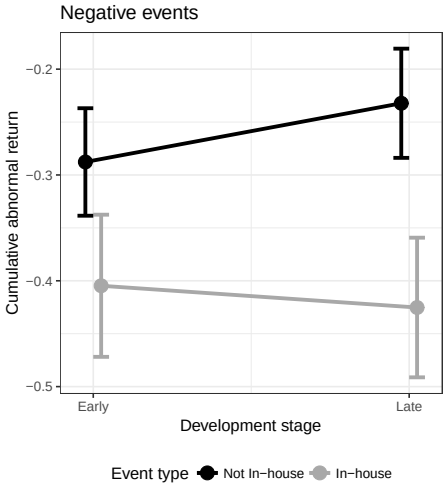
Our results suggest that one effective way to reduce market reactions around R&D announcements is to lower the relative importance of single products in the product portfolio. Yet, especially for companies with a limited product pipeline, reducing product specific dependencies is not trivial. One possibility that leaves the size of the portfolio unchanged is to form strategic alliances during the product development and thus to share product specific risks. We report some suggestive evidence in figure 5.4 showing that the absolute market reactions after negative news are lower when associated to products that share the risks and merits of co-development, even more so for late development stages (subfigure b). On the contrary, market reactions that follow positive news do not show differences in whether they relate to in-house developed products or not (subfigure a). Increased involvement in a market for technologies (Arora et al., 2001) may thus mitigate financial effects induced by negative innovation announcements, while at the same time it does not substantially affect market reactions that follow positive events.[22] Yet, strong reliance on collaborations may reduce other long-term risks, such as becoming

[22] We derive the marginal effects of the variable *In-house* and *Not in-house* interacted with *Positive* and *Late stage* on CAR using a linear regression that includes the probability, the portfolio measure, and all other specified control variables. We find that for negative events the predicted CAR for *In-house* products is significantly lower than for products not developed in-house, on the 10% level. For positive events we do not find a significant difference, as suggested in the figures. The predicted CAR estimates are reported in table D.17 in Appendix D.2.

Figure 5.4: $\overline{CAR}$ of in-house and not in-house developed products in early and late stages



(a) Positive events



(b) Negative events

a target for acquisition or foster competitive forces (also referred to as co-opetition, Gnyawali and Park, 2011). We encourage future work to further explore these preliminary findings.

Another possibility to reduce adverse market reactions around R&D announcements is to manage investor expectations in a way that they reflect objective evidence. Overconfidence frequently attributed to financial investors (Allen and Evans, 2005) has been assumed to be one reason for the observed return gap between positive and negative events (A. Sharma and Lacey, 2004). If part of the price drop that follows negative events is due to investors' overly optimistic speculations, then companies announcing interim results should try to avoid fostering overconfidence of financial analysts, as it may backfire in the long run.

In this work, we show that using two sophisticated measures on the market specific product success probability and the product portfolio importance helps us to better explain market reactions that follow from innovation announcements. Yet, there still remains a substantial share of announcement risk that needs to be properly managed. The combination of data mining tools and high-frequency real-time information that relates to market expectations could become an interesting venue to elevate our methodology further.

Chapter 6

Conclusion

The field of economics has developed a long tradition of approaching complex real-world relationships by formulating mathematical models that naturally require simplifying assumptions. An especially often debated set of assumptions concerns how to model the human side of the equation. Over the years, economists have increasingly refrained from the paradigm that human decision-making across settings can primarily be explained by simple forms of utility maximization. Breaking with this premise, understandably raises new questions, for example, "how does decision-making and behavior change given different situations?" or, if we assume decisions are not optimal, "how can they be improved?".

These questions become particularly relevant in environments for which decision-making is not straightforward. For example, competitive environments with uncertain outcomes potentially hinder successful rational decision-making (Dosi et al., 2001). If so, then we should carefully investigate such environments to better understand their impact on individual decisions and, moreover, provide means for improving them. The first part of the thesis approaches the challenges that situations of outcome uncertainty pose on individuals by analyzing experimental data of contest experiments. Chapter 2 provides insights on how individual efforts depend on the structure of the contest, whereas chapter 3 discusses an econometric approach to develop a taxonomy of contestants based on

their observed behavior. In the second part of the thesis we present a use case for improving decision-making in a real-world setting that is characterized by high outcome uncertainty: the product development process of the pharmaceutical industry. Chapter 4 investigates the use of several supervised learning methods to classify the stage-dependent success of pharmaceutical projects in development. Chapter 5 picks up the previously developed methodological approach to explain the drivers of market reactions after product-specific innovation announcements. The following concluding remarks summarize for each chapter its findings and briefly discuss its implications, limitations, and suggestions for future extensions.

Chapter 2 analyzes the efforts of individuals that repeatedly compete in contest situations by means of a laboratory experiment. The contest settings vary in the number of contestants per group (three and five) and the prize assignment mechanism (lottery and share contest). That way we can analyze aggregate behavioral differences that emerge from differences in competition and outcome uncertainty. We find that zero expenditures are more frequent in situations with high outcome uncertainty and high competition. The increasing share of zero expenditures in five-player lottery contests also explains, contrary to theoretical predictions, why we find no significant difference between average group expenditures between three-player and five-player lottery treatments. Since only a fraction of zero expenditures can be explained by weighted forms of fictitious play, we estimate an experience weighted attraction model (C. Camerer and T. H. Ho, 1999) to better understand differences in learning processes across contest settings. Consistent with the increasing share of zero expenditures, we find that lottery contestants rather apply experiential learning rules that, with the accumulation of lost rounds, tend to discourage positive expenditure levels.

Our finding that average group expenditure in lotteries does not significantly increase with group size has direct implications for contest designers depending on whether their goal is to maximize or minimize overall contestant effort. A contest designer who aims at increasing participation can do so by providing multiple prizes, as it is common in lotteries or raf-

fles, by making the prize more desirable, or by making participation less costly. In the pharmaceutical industry, for example, regulators promote research on rare diseases by granting companies tax deductions (i.e. substitutions of efforts) and market exclusivity (i.e. higher prize) (Seoane-Vazquez et al., 2008). Continuing in the spirit of applying our findings to the pharmaceutical drug development, which is exemplary for a competitive environment under outcome uncertainty (Sutton, 1998), we would expect that misallocation in terms of frequent R&D overspending occurs also on the company level. Especially in competitive research areas we could expect individual firms to spent excessively in order to become the first to develop a breakthrough treatment. Only after considerable time, after having learned from previous investment decisions companies would abstain from unprofitable investments. In fact, there are numerous industry examples in which companies collectively invest in the same area, such as gene therapy, with limited success (another prominent example in the pharmaceutical context is the accumulation of Alzheimer trials based on the amyloid hypothesis, Makin, S., 2018).

Another finding is that less complex decision rules better explain the observed choices for situations in which outcome uncertainty is high (see also Alós-Ferrer and Ritschel, 2018). Here future work needs to further investigate if the observed behavior is due to the decision formation mechanism we propose, or if the observed behavior is caused by other underlying choice rules. More generally speaking, experimental economics has made substantial contributions to show how individuals behave compared to theoretical predictions. Yet, since the decision-making process is commonly not be measured, motivating the reasons behind the observed behaviors stays vague by construction. Besides exploring behavioral differences in various situations, it becomes essential to find out why these differences occur.

In chapter 3 we develop a player taxonomy for contest experiments using the heterogeneity of observed individual efforts. We use the C-Lasso method (Su et al., 2016) to shrink individual regression coefficients to a set of group coefficients and thus categorize contestants into types based on their reactions to previously received information. For this

purpose we expand the base model in (Su et al., 2016) to be used in all sorts of Tobit applications in which the range of the dependent variable is restricted. We analyze data of six contest experiments that differ in whether groups of four contestants stay fixed over the experiment or are randomly reassigned after each round. In both contest regimes the optimal number of latent types is estimated to be three. In fixed matched treatments the largest group of contestants shows reciprocal traits. Contest groups with more *reciprocators* display lower average efforts, which hints at the collaborative nature of this type. For treatments in which contestants are randomly matched after each round, *reciprocators* appear less frequent. Overall, we find that identified types consider information of previous rounds differently than predicted by myopic best reply considerations.

Our contribution stresses the importance of group heterogeneity that tends to be overlooked in mainstream regression analysis with focus on average effects. If we have reason to believe that a dependent variable relates systematically different to a set of explanatory variables for different unobserved groups of observations, then implications based on average effects can be misleading. Game theoretical models vastly acknowledge that individual decision rules differ in their sophistication (e.g. Stahl, 1993; Stahl II and Wilson, 1994; C. F. Camerer et al., 2004), which makes correct type identification particularly important in game theoretical experiments. In many experiments, behavioral heterogeneity is rather the norm than the exception. Yet average effects across participants are frequently used to report results which resonates a notion of participant homogeneity that might be too simplistic. To foster a discussion on player types in contest experiments, we bundle contest data from different experiments and apply a novel econometric method that identifies latent type structures in panel data. One downside of such an ensuing data-driven approach is that differences in identified types cannot be further explored during the course of the experiment. Future work on similar experiments could in a first step identify player types and next explicitly investigate the impact of different type compositions on group behavior. Based on our findings, we expect that the group composition

influences group expenditure dynamics. For example, it would be interesting to observe the effort dynamics in contest experiments that emerge over time when exogenously assigning groups based on previously identified types. A similar line of thoughts is worth to explore across game theoretical experiments that involve group settings.

The first half of the thesis showed that even in controlled experimental environments individual decisions in competitive situations under outcome uncertainty are heterogeneous and often deviate from standard theoretical predictions. In chapter 4 we therefore explore how to support decision-makers in similar, real-world situations. As our target setting we choose the pharmaceutical drug development for which we attempt to predict stage-specific development outcomes exploiting a rich information base by using supervised learning algorithms. Each algorithm is trained on a large feature set involving company-, product-, market-, and regulatory characteristics. We find that the BART algorithm (Kapelner and Bleich, 2013) provides the most accurate classification results across a variety of methods and is thus used to explore the success rates of the current development pipeline and to assess which types of products are predicted to be most successful.

Estimating with higher accuracy the development outcome of drugs under development is of substantial interest for pharmaceutical companies, financial analysts, and regulators. It is therefore not surprising that Evaluate Ltd. (Evaluate Ltd., 2019), with whom we frequently exchanged ideas and information over the course of the project, has since then built on our methodological approach to estimate product-specific success probabilities. In fact, the integration of a supervised learning-based classification tool into a dynamic data base would be the next logical step to apply our approach. Outcome predictions that consider multiple informational dimensions are shown to be, on average, more accurate than measures that relate solely to historic success rates. Using sophisticated data-based methods to gauge clinical success may have important implications for various stakeholders in the pharmaceutical industry. For example, the usage of more accurate success measures allows company managers to make more informed decision about the allocation

of resources. It also delivers novel insights to regulators about the successes awaiting in the current development pipeline and provides a tool to assess the prospect of success on the project level, particularly valuable to companies looking for collaboration or acquisition opportunities. Yet, safety and efficacy of a single drug, two dimensions that ultimately determine its approval, are difficult to gauge by solely relying on publicly available information. Supervised learning algorithms that combine high-dimensional data with project-specific expert opinions might be the next step to improve predictions across many applications.

It needs to be stressed that providing decision makers with sophisticated tools that address the probability of R&D success will not trivialize the decision making process on how to allocate resources. First, resource allocation does not solely depend on the odds of success but also on future revenue projections and the strategic interaction with other companies. Second, since each company is different in their in-house expertise and their inherent risk profile, we cannot expect that an entire industry follows collectively the suggestions of artificial intelligence approaches (i.e. investing only in projects that are commonly estimated to be likely to become successful). Therefore it remains ambiguous how the usage of algorithms used to estimate pharmaceutical successes will impact the R&D pipeline on an industry level and how this will affect society as a whole. In fact, the impact that artificial intelligence will continue to have on our societies is multifaceted, difficult to objectively assess, even more difficult to quantify, and thus remains an interesting and important venue for further research.

The outcome uncertainty of the pharmaceutical drug development process directly translates to substantial reactions of companies' stock prices that follow product innovation announcements. In chapter 5 we aim at identifying common effects that explain these reactions. According to our theoretical framework, company stock market reactions that follow from project innovation announcements should be subject to a probability effect, that is the change in the perceived probability of market launch, and a portfolio effect, that is the relative importance of a product in a companies portfolio in terms of future cash flows. Measuring

both effects directly poses practical difficulties, which we try to overcome by estimating the project-specific success probability using a combination of supervised learning methods, and by measuring a products portfolio importance via net present values. We find our hypotheses confirmed, namely that projects with high success probability (portfolio importance) show lower (higher) market reactions after product innovation announcements.

Our results are of applied interest to pharmaceutical companies that want to hedge against negative announcement risks and thus wish to anticipate the possible market impact of public disclosure. Negative project news might lead to a sustained drop in company equity which could, especially for companies with a narrow project portfolio, influence the managerial decision to discontinue a precarious project. Financial investors that aim at buying company stock are another group of stakeholders that benefits from understanding how the success probability of products is perceived in the market. Our methodological approach, which relies on an ensemble learner consisting of several SL-algorithms to gauge project success probabilities, could be of great use, as well. In fact, estimating accurately the probability of success may complement the traditional DCF analysis that focuses on estimating future cash flows but commonly accounts for their risk in a simplistic fashion (i.e. by adding a risk premium to the discount factor). An interesting venue for further research would be to compare different investment strategies in terms of profitability across markets and time, e.g., one strategy using a sophisticated measure of success probabilities, another focusing on accurately forecasting earning trajectories, and a third combining both. Another opportunity for future research concerns improving the measurement of perceived success probabilities and thus increasing the explained variation in market returns after product innovation announcements. To do so, one needs to comprehend how investors establish success probability estimates and how these change with the arrival of new information. Such fundamental questions could be addressed first of all using experimental settings that are able to control the characteristics and timing of new information.

Appendix A

Supplementary Material for Chapter 2

A.1 Risk Assessment

The equilibrium predictions are derived in both contest types under the assumption of risk neutrality. However, difference in individual risk profiles i.e. risk aversion might alter the expected effort levels and hence the conclusions drawn from a direct comparison between contest types.

Theoretical evidence of the effect of risk aversion on efforts is ambiguous (Konrad and Schlesinger, 1997), since results vary with the assumptions on risk aversion, contest success functions, and homogeneity of rent-seekers (Treich, 2010).

Early experimental work by Millner and Pratt, 1991 on lottery contests finds lower mean dissipation rates for risk-averse contest groups, not significantly different from risk-neutral equilibrium predictions. Other authors find risk aversion to significantly reduce efforts in lottery contests, but averages remain still above the risk-neutral equilibrium predictions (e.g. Anderson and Freeborn, 2010; Mago, Sheremeta, et al., 2013; Sheremeta, 2011). In contrast, a direct comparison of lottery and share contests does not find risk aversion to significantly drive down efforts (Shupp et al., 2013).

To control for risk aversion in our experiment, we ask subjects to evaluate how willing they are to take risks, in general, using a Likert scale from 1 (unwilling to take risks) to 7 (fully prepared to take risks). Dohmen et al., 2011 validate this measure in a representative subject pool to measure individual risk inclination. The average "risk score", reported in table A.1, is not significantly different between share contests and lottery contests treatments. Hence, differences in efforts between treatments of equal group size should not be due to different risk score distributions across treatments.

Table A.1: Average responses on risk assessment overall and for each treatment

	Overall	3S	3L	5S	5L
Mean risk score	4.60	4.97	4.35	4.87	4.38
	(0.12)	(0.27)	(0.24)	(0.26)	(0.20)
p-value 3S-3L			0.11		
p-value 5S-5L					0.14

Mean risk score (1 low subjective risk score, 7 high subjective risk score) and standard error (in parentheses). Between-treatment comparisons are based on Pearson's Chi-squared test at individual level.

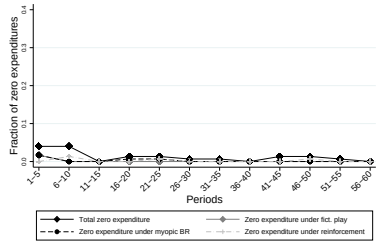
A.2 Additional Tables and Figures

Table A.2: Mean fraction of plays that imitate previous average opponent expenditure (+/-50)

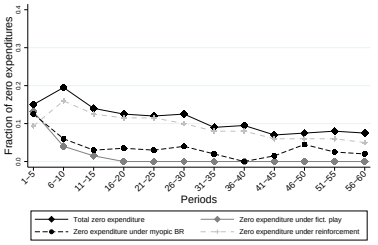
	3S	5S	3L	5L	p-value 3S-3L	p-value 5S-5L
Mean All	0.27	0.25	0.12	0.13	0.00	0.01
Mean 1-30	0.24	0.21	0.10	0.12		
Mean 31-60	0.31	0.29	0.13	0.14		
Within p-value	0.05	0.12	0.39	0.26		

Reported p-values for within-treatment comparisons between the two halves of the experiment are based on Wilcoxon matched-pairs signed-rank tests. Reported p-values for between-treatment comparisons are based on two-sided Wilcoxon rank-sum tests.

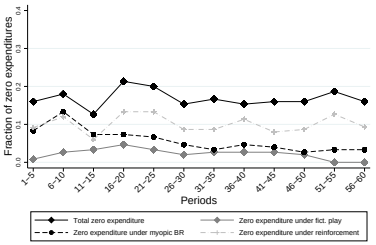
Figure A.1: Fraction of zero expenditures across treatments over time - including explained zeros under reinforcement learning



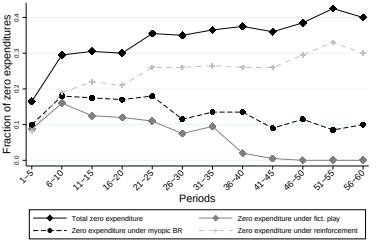
(a) 3S



(b) 5S



(c) 3L



(d) 5L

Table A.3: EWA estimation results across treatments (no clustering adjustment of standard errors)

Treatment	3S		5S		3L		5L	
Periods	1-30	All	1-30	All	1-30	All	1-30	All
$N0$	0.26**	0.19	0.14	0.14	0.53***	0.18***	0.40**	0.08
	(0.11)	(0.14)	(0.10)	(0.11)	(0.13)	(0.07)	(0.08)	(0.10)
	[0.11 0.52]	[0.04 0.58]	[0.03 0.46]	[0.03 0.48]	[0.29 0.75]	[0.08 0.35]	[0.13 0.75]	[0.01 0.57]
κ	1.00***	0.95***	0.55***	0.58***	1.00***	0.44***	1.00***	0.64**
	(0.00)	(0.34)	(0.16)	(0.16)	(0.00)	(0.17)	(0.00)	(0.33)
	[1.00 1.00]	[0.08 1.00]	[0.26 0.81]	[0.27 0.84]	[1.00 1.00]	[0.17 0.75]	[1.00 1.00]	[0.10 0.97]
ϕ	0.73***	0.75***	0.75***	0.75***	0.71***	0.79***	0.79***	0.82***
	(0.03)	(0.02)	(0.02)	(0.01)	(0.02)	(0.01)	(0.02)	(0.01)
	[0.67 0.78]	[0.71 0.78]	[0.71 0.79]	[0.72 0.78]	[0.67 0.75]	[0.76 0.82]	[0.75 0.82]	[0.80 0.84]
λ	0.80***	0.77***	0.70***	0.69***	0.88***	0.83***	0.88***	0.86***
	(0.02)	(0.05)	(0.04)	(0.04)	(0.01)	(0.04)	(0.01)	(0.05)
	[0.76 0.84]	[0.68 0.88]	[0.62 0.78]	[0.61 0.78]	[0.86 0.90]	[0.76 0.90]	[0.87 0.90]	[0.77 0.96]
δ	0.61***	0.67***	0.59***	0.62***	0.02	0.11***	0.13***	0.18***
	(0.03)	(0.02)	(0.02)	(0.02)	(0.06)	(0.04)	(0.04)	(0.03)
	[0.55 0.66]	[0.63 0.70]	[0.54 0.63]	[0.59 0.65]	[0.00 0.84]	[0.06 0.21]	[0.06 0.24]	[0.13 0.25]
$Obs.$	900	1800	1200	2400	900	1800	1200	2400
$-LL$	1494.0	2677.3	1766.6	3131.6	1444.3	2768.7	1824.0	3210.6
AIC	2998.0	5364.7	3543.2	6273.2	2898.6	5547.4	3658.0	6431.2
BIC	3022.0	5392.2	3568.6	6302.2	2922.6	5574.8	3683.5	6460.1

EWA Maximum Likelihood estimation with restricted parameters reported. Standard errors of the restricted metric, in parenthesis; 95% confidence interval of restricted estimates in brackets. P-values ($H_0 = 0$): *$\leqslant$ 0.10, **$\leqslant$ 0.05, ***$\leqslant$ 0.01, Obs. is the number of observations, -LL is the negative log-pseudolikelihood, AIC and BIC are the Aikaike- and Bayesian information criterion, respectively.

Table A.4: Delta estimates for different EWA model specifications

δ estimates	$N0$=0	$N0 = 0$, $Attr_0 = 0$	$N0 = 0$, $Attr_0 = 0$, $\kappa = 0$	Reported δ values in main text
3S 1-30	0.62	0.62	0.65	0.61
3S All	0.67	0.67	0.68	0.67
5S 1-30	0.59	0.59	0.58	0.59
5S All	0.62	0.62	0.61	0.62
3L 1-30	0.04	0.04	0.05	0.02
3L All	0.12	0.12	0.11	0.11
5L 1-30	0.14	0.14	0.15	0.13
5L All	0.19	0.19	0.17	0.18

Table A.5: Random effects logit regression on zero expenditures in lottery contests

Dependent variable:	3L		5L	
Zero expenditure$_t$	(1)	(2)	(3)	(4)
Expenditures$_{t-1}$	-1.38***	-1.31**	-1.02***	-0.61*
	(0.53)	(0.52)	(0.35)	(0.35)
Other	2.00*	1.84*	1.00**	0.69*
Expenditures$_{t-1}$	(1.12)	(1.10)	(0.43)	(0.41)
Other	-0.70*	-0.65	-0.18	-0.10
Expenditures$^2_{t-1}$	(0.42)	(0.42)	(0.14)	(0.13)
Period	-0.00	-0.01	0.03***	0.03***
	(0.01)	(0.01)	(0.01)	(0.01)
Loss streak		-0.21		-0.58***
		(0.14)		(0.14)
Constant	-3.71***	-3.73***	-3.03***	-3.39***
	(0.98)	(0.97)	(0.66)	(0.70)
$Obs.$	1770	1770	2360	2360

Random effects logit regression; Standard error clustered at player level in parenthesis. P-values: *$\leqslant$ 0.10, **$\leqslant$ 0.05, ***$\leqslant$ 0.01, Obs. is the number of observations. All effort related regressors are normalized (divided by 1000).

A.3 Instructions

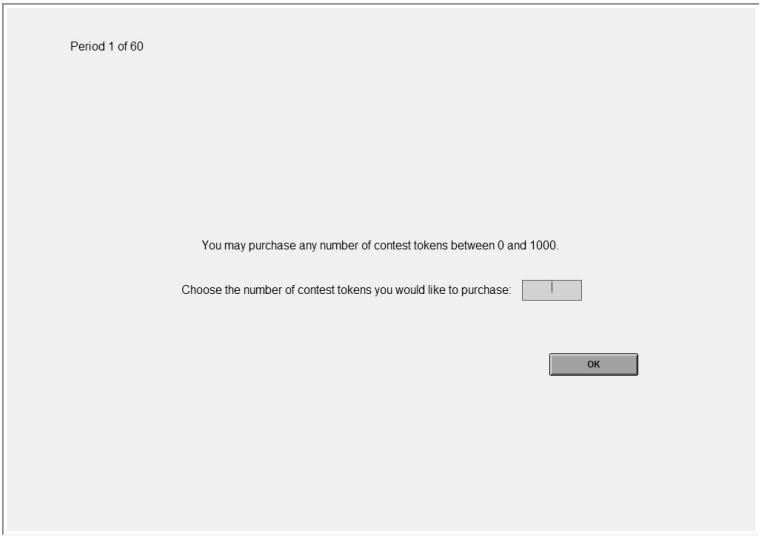
Welcome! You are about to participate in an experiment in the economics of decision making. Please do not talk to any of the other participants until the experiment is over. If you have a question at any time please raise your hand and an experimenter will come to your desk to answer it. The experiment will consist of 60 periods. In each period you will have the chance to earn points. At the end of the experiment each participant's accumulated point earnings from all periods will be converted into cash at the exchange rate of 0.015 pence per point. Each participant will be paid in cash and in private.
At the beginning of the experiment you will be matched with two [four] other people, randomly selected from the participants in this room, to form a group of three [five]. The composition of the group will stay the same throughout the experiment, i.e. you will form a group with the same two [four] other participants during the whole experiment. Your earnings will depend on the decisions made within your group, as described below. Your earnings will not be affected by decisions made in other groups. All decisions are made anonymously and you will not learn the identity of the other participants in your group.

Decision task in each period

Each period has the same structure. In each period the three [five] participants in each group will be competing for a prize of 1000 points.
At the beginning of the period each participant will be given an endowment of 1000 points. Each participant has to decide how many of these points they want to use to buy "contest tokens". Each contest token costs 1 point, so each participant can purchase up to 1000 of these tokens. Any part of the endowment that is not spent on contest tokens is kept by the participant. Each participant must enter his or her decision via the computer. An example screenshot is shown below.

Once everybody has chosen how many contest tokens to purchase, the computer will calculate each participant's share of the prize of 1000 points.

Figure A.2: Experiment screenshot- expenditure choice



Your share of the prize will depend on how many contest tokens you have purchased and the total number of contest tokens purchased in your group.

[In Share Contest Description]
If nobody in your group purchases any contest tokens, none of you will receive a share of the prize. Otherwise, the computer will calculate each participant's share of the prize so that your share of the prize will be equal to the number of contest tokens that you have purchased divided by the total number of contest tokens purchased in your group. That is, if you buy a number of X contest tokens and if the other two participants in your group buy Y and Z contest tokens each, then your share of the prize will be X/(X+Y+Z). [That is, if you buy a number of V contest tokens and if the other four participants in your group buy W, X, Y and Z contest tokens each, then your share of the prize will be V/(V+W+X+Y+Z).] Your contest earnings will be your share times 1000 points (rounded to the nearest point).

[In Lottery Contest Description]
If nobody in your group purchases any contest tokens, none of you will win the prize. Otherwise, the computer will determine which participant wins the prize in a way that will ensure that the probability that you will win the prize is equal to the number of contest tokens that you have purchased divided by the total number of contest tokens purchased in your group. That is, if you buy a number of V contest tokens and if the other two participants in your group buy W and X contest tokens each, then the probability that you win the prize will be V/(V+W+X). [That is, if you buy a number of V contest tokens and if the other four participants in your group buy W, X, Y, Z contest tokens each, then the probability that you win the prize will be V/(V+W+X+Y+Z).] Your contest earnings will be either 0 (if you do not win the prize), or 1000 (if you win the prize).

Your point earnings for the period will be calculated as follows:

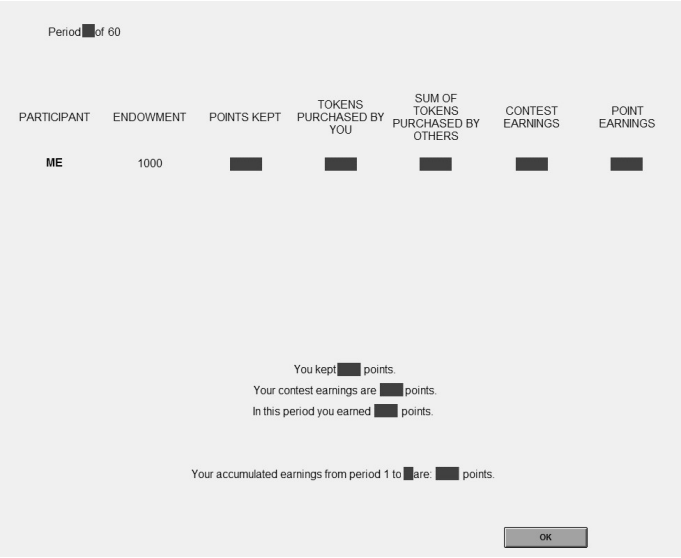
point earnings = 1000 – contest tokens purchased + contest earnings

After all participants have made a decision, a result screen will appear. An example screenshot is shown below. This is like the screen you will see during the experiment except that the blacked out fields will be filled in according to the decisions made in that round.

Each participant will be informed of the points remaining from their endowment after making their purchase, the number of contest tokens they have purchased, the sum of tokens purchased by the other participants in their group, their contest earnings and their point earnings for the period. In addition, the results screen will inform each participant of his or her accumulated points from all periods so far.

Beginning of the experiment
If you have any questions please raise your hand and an experimenter will come to your desk to answer it.

Figure A.3: Experiment screenshot- information feedback



We are now ready to begin the decision-making part of the experiment. Please look at your computer screen and begin making your decisions.

A.4 QRE Model

In result section 2.4.2 we report that choices made by lottery contestants rely more on own past experience than the ones of share contestants. This supplementary section adds to the behavioral differences across treatments by exploring the degree of contestants' "sophistication" using the homogeneous Quantal Response Equilibrium (QRE) model of McKelvey and Palfrey, 1995. Sophisticated players base their choices on the evaluation of expectations on future actions of opponents (C. F. Camerer et al., 2002). The estimation results complement the EWA findings: sophistication increases over time, yet share contestants act more sophisticated than lottery contestants.

$$P(x_k) = \frac{e^{\lambda^Q E_p[\pi(x_k)]}}{\sum_{j=1}^{K} e^{\lambda^Q E_p[\pi(x_j)]}} \qquad \text{for} \quad k = 1 \cdots 11 \tag{A.1}$$

Let $P(x_k)$ denote the probability of choosing the k^{th} bin x_k for an arbitrary player. Let $E_p[\pi(x_k)]$ be the expected payoff from choosing bin x_k conditional on all $N-1$ opponents playing the mixed strategy p over the K bins.[1] The K probabilities are the solutions of a logit QRE specification in equation A.1. $\lambda^Q \in [0, \infty)$ is called precision parameter. While λ in the EWA model measures the impact of past attractions, λ^Q measures the impact of expectations on choice probabilities. Table A.6 describes all reported QRE parameters.

The estimation results in table A.7 show that the lottery treatments exhibit a lower value of λ^Q, indicating that players' choices depend less on payoff expectations given the anticipation of opponents mixed strategies and are therefore, on average, less sophisticated.

We use likelihood ratio tests to verify heterogeneity in λ^Q within- and between-treatments (see Lim et al., 2014. To test whether the impact of expected payoffs on choices changes over time, we compare the log-likelihood of a constant λ^Q across all periods with the log-likelihood of a differing λ^Q between the two halves of the session (equation A.2).

[1] To be coherent with the EWA estimation, we use 11 bins of equal distance.

Table A.6: Description of reported QRE values

Parameter	Description
λ^Q	Sensitivity measure of attractions. The higher λ^Q, the more best-responses rely on forecasts of opponents behavior. For $\lambda^Q = 0$, equation A.1 simplifies to $P(x_k) = 1/K$, i.e. every choice is equally likely to occur and does not depend on the relative expected payoff. Contrariwise, as $\lambda^Q \to \infty$ the QRE prediction converges to the Nash equilibrium.
$-LL$	λ^Q is chosen such that it maximizes the log-likelihood of the theoretical choice probabilities of the QRE model given real choice frequencies. The corresponding log-likelihood value is given by $-LL$.
$-LL_{Uni}$	The log-likelihood that would result from all choices equally likely.
$-LL_{Max}$	The log-likelihood that would result from an exact correspondence between model estimations and the observed data.
Q	Q-statistic ($\frac{LL-LL_{Uni}}{LL_{Max}-LL_{Uni}}$) measures the fit of the QRE model according to Lim et al., 2014. If $Q = 0$, then the best fit is achieved using a uniform choice distribution. If $Q = 1$, the model predicts perfectly the empirical distribution.

Similarly, for between-treatment comparisons (3L-5L, 3S-5S, 3S-3L and 5S-5L) we take for each λ^Q the sum of the log-likelihood of both treatments and find the new optimal λ^Q and its corresponding log-likelihood (equation A.3). We then compare the resulting LL_{T1+T2} with $LL_{T1} + LL_{T2}$. Meaning we estimate the λ^Q parameters for each treatment separately and then sum the corresponding log-likelihood estimations. Both test statistics are asymptotically chi-squared distributed with 1 degree of freedom. The summary of all likelihood ratio tests can be found in table A.8.

$$D(within) = -2\big[LL_{ALL} - (LL_{1-30} + LL_{31-60})\big] \tag{A.2}$$

$$D(between) = -2\big[LL_{T1+T2} - (LL_{T1} + LL_{T2})\big] \tag{A.3}$$

Estimations show a significant increase in the precision parameters over time: the estimated λ^Q in 3S increases from 0.33 in the first half to 0.76

in the second half ($D = 109.8$, $p = 0.00$). Comparable increases are observed for all other treatments (5S: from 0.31 to 0.51 ($D = 49.3$, $p = 0.00$), 3L: from 0.09 to 0.14 ($D = 18.8$, $p = 0.00$), 5L: from 0.16 to 0.24 ($D = 35.0$, $p = 0.00$).

When comparing the results for different group sizes of the same contest, one might expect that an increase in group size could make the game more complex since the possible set of opponents' expenditures increases, leading thus to a lower lambda in larger groups.[2] We find support for this assumption in the overall periods of the share contest ($D = 18.6$, $p = 0.00$), yet in the first half of the share contests results are not significant ($D = 1.0$, $p = 0.31$). In the lottery contest we find significantly higher λ^Q for the five-player treatment (1st half: $D = 42.3$, $p = 0.00$, all: $D = 87.0$, $p = 0.00$).

Using the same approach for a between contest type comparison, we discover that λ^Q is significantly higher for the share contest in all evaluated treatments and subsamples (1st half three-player: $D = 185.5$, all three-player: $D = 509.1$, 1st half five-player: $D = 85.4$, all five-player: $D = 178.1$). This suggests that expected payoffs matter less for lottery contest players when deciding on their next investment, especially during early periods. Similar to Chowdhury, Sheremeta, et al., 2014, we find a higher Q-statistic for the share contest in all treatments suggesting that the estimated QRE model better fits the share contest expenditures. Expenditures in the lottery contain a higher share of zero expenditures, which may explain the lower goodness of fit.

Stylizing the main results of the QRE model, we find that players across treatments become more sophisticated over time meaning that they increasingly rely on forecasting opponents' behavior when determining their choices. Additionally, a higher number of players leads to an increase in λ^Q for lottery treatments, while the effect in the share contest is ambiguous. The choice patterns observed in lottery treatments display less sophistication than in the share contest which might be due to the

[2] Lim et al., 2014 estimate a QRE model for Tullock contests and find a lower lambda in larger groups.

probabilistic prize assignment that makes it challenging to form sophisticated expectations.

Table A.7: QRE estimation results across treatments

	Treatment	3S		5S		3L		5L	
	Periods	1-30	All	1-30	All	1-30	All	1-30	All
QRE	λ^Q	0.33	0.47	0.31	0.38	0.09	0.11	0.16	0.20
	$-LL$	1816	3403	2279	4412	2097	4126	2567	4926
	$-LL_{Uni}$	2158	4316	2877	5755	2158	4316	2877	5.755
	$-LL_{Max}$	1774	3328	2184	4296	1921	3862	2327	4350
	Q	0.89	0.92	0.86	0.92	0.26	0.42	0.57	0.59

λ^Q is the precision parameter of the QRE model which maximizes the log-likelihood (LL) to observe the relative frequency of actual choices. See table A.6 for further parameter explanations.

Table A.8: QRE model- likelihood ratio test results

Treatment	Period	λ^Q	-LL	D-statistic	p-value
3S	1-30	0.33	1816		
	31-60	0.76	1532	109.8	0.000
	All	0.47	3403		
5S	1-30	0.31	2279		
	31-60	0.51	2109	49.3	0.000
	All	0.38	4412		
3L	1-30	0.09	2097		
	31-60	0.14	2020	18.8	0.000
	All	0.11	4126		
5L	1-30	0.16	2567		
	31-60	0.24	2342	35.0	0.000
	All	0.20	4926		
5S+3S	1-30	0.32	4095	1.0	0.307
	31-60	0.60	3654	26.9	0.000
	All	0.42	7824	18.6	0.000
5L+3L	1-30	0.13	4685	42.3	0.000
	31-60	0.20	4384	44.2	0.000
	All	0.16	9096	87.0	0.000
3L+3S	1-30	0.16	4005	185.5	0.000
	31-60	0.27	3746	386.9	0.000
	All	0.21	7784	509.1	0.000
5L+5S	1-30	0.22	4888	85.4	0.000
	31-60	0.32	4506	112.5	0.000
	All	0.26	9427	178.1	0.000

λ^Q is the precision parameter of the QRE model which maximizes the log-likelihood (LL) to observe the relative frequency of actual choices. The D-statistic for the first four treatments is the within-LR test statistic of equation A.2. The D-statistic for the last four treatment combinations is based on the between-LR test statistic of equation A.3.

A.5 Mathematical Demonstrations

A.5.1 Nash Equilibrium of Share and Lottery contests

1. Show equality of expected values in lottery and share contest

The expected payoff of player i in the lottery contest is given by

$$\begin{aligned} E(\pi_i) &= \frac{x_i}{X}(e - x_i + V) + \frac{X - x_i}{X}(e - x_i) \\ &= \frac{x_i}{X}V + \frac{X}{X}(e - x_i) \\ &= e - x_i + V(\frac{x_i}{X}) \end{aligned}$$

which is the individual payoff function of the share contest. Since the share contest is not probabilistic, its payoff function is equivalent to its expected payoff function.

2. Derive the Nash equilibrium starting from the expected value

Replace X by $\sum_{i=1}^{N} x_i$. The expected value of payoff function can then be expressed as:

$$E(\pi) = \frac{x_i}{\sum_{i=1}^{N} x_i}V - x_i + e$$

The First order condition with respect to own effort x_i is given by:

$$\begin{aligned} FOC_{x_i} = \frac{\sum_{i=1}^{N} x_i - x_i}{(\sum_{i=1}^{N} x_i)^2}V - 1 &\stackrel{!}{=} 0 \\ (\sum_{i=1}^{N} x_i - x_i)V &= (\sum_{i=1}^{N} x_i)^2 \\ V\sum_{i=1}^{N} x_i - (\sum_{i=1}^{N} x_i)^2 &= x_i V \end{aligned}$$

In equilibrium this expression needs to hold for every individual i. Since the sum of expenditures (left term) is the same for all i, due to equality also the right term is the same for all i. We thus replace x_i with x^* and obtain:

$$\begin{aligned} V\sum_{i=1}^{N} x^* - (\sum_{i=1}^{N} x^*)^2 &= x^*V \\ Vnx^* - (nx^*)^2 &= x^*V \\ V(n-1)x^* &= (nx^*)^2 \\ \frac{V(n-1)}{n^2} &= x^* \end{aligned}$$

which is the Nash equilibrium for individual efforts of the lottery and share contest. The equilibrium contribution for n players in a group is then $x^* = nV(n-1)/n^2$. Note that for an increase in the number of players, the individual level of equilibrium efforts decreases but the level of group effort increases.

A.5.2 Myopic Best Response Function

The best response (BR) function relates the opponent efforts c to the payoff maximizing level of own efforts x. A myopic best response considers only the opponent efforts of the last round when choosing an own effort. Here we show how the best response function is calculated and why it is, according to the myopic BR, not optimal to invest if aggregate opponent efforts were higher in the previous round than the prize V. We start with the expected value of the contests:

$$E(\pi(x)) = e - x + V(\frac{x}{x+c})$$

Since we are interested in the (expected) payoff maximizing level of own expenditures, we take the derivative with respect to x and set it equal to zero.

$$\frac{\partial E(\pi(x))}{\partial x} = -1 + \frac{V}{x+c} - \frac{Vx}{(x+c)^2} \stackrel{!}{=} 0$$

$$(x+c)^2 = Vc$$

$$x_{1,2} = -\frac{2c}{2} \pm \sqrt{(\frac{2c}{2})^2 - (c^2 - Vc)}$$

$$x_{1,2} = -c \pm \sqrt{Vc}$$

Since x and c are ≥ 0, only the solution $x_1 = -c + \sqrt{Vc}$ is of practical relevance. The best response function given opponent effort level c and prize V is hence given by $BR(c) = -c + \sqrt{Vc}$. If $c = V$, then the best response is zero, meaning not to put any effort in the contest. In case $c > V$, the $BR(c)$ becomes negative, but since efforts are by definition greater or equal to zero, zero efforts remain, in practice, the best response. If an individual expects $c \geq V$, where in our experiment $V = 1000$, then own efforts greater zero are not maximizing the expected payoff.

Appendix B

Supplementary Material for Chapter 3

B.1 Additional Tables and Figures

Table B.1: Likelihood ratio test: p-values for model selection

Additionally included variables	FM	RM
$l.othereffort^2$	0.02	0.01
$l2.othereffort$	0.12	0.26
$l2.othereffort, l2.othereffort^2$	0.20	0.41

Table B.1 shows the p-values of the likelihood ratio tests between the base model and models with additionally included variables for FM and RM treatments. For p-values below 0.05, we reject the Null hypothesis that both models are equivalent in terms of log likelihood and include the additional variable.

Table B.2: Tobit regression estimates for FM type "others"

Dep. variable: *effort*	*(1)*	*(2)*	*(3)*
l.othereffort	0.031	0.088	0.116
	(0.102)	(0.1044)	(0.105)
$l.othereffort^2$	0.008	-0.0132	-0.026
	(0.039)	(0.040)	(0.040)
l2.othereffort	0.370***	0.375***	0.379***
	(0.099)	(0.102)	(0.103)
$l2.othereffort^2$	-0.120***	-0.125***	-0.126***
	(0.037)	(0.038)	(0.039)
period	-0.001	0.000	0.001
	(0.003)	(0.003)	(0.003)
l.win	0.105***	0.114***	-
	(0.034)	(0.035)	
l2.win	0.179***	-	-
	(0.034)		
σ_ϵ	0.342***	0.350***	0.354***
	(0.011)	(0.012)	(0.012)
Obs; N	558;31	558;31	558;31

Standard errors in parenthesis; p-values: *$\leqslant$ 0.10, **$\leqslant$ 0.05, ***$\leqslant$ 0.01; Obs. is the total number of observations; N is the number of contestants in each group.

Figure B.1: Marginal effect of l.othereffort on effort in FM treatments

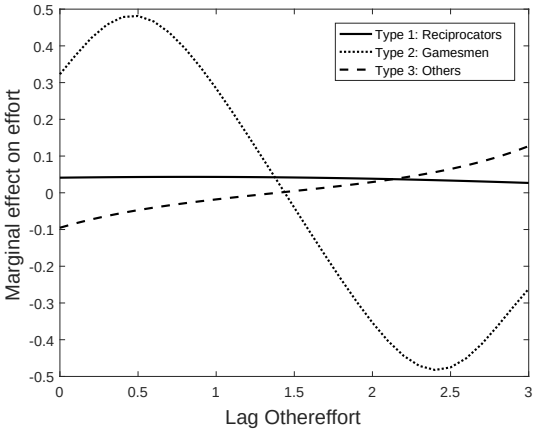
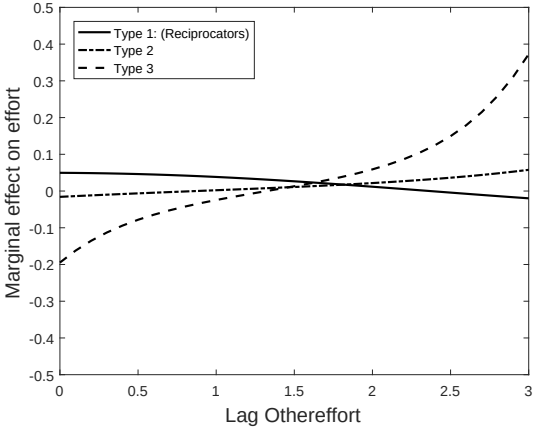


Figure B.2: Marginal effect of l.othereffort on effort in RM treatments



Appendix C

Supplementary Material for Chapter 4

C.1 Hyperparameter Tuning

More complex SL-algorithms often rely on an array of exogenously supplied parameters that remain unchanged during the actual training procedure. These so called hyperparameters are usually estimated using cross validation approaches and chosen such that they minimize the out of sample error. For PROBIT and DT there are no hyperparameters available to tune. For the other algorithms we restrict hyperparameter tuning to one (RF, C5.0) or two values (BART) for each data set, because we observe that the performance gain from fine-tuning hyperparameters is low while computation times are high. All hyperparameters are tuned using five-fold cross validation. Hyperparameters that are not tuned are kept at their default values specified by the respective *R* package. The hypertuning results of the BART cut-off probability, the BART number of trees, the RF randomly sampled variables at each split, and the C5.0 number of boosting iterations are shown in table C.1, table C.2, table C.3, and table C.4, respectively.

Table C.1: Hyperparameter tuning: Error ratio for BART cut-off probability

	Cut-off probability value				
	0.3	0.4	0.5	0.6	0.7
PI	0.1714	0.1596	**0.1524**	0.1598	0.183
PII	0.119	0.108	**0.0974**	0.1084	0.1334
PIII	0.1442	**0.1082**	0.1212	0.1492	0.2082
Filed	0.033	0.033	**0.033**	0.033	0.033

Table C.1 shows the out-of-sample error ratio for BART cut-off probability values at each data set. The lowest error ratio for each data set is highlighted in bold.

Table C.2: Hyperparameter tuning: BART number of trees

	Number of trees			
	50	100	150	200
PI				
Accuracy	0.8499	0.8554	**0.8595**	0.8562
Accuracy SD	0.0076	0.0152	0.0077	0.0065
PII				
Accuracy	0.8990	**0.9006**	0.8995	0.8975
Accuracy SD	0.0076	0.0069	0.0101	0.0094
PIII				
Accuracy	0.8750	**0.8767**	0.8701	0.8701
Accuracy SD	0.0545	0.0413	0.0382	0.0432
Filed				
Accuracy	0.9671	0.9671	0.9671	**0.9671**
Accuracy SD	0.0066	0.0066	0.0066	0.0066

Table C.2 shows the accuracy and its standard deviation (SD) for BART number of tree values at each data set. The highest accuracy for each data set is highlighted in bold.

Table C.3: Hyperparameter tuning: RF number of sampled variables at each split

	Number of sampled variables at each split			
Accuracy (SD)	7	8	9	10
PI	0.8407 (0.0120)	0.8411 (0.0138)	**0.8415** (0.0135)	0.8396 (0.0147)
PII	0.8708 (0.0194)	0.8688 (0.0196)	**0.8739** (0.0204)	0.8719 (0.0159)
PIII	0.8455 (0.0390)	0.8422 (0.0380)	0.8439 (0.0364)	**0.8472** (0.0353)
Filed	0.9672 (0.0160)	0.9672 (0.0160)	**0.9702** (0.0179)	0.9642 (0.0129)

Table C.3 shows the accuracy and its standard deviation (SD) for RF number of sampled variables at each split. The highest accuracy for each data set is highlighted in bold.

Table C.4: Hyperparameter tuning: C5.0 number of boosting iterations

	PI		PII		PIII		Filed	
Iterations	ACC	SD	ACC	SD	ACC	SD	ACC	SD
10	0.7930	0.0211	0.8473	0.0154	0.7948	0.0288	**0.9789**	0.0173
20	0.8040	0.0162	0.8462	0.0205	0.7948	0.0307	0.9760	0.0172
30	0.8073	0.0195	0.8498	0.0183	0.8063	0.0233	0.9760	0.0172
40	0.8114	0.0150	0.8514	0.0231	0.8112	0.0197	0.9760	0.0172
50	0.8114	0.0161	0.8519	0.0183	**0.8128**	0.0210	0.9760	0.0172
60	0.8099	0.0175	0.8493	0.0195	0.8095	0.0219	0.9760	0.0172
70	0.8121	0.0172	0.8503	0.0174	0.7997	0.0238	0.9760	0.0172
80	0.8125	0.0180	0.8524	0.0157	0.8030	0.0217	0.9760	0.0172
90	**0.8139**	0.0182	0.8514	0.0131	0.8030	0.0255	0.9760	0.0172
100	0.8121	0.0199	**0.8529**	0.0125	0.8047	0.0220	0.9760	0.0172

Table C.4 shows the accuracy (ACC) and its standard deviation (SD) for C5.0 number of boosting iterations. The highest accuracy for each data set is highlighted in bold.

C.2 Additional Tables and Figures

Table C.5: Inclusion and exclusion criteria of data in main dataset

Description	The dataset includes FDA monitored drugs of active pharmaceutical companies with at least one reported indication status in the US.
Excluded	**Included**
Companies outside the pharmaceutical sector: • AGBio & Chemicals • Food & Consumer • Hospitals • Government Agencies • Non-profit • Universities • Investors	*Companies within the pharmaceutical sector:* • Global Majors (Pharma) • Regional Majors (Pharma) • Speciality & Genomics (Pharma) • Biotechnology
Firms with no core business in R&D and marketing of pharmaceuticals: • MedTech • CRO • API (Manufacturers)	*Firms with core business in R&D and marketing of pharmaceuticals:* • All companies with core business in Pharma
Products: • Generics • OTC Products • New drug applications • New Derivative • Biosimilars • Unclassified	*Products:* • New molecular entities
Indications: • Not in R&D or marketed for the US • Status unclassified • Last status before 2008 - after 2017	*Indications:* • In R&D or marketed for the US • Status Assigned • Last status between 2008 and 2017

Table C.6: Validation results of the five SL-algorithms across data sets (one repetition only)

		AUC	AUCL	AUCH	SENS	SPEC
PROBIT	PI	0.8497	0.8272	0.8722	0.9093	0.5630
	PII	0.8887	0.8657	0.9116	0.6538	0.9141
	PIII	0.8767	0.8331	0.9204	0.9441	0.5960
	Filed	0.7000	0.4971	0.9029	1.0000	0.0000
DT	PI	0.8096	0.7835	0.8358	0.9320	0.4048
	PII	0.8416	0.8125	0.8707	0.5737	0.9408
	PIII	0.8013	0.7461	0.8564	0.9255	0.5657
	Filed	0.5000	0.5000	0.5000	1.0000	0.0000
RF	PI	0.8808	0.8611	0.9005	0.9559	0.3941
	PII	0.9340	0.9165	0.9516	0.7917	0.9179
	PIII	0.9098	0.8749	0.9448	0.9317	0.6667
	Filed	**0.9781**	0.9540	1.0000	1.0000	0.0000
BART	PI	**0.9420**	0.9294	0.9547	0.9219	0.7480
	PII	**0.9618**	0.9480	0.9756	0.8429	0.9523
	PIII	**0.9401**	0.9142	0.9659	0.8882	0.8081
	Filed	0.9518	0.8964	1.0000	1.0000	0.0000
C5.0	PI	0.8686	0.8471	0.8902	0.9307	0.5416
	PII	0.9111	0.8911	0.9312	0.7115	0.9179
	PIII	0.8924	0.8539	0.9309	0.9255	0.5960
	Filed	0.6686	0.3487	0.9886	0.9927	0.0000

Table C.6 shows the AUC, its lower and upper 95% confidence interval (AUCL, AUCH), the sensitivity (SENS) and the specificity (SPEC) on the validation set for each SL-algorithm and data set. The algortihm with the highest AUC of each data set is highlighted in bold.

Figure C.1: Stylized drug development process in the US

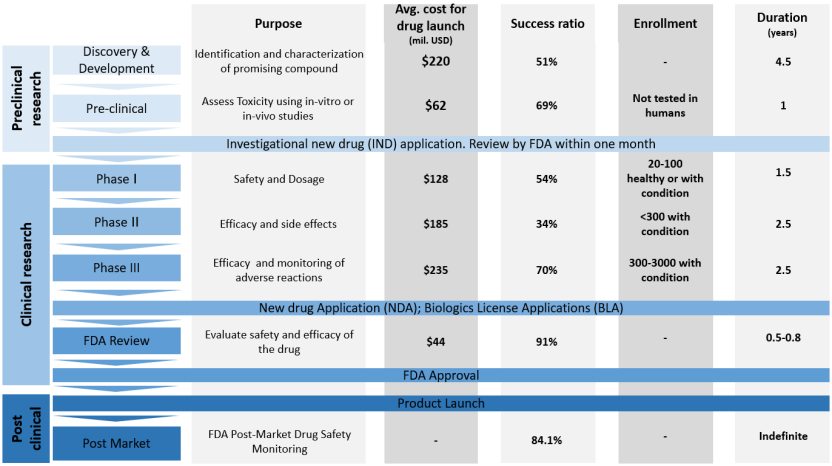


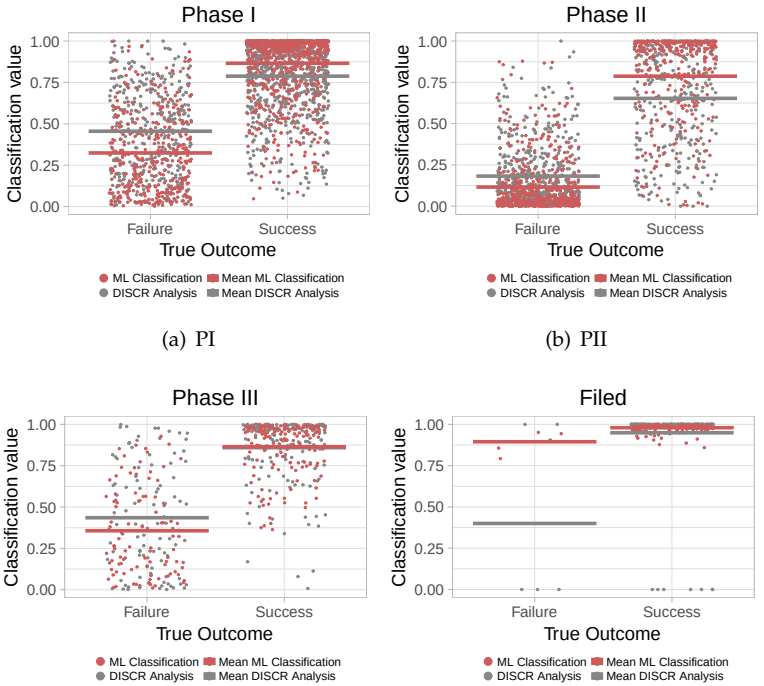
Figure C.1 stylizes the US drug development process and describes each stage according to US Food and Drug Agency, 2019. Estimates concerning the average cost for one drug launch, the success ratio, and the duration of each stage are found in Paul et al., 2010. In stage "Discovery and Development", cost, success ratio, and duration reflect estimates after selection of a target compound. The post launch success ratio is found in Qureshi et al., 2011.

Table C.7: Meta-analysis of phase success rates

Source	Number of observations	Time	Success rates in Phase I	in Phase II	in Phase III	PI-PIII Completion
Wong and Lo (2019)	15102	2000-2015	66.4%	48.6%	59.0%	19.0%
EvaluatePharma (2018)	16000	2000-2018	66.8%	33.1%	57.5%	13.7%
Thomas et al. (2016)	7455	2006-2015	63.2%	30.7%	58.1%	11.3%
Hay et al. (2014)	4451	2003-2011	64.5%	32.4%	60.1%	13.6%
Weighted average			**65.8%**	**38.1%**	**58.4%**	**14.7%**

Table C.7 shows a meta analysis of success rates in clinical phases. The weighted average weighs the success rate of each source by its share of observations.

Figure C.2: Best-in-class SL and DISCR classification values separated by phase and true outcome



(a) PI

(b) PII

(c) PIII

(d) Filed

Figure C.3: Predicted success rates of current PII and PIII pipeline by technology

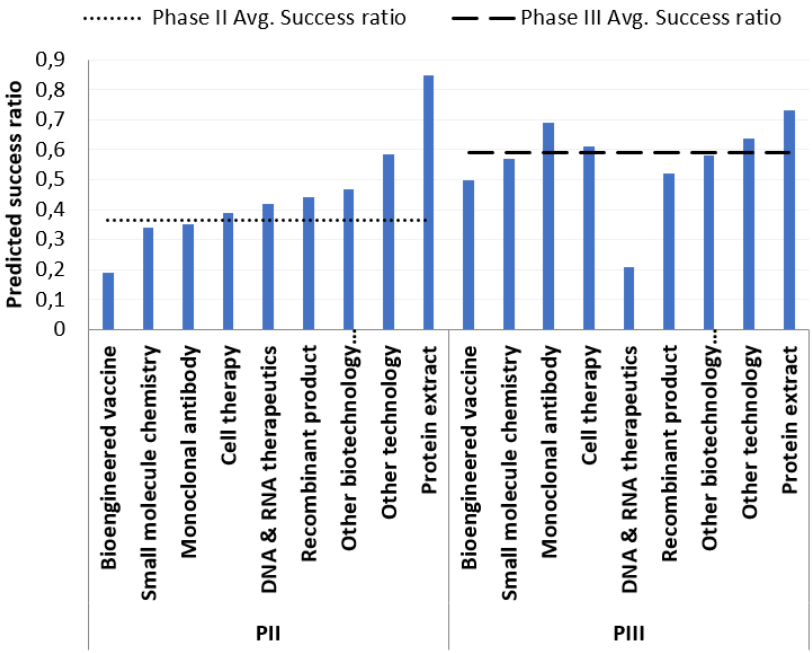


Figure C.3 contains the predicted success ratio on the current PII/ PII pipeline using the BART algorithm. Success ratios are split by employed technology. Only categories with more than 10 observations reported.

Figure C.4: Predicted success rates of current PII and PIII pipeline by indication

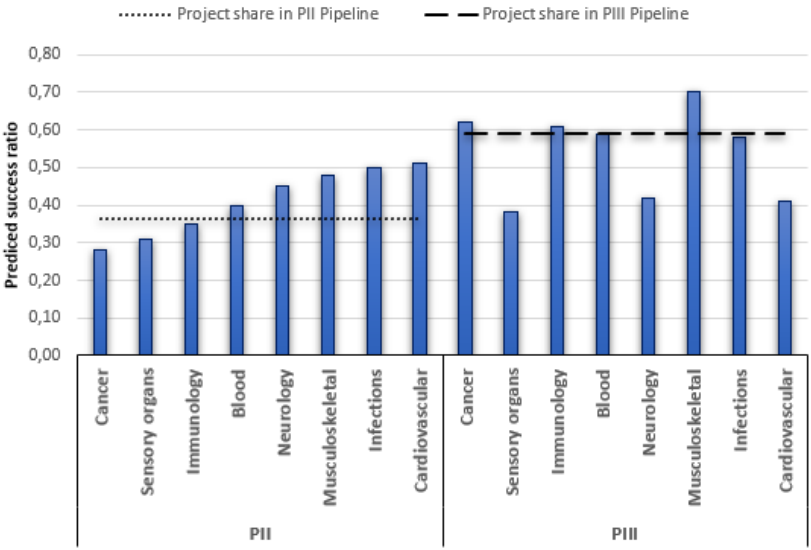


Figure C.4 contains the predicted success ratio on the current PII/ PII pipeline using the BART algorithm. Success ratios are split by indication level 1. Only categories with more than 20 observations reported.

Appendix D

Supplementary Material for Chapter 5

D.1 Product Success Probability Estimation via Supervised Learning

This section presents each method used in the SL procedure in more detail. We also report a description of the feature set (table D.1), information on the optimal weights across methods and stages (table D.2), and information on the out-of-sample performance (table D.3).

The $HIST$ measure is built on the stage-specific historical success ratio of a therapeutic market. For example, a Phase II product in the therapeutic market "breast cancer" is assigned a stage-specific success probability of 0.24 based on the ratio of products in the same stage and same market that passed Phase II. A product in the same market, but in Phase III (Filed) would be assigned the historical success probability of 0.5 (1). Then, for obtaining the success probability of receiving market approval the stage-specific probabilities are multiplied. The historic success probability of a breast cancer product in Phase II to be approved is thus $0.24 * 0.5 * 1 = 0.12$. Our data sample covers success rates of more than 200 therapeutic markets that are different across development stages. We use

the $HIST$ method as it presents an easy rule-of-thumb investors could use to form probability estimates.

The intuition of the FFT measure is best explained by an example. Figure D.1 shows the FFT tree structure of the decision rules used to estimate Phase I success probabilities. The FFT decision tree is computed during the training phase of the algorithm (see Phillips et al., 2017 for details). For each observation in the testing set, the decision tree sequentially evaluates its features (nodes - rectangular boxes) based on decision rules (in orange). At each decision rule the FFT tree either predicts the stage-specific outcome of a product or moves to the next feature and decision rule. The classifying branches of decision rules are also called leaves. For example, the algorithm is trained such that in Phase I all products whose Mechanism of Action (MoA) has been previously approved are classified as successful. If the MoA has not been approved yet, the next feature "Proprietary Level" is evaluated. At the last decision rule all remaining observations are classified. The percentage in the blue circle denotes the accuracy for each decision rule which is the number of correct classifications over all classifications. The way the FFT algorithm proceeds is similar to an investor who evaluates one feature at a time and decides based on a previously learned cutoff rule whether to form an opinion about the product outcome or to evaluate another (less relevant) feature.

$LOGIT$ models are commonly used in regression analysis. The binary variable (success or failure) at each stage of the development process is modeled by a logistic function that maps values from the real domain into a range between zero and one, which can be thought of as probabilities ($f : \mathbb{R} \to (0, 1)$).

$$f(x) = \frac{1}{1 + e^{-(\mathbf{X}'\beta)}}$$

Using a maximum likelihood procedure, the vector β is chosen such that, given the feature set $\mathbf{X}$, the mean squared error between $f(x)$ and the outcomes of the training set is minimized. We include the $LOGIT$ method since it resembles an investor who estimates the success probability of a product by assigning weights to every product feature.

Figure D.1: FFT decision rules used to estimate phase I success probabilities

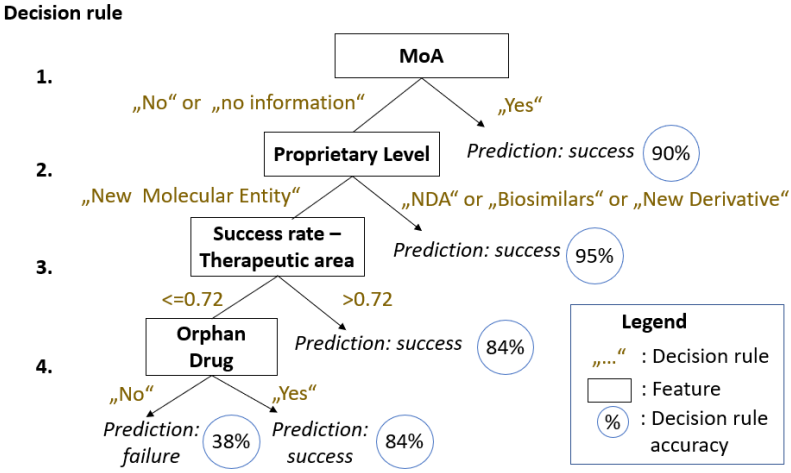


Figure D.1 shows the FFT decision tree and the accuracy at each leaf in the Phase I testing set.

Lastly we use a $BART$ algorithm, as it resembles a sophisticated supervised learning method that can consider non-linear relationships and interactions between features. As the FFT algorithm, $BART$ belongs to the family of tree algorithms, but rather than estimating just one single tree it estimates an ensemble of trees. The contribution of every tree to classify observations is additive. The structure of the trees and the leaf parameters (decision nodes) of each tree are determined by prior distributions that are data set dependent. This ensures that all trees in the ensemble contribute to the classification. The posterior distribution of the tree structure and its leaves is drawn using a Gibbs sampler. We use the R package *bartMachine* of Kapelner and Bleich, 2013 for our implementation.

Once we have trained all algorithms, we derive the 4×1 vector of optimal weights w that minimizes the mean squared error between the matrix of probability estimates of each method A at a specific stage and the vector of true outcomes y solving the following quadratic optimization problem:

$$\begin{aligned} \min_{w} \frac{1}{2}||Aw - y||^2 &= \frac{1}{2}(Aw - y)^T(Aw - y) = \\ &= \frac{1}{2}w^T A^T Aw - y^T Aw + \frac{1}{2}y^T y \\ s.t. \quad w &\geq 0, \\ \mathbf{1}^T w &= 1 \end{aligned}$$

The weights are hence chosen such that their linear combination with the probability estimates minimizes the distance to the vector of true outcomes y while restricting them to sum up to one and being positive. We reserve 15% of the observations in each sample for deriving the optimal weights between algorithms. On these 15% we use a 10-fold cross validation, meaning we optimize ten times using randomly chosen samples of 10% and average the weighting estimates. This procedure is performed for each development stage leading to four different weighting vectors. Table D.2 shows for each stage the weighting vector assigned to the success probability estimates of the different methods. Whilst for Phase I and Phase II prediction performance is optimized using a high share of $BART$, we see that the weight of FFT and $LOGIT$ increases in the late stage of development. The weight of $HIST$ is very low for probability estimates across phases. Thus the $HIST$ measure is most different with respect to the weighted average that considers all methods. Replacing the weighted average with $HIST$, as we do in the regression robustness check, generates more distortion from the original $Prob$ measure than using any other method in isolation.

Not only do the methods used in the supervised learning procedure differ in their complexity, they also perform differently when predicting outcomes on a separately kept testing sample (the remaining 15%). Table D.3 contains for each method the out-of-sample classification results using the area under the receiver operating curve (AUC) measure. The AUC measure ranges between 0 and 1 and quantifies how well the classification approach solves the trade-off between type one and type two errors. Roughly speaking, an AUC of 0.5 is equivalent to randomly guessing the outcome. We can observe that more complex methods are associated to better out-of-sample classification results. In Phase I till Phase

III $BART$ performs best, whereas in Filed FFT delivers the best prediction results. Across data sets the classification performance of $HIST$ is lowest.

Table D.1: Description of product success features used in SL-methods

Feature category	Feature Name	Description
Product	MoA	Classifies whether the underlying mechanism of action (MoA) of the product has been already approved in the US market. New products that rely on already approved MoA are usually less risky.
Product	Proprietary level	Classifies the product by its type of novelty. The following classes are considered: New Molecular Entity, New Drug Application (NDA), Biosimilar, and New Derivative.
Regulatory	Orphan Drug	Classifies whether the product has been granted an orphan drug status. Orphan status is granted to drugs that are developed for rare diseases and usually encompasses tax benefits and market exclusivity.
Regulatory	Expedited	Classifies whether the product has been granted an expedited status. Promising drug candidates can apply for expedited FDA treatment, which implies an accelerated review process.
Technology	Success rate - Technology	The historical phase-specific success rate of products that share the same type of technology e.g. "Small molecule Medicine" or "Cell therapy".
Therapeutic Market	Success rate - Therapeutic area	The historical phase-specific success rate of products that share the same therapeutic area e.g. "Alzheimer's disease" or "Narcolepsy".

Table D.2: Optimal weights (rounded) of SL methods for each phase

Method	Phase I	Phase II	Phase III	Filed
HIST	0.03	0.01	0.03	0.01
FFT	0.02	0.02	0.17	0.36
LOGIT	0.10	0.06	0.17	0.25
BART	0.86	0.91	0.62	0.37
Sum	1	1	1	1

Table D.3: Out-of-sample AUC of each SL method across development phases

Method	Phase I	Phase II	Phase III	Filed
HIST	0.58	0.68	0.65	0.53
FFT	0.68	0.82	0.68	0.68
LOGIT	0.73	0.87	0.74	0.62
BART	0.76	0.88	0.77	0.65

D.2 Additional Tables and Figures

Figure D.2: Distribution of cumulated abnormal returns split by negative and positive announcements

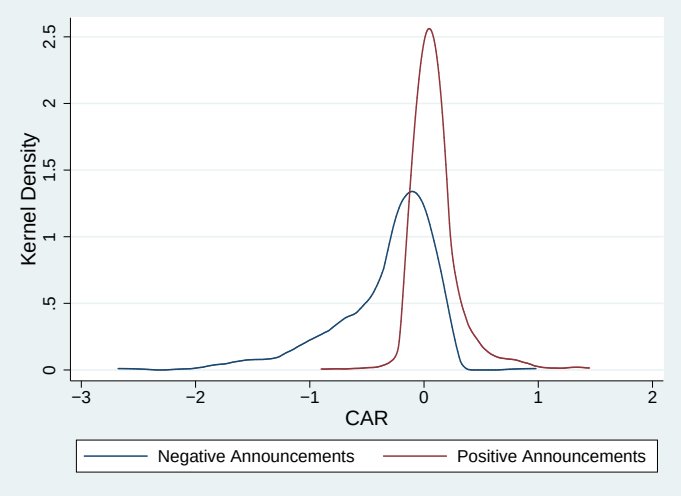


Figure D.2 shows the dispersion ofCAR for negative and positive announcements.

Table D.7: Model 1 - Effect of product portfolio importance $Port$ on $Return$ and CAR - with control variables

	Dependent variable: $Return$				Dependent variable: CAR			
	Spec. 1	Spec. 2	Spec. 3	Spec. 4	Spec. 1	Spec. 2	Spec. 3	Spec. 4
$Port$	-0.27***	-0.32***	-0.32***	-0.34***	-0.44***	-0.47***	-0.35***	-0.29***
	(0.05)	(0.06)	(0.05)	(0.03)	(0.05)	(0.10)	(0.10)	(0.04)
$Positive$	0.09***	0.10***	0.03**	0.07***	0.07*	0.06*	0.03***	0.09***
	(0.03)	(0.03)	(0.02)	(0.02)	(0.04)	(0.03)	(0.01)	(0.02)
$Port \times Positive$	0.47***	0.46***	0.41***	0.45***	0.54***	0.55***	0.42***	0.38***
	(0.05)	(0.07)	(0.06)	(0.03)	(0.06)	(0.10)	(0.10)	(0.04)
Constant	-0.06**	-0.04	0.05**	0.00	-0.04	0.01	0.03	-0.05
	(0.03)	(0.05)	(0.02)	(0.03)	(0.03)	(0.06)	(0.03)	(0.04)
Acquisition		0.04	-0.01	-0.01		0.03	0.00	0.01
		(0.04)	(0.01)	(0.01)		(0.04)	(0.01)	(0.02)
In-licensed		0.02	0.01	0.02		0.03	0.01	0.04**

Table D.7 – continued from previous page

	Dependent variable: $Return$				Dependent variable: CAR			
	Spec. 1	Spec. 2	Spec. 3	Spec. 4	Spec. 1	Spec. 2	Spec. 3	Spec. 4
		(0.02)	(0.00)	(0.01)		(0.03)	(0.01)	(0.02)
Other		0.15**	0.11	0.12		0.24	0.05	0.11
		(0.07)	(0.30)	(0.07)		(0.15)	(0.75)	(0.09)
Filed		-0.03	-0.06***	-0.04		-0.07*	-0.06**	-0.05*
		(0.03)	(0.01)	(0.02)		(0.04)	(0.03)	(0.03)
Phase II		0.06	-0.03**	0.00		0.03	-0.04	-0.01
		(0.04)	(0.01)	(0.02)		(0.04)	(0.03)	(0.03)
Phase III		-0.00	-0.05***	-0.02		-0.07*	-0.06**	-0.03
		(0.03)	(0.01)	(0.02)		(0.04)	(0.03)	(0.03)
Blood		0.02	0.02**	0.03		0.04	0.03	0.03
		(0.03)	(0.01)	(0.03)		(0.04)	(0.03)	(0.03)
Cardiovascular		-0.03	-0.01	-0.04		-0.15	-0.01	0.04
		(0.06)	(0.01)	(0.03)		(0.12)	(0.01)	(0.04)
Immunology		-0.01	0.01	0.01		0.01	0.00	0.01
		(0.03)	(0.01)	(0.03)		(0.03)	(0.01)	(0.03)
Infections		-0.03	-0.00	-0.01		0.01	0.02*	0.01
		(0.04)	(0.01)	(0.03)		(0.05)	(0.01)	(0.04)
Miscellaneous		0.01	-0.01	-0.02		0.02	-0.00	-0.01
		(0.03)	(0.01)	(0.02)		(0.03)	(0.01)	(0.02)
Musculoskeletal		-0.02	-0.01	-0.02		-0.04	0.00	-0.00
		(0.03)	(0.01)	(0.03)		(0.05)	(0.01)	(0.03)
Neurology		0.04	0.02	0.03		0.06	0.03**	0.04
		(0.04)	(0.01)	(0.02)		(0.04)	(0.01)	(0.02)
Respiratory		0.04	0.01	0.03		0.07	0.02**	0.05
		(0.04)	(0.01)	(0.03)		(0.05)	(0.01)	(0.03)
Europe		0.01	0.00	-0.00		0.02	0.00	-0.01
		(0.02)	(0.01)	(0.01)		(0.02)	(0.01)	(0.02)
Other		0.19	-0.00	-0.01		0.21	-0.01	0.00
		(0.21)	(0.01)	(0.04)		(0.17)	(0.02)	(0.05)
Global Majors		-0.06**	-0.01	-0.03		-0.03	-0.01	-0.00
		(0.03)	(0.01)	(0.02)		(0.03)	(0.01)	(0.02)
Other		0.11**	0.01	0.05		0.21***	0.03	0.07
		(0.05)	(.)	(0.12)		(0.08)	(.)	(0.15)
Regional Major		-0.06*	-0.01	-0.02		-0.06	-0.01	-0.01
		(0.03)	(0.01)	(0.03)		(0.04)	(0.01)	(0.03)
Specialty		-0.06	-0.00	-0.01		-0.08	0.00	0.01
		(0.04)	(0.01)	(0.02)		(0.05)	(0.01)	(0.02)
2017Q2		0.00	0.00	0.01		0.02	0.02**	0.03
		(0.03)	(0.01)	(0.02)		(0.03)	(0.01)	(0.03)
2017Q3		0.08*	0.00	0.02		0.07*	0.01*	0.04
		(0.05)	(0.01)	(0.02)		(0.04)	(0.01)	(0.03)
2017Q4		-0.01	-0.00	-0.01		-0.02	0.00	-0.02

Table D.7 – continued from previous page

	Dependent variable: $Return$				Dependent variable: CAR			
	Spec. 1	Spec. 2	Spec. 3	Spec. 4	Spec. 1	Spec. 2	Spec. 3	Spec. 4
		(0.03)	(0.01)	(0.02)		(0.03)	(0.01)	(0.03)
2018Q1		-0.03	0.01	-0.00		-0.06	0.01*	0.01
		(0.03)	(0.01)	(0.02)		(0.06)	(0.01)	(0.03)
2018Q2		0.06	0.00	0.03		0.06	0.01	0.03
		(0.05)	(0.01)	(0.02)		(0.04)	(0.02)	(0.03)
2018Q3		-0.02	-0.01	-0.01		0.00	0.01	0.03
		(0.03)	(0.01)	(0.02)		(0.04)	(0.01)	(0.03)
2018Q4		-0.07**	-0.02	-0.04*		-0.04	-0.00	-0.03
		(0.03)	(0.01)	(0.02)		(0.04)	(0.01)	(0.03)
Many Markets		-0.05**	-0.01	-0.04**		-0.10**	-0.01	-0.05**
		(0.02)	(0.01)	(0.02)		(0.04)	(0.01)	(0.02)
Observations	510	510	510	510	510	510	510	510
Adjusted R^2	0.359	0.398	-	0.601	0.353	0.407	-	0.475
Control variables	No	Yes	Yes	Yes	No	Yes	Yes	Yes
Outlier correction	No	No	Yes	Yes	No	No	Yes	Yes

Table D.7 contains the regression results for model 1 concerning the effect of product portfolio importance $Port$ on $Return$ (column 2-5) and on CAR (column 6-9). All control variables included. Standard errors in parenthesis. P-value $\leq 0.1 :^{*}$ $\leq 0.05 :^{**}$ $\leq 0.01 :^{***}$.

Table D.8: Model 2 - Effect of product success probability $Prob$ on $Return$ and CAR - with control variables

	Dependent variable: $Return$				Dependent variable: CAR			
	Spec. 1	Spec. 2	Spec. 3	Spec. 4	Spec. 1	Spec. 2	Spec. 3	Spec. 4
$Prob$	0.10	0.12*	0.23***	0.12**	0.06	0.07	0.22**	0.12*
	(0.07)	(0.07)	(0.08)	(0.05)	(0.07)	(0.10)	(0.11)	(0.06)
$Positive$	0.66***	0.65***	0.47***	0.53***	0.58***	0.56***	0.41***	0.52***
	(0.06)	(0.08)	(0.05)	(0.04)	(0.05)	(0.07)	(0.09)	(0.04)
$Prob \times Positive$	-0.40***	-0.35***	-0.31***	-0.27***	-0.23***	-0.21**	-0.28***	-0.24***
	(0.09)	(0.10)	(0.08)	(0.05)	(0.08)	(0.10)	(0.11)	(0.06)
Constant	-0.31***	-0.35***	-0.31***	-0.32***	-0.37***	-0.36***	-0.28***	-0.34***
	(0.05)	(0.06)	(0.05)	(0.04)	(0.04)	(0.07)	(0.09)	(0.05)
Acquisition		0.08**	0.00	0.04**		0.06*	-0.00	0.03
		(0.04)	(0.01)	(0.02)		(0.04)	(0.01)	(0.02)
In-licensed		-0.02	0.00	0.02		0.03	0.00	0.03

Table D.8 – continued from previous page

	Dependent variable: $Return$				Dependent variable: CAR			
	Spec. 1	Spec. 2	Spec. 3	Spec. 4	Spec. 1	Spec. 2	Spec. 3	Spec. 4
		(0.03)	(0.01)	(0.02)		(0.03)	(0.01)	(0.02)
Other		0.11	0.03***	0.10**		0.15*	0.05	0.11*
		(0.10)	(0.01)	(0.05)		(0.09)	(0.05)	(0.05)
Filed		-0.07	-0.04**	-0.02		-0.07	-0.04*	-0.05
		(0.06)	(0.02)	(0.04)		(0.05)	(0.02)	(0.04)
Phase II		-0.03	-0.02	0.00		-0.05	-0.01	-0.03
		(0.06)	(0.02)	(0.03)		(0.04)	(0.02)	(0.03)
Phase III		-0.07	-0.04***	-0.02		-0.09**	-0.05**	-0.05
		(0.05)	(0.01)	(0.03)		(0.04)	(0.02)	(0.04)
Blood		0.07	0.04***	0.03		0.04	0.02	0.02
		(0.06)	(0.01)	(0.04)		(0.05)	(0.02)	(0.04)
Cardiovascular		0.12	0.03	-0.01		-0.08	-0.00	-0.01
		(0.12)	(0.03)	(0.04)		(0.12)	(0.05)	(0.05)
Immunology		-0.05	-0.01	-0.04		-0.07	-0.01	-0.04
		(0.04)	(0.01)	(0.04)		(0.05)	(0.01)	(0.04)
Infections		0.02	-0.01	-0.01		-0.03	0.00	-0.03
		(0.04)	(0.02)	(0.04)		(0.06)	(0.01)	(0.04)
Miscellaneous		0.06*	0.01	0.03		0.02	-0.00	0.01
		(0.04)	(0.01)	(0.02)		(0.03)	(0.01)	(0.03)
Musculoskeletal		0.01	0.00	-0.01		-0.03	-0.00	-0.03
		(0.04)	(0.01)	(0.03)		(0.05)	(0.01)	(0.04)
Neurology		0.09**	0.04**	0.04		0.04	0.03**	0.04
		(0.04)	(0.02)	(0.03)		(0.04)	(0.01)	(0.03)
Respiratory		0.20	0.07*	0.05		0.15**	0.07*	0.10**
		(0.15)	(0.04)	(0.03)		(0.06)	(0.04)	(0.04)
Europe		0.03	0.00	0.01		0.03	-0.00	0.01
		(0.04)	(0.01)	(0.02)		(0.03)	(0.01)	(0.02)
Other		0.07	-0.03	-0.03		0.03	-0.02	0.04
		(0.12)	(0.03)	(0.05)		(0.13)	(0.02)	(0.06)
Global Majors		-0.07***	-0.04***	-0.04**		-0.01	-0.03**	-0.01
		(0.03)	(0.01)	(0.02)		(0.03)	(0.01)	(0.02)
Other		-0.07	-0.14	-0.01		0.11**	-0.16	0.07
		(0.09)	(0.17)	(0.14)		(0.05)	(0.36)	(0.17)
Regional Major		-0.09*	-0.05***	-0.04		-0.02	-0.02	-0.01
		(0.05)	(0.01)	(0.04)		(0.05)	(0.02)	(0.05)
Specialty		-0.07*	-0.02	-0.03		-0.04	-0.02	-0.03
		(0.04)	(0.02)	(0.02)		(0.04)	(0.01)	(0.03)
2017Q2		0.04	0.01	0.03		0.06	0.02	0.05
		(0.03)	(0.01)	(0.03)		(0.04)	(0.01)	(0.04)
2017Q3		0.04	0.00	0.03		0.06	0.00	0.04
		(0.04)	(0.01)	(0.03)		(0.04)	(0.01)	(0.04)
2017Q4		0.05	0.00	0.00		-0.00	0.00	0.01

Table D.8 – continued from previous page

	Dependent variable: $Return$				Dependent variable: CAR			
	Spec. 1	Spec. 2	Spec. 3	Spec. 4	Spec. 1	Spec. 2	Spec. 3	Spec. 4
		(0.04)	(0.01)	(0.03)		(0.05)	(0.01)	(0.04)
2018Q1		-0.03	0.02	-0.02		-0.06	0.00	-0.01
		(0.04)	(0.01)	(0.03)		(0.06)	(0.01)	(0.04)
2018Q2		0.02	0.01	0.00		0.03	0.00	0.03
		(0.05)	(0.01)	(0.03)		(0.05)	(0.02)	(0.04)
2018Q3		0.09*	0.01	0.03		0.09*	0.03***	0.08**
		(0.06)	(0.01)	(0.03)		(0.05)	(0.01)	(0.04)
2018Q4		0.03	0.01	-0.01		0.02	-0.00	0.01
		(0.05)	(0.01)	(0.03)		(0.04)	(0.01)	(0.04)
Many Markets		-0.04	-0.02**	-0.05***		-0.09*	-0.03**	-0.05**
		(0.04)	(0.01)	(0.02)		(0.04)	(0.01)	(0.02)
Observations	703	703	703	703	703	703	703	703
Adjusted R^2	0.268	0.282		0.453	0.314	0.336		0.374
Control variables	No	Yes	Yes	Yes	No	Yes	Yes	Yes
Outlier correction	No	No	Yes	Yes	No	No	Yes	Yes

Table D.8 contains the regression results for model 2 concerning the effect of product success probability $Prob$ on $Return$ (column 2-5) and on CAR (column 6-9). All control variables included. Standard errors in parenthesis. P-value $\leq 0.1 :^{*} \quad \leq 0.05 :^{**} \quad \leq 0.01 :^{***}$.

Table D.9: Model 3 - Effect of product success probability $Prob$ and portfolio importance $Port$ on $Return$ and CAR - with control variables

	Dependent variable: $Return$				Dependent variable: CAR			
	Spec. 1	Spec. 2	Spec. 3	Spec. 4	Spec. 1	Spec. 2	Spec. 3	Spec. 4
$Positive$	0.23***	0.23***	0.17***	0.17***	0.08	0.07	0.11***	0.18***
	(0.06)	(0.06)	(0.03)	(0.04)	(0.07)	(0.08)	(0.03)	(0.04)
$Port$	-0.26***	-0.30***	-0.25***	-0.32***	-0.45***	-0.47***	-0.32***	-0.27***
	(0.05)	(0.06)	(0.05)	(0.03)	(0.05)	(0.10)	(0.07)	(0.04)
$Port \times Positive$	0.43***	0.43***	0.34***	0.43***	0.53***	0.55***	0.38***	0.36***
	(0.05)	(0.07)	(0.05)	(0.03)	(0.06)	(0.10)	(0.07)	(0.04)
$Prob$	0.07	0.15**	0.17***	0.11***	-0.10	-0.02	0.10**	0.09*
	(0.06)	(0.07)	(0.05)	(0.04)	(0.07)	(0.10)	(0.05)	(0.05)
$Prob \times Positive$	-0.19***	-0.20**	-0.18***	-0.14***	0.00	-0.00	-0.11**	-0.13**
	(0.07)	(0.08)	(0.05)	(0.04)	(0.08)	(0.11)	(0.05)	(0.05)
Constant	-0.11**	-0.13**	-0.08**	-0.07*	0.02	0.02	-0.03	-0.11**

Table D.9 – continued from previous page

	Dependent variable: $Return$				Dependent variable: CAR			
	Spec. 1	Spec. 2	Spec. 3	Spec. 4	Spec. 1	Spec. 2	Spec. 3	Spec. 4
	(0.05)	(0.06)	(0.03)	(0.04)	(0.06)	(0.09)	(0.04)	(0.05)
Acquisition		0.05	-0.00	-0.00		0.03	-0.00	0.01
		(0.04)	(0.01)	(0.01)		(0.04)	(0.01)	(0.02)
In-licensed		0.01	0.01**	0.02		0.03	0.00	0.04**
		(0.02)	(0.00)	(0.01)		(0.03)	(0.01)	(0.02)
Other		0.12*	0.05	0.10		0.24	0.04	0.08
		(0.07)	(0.12)	(0.07)		(0.15)	(0.62)	(0.09)
Filed		-0.03	-0.06***	-0.03		-0.06	-0.07**	-0.04
		(0.04)	(0.02)	(0.03)		(0.05)	(0.03)	(0.03)
Phase II		0.07*	-0.03*	0.00		0.03	-0.04	-0.01
		(0.04)	(0.02)	(0.02)		(0.04)	(0.03)	(0.03)
Phase III		-0.01	-0.05***	-0.03		-0.06	-0.06*	-0.03
		(0.03)	(0.01)	(0.02)		(0.04)	(0.03)	(0.03)
Blood		0.02	0.02***	0.03		0.04	0.02	0.03
		(0.03)	(0.01)	(0.03)		(0.04)	(0.03)	(0.03)
Cardiovascular		-0.04	-0.01	-0.05*		-0.15	-0.01	0.04
		(0.07)	(0.01)	(0.03)		(0.12)	(0.03)	(0.04)
Immunology		-0.01	0.01	0.01		0.01	0.00	0.01
		(0.03)	(0.01)	(0.03)		(0.03)	(0.01)	(0.03)
Infections		-0.03	-0.00	-0.01		0.01	0.01	0.01
		(0.04)	(0.01)	(0.03)		(0.05)	(0.01)	(0.04)
Miscellaneous		0.01	-0.01	-0.02		0.02	-0.01	-0.01
		(0.03)	(0.00)	(0.02)		(0.03)	(0.01)	(0.02)
Musculoskeletal		-0.02	-0.01	-0.02		-0.04	-0.00	-0.01
		(0.03)	(0.01)	(0.03)		(0.05)	(0.02)	(0.03)
Neurology		0.04	0.02*	0.03		0.06	0.02**	0.03
		(0.04)	(0.01)	(0.02)		(0.05)	(0.01)	(0.02)
Respiratory		0.02	-0.00	0.03		0.07	0.02	0.04
		(0.04)	(0.01)	(0.03)		(0.05)	(0.02)	(0.03)
Europe		0.01	0.01	-0.00		0.02	0.00	-0.01
		(0.02)	(0.01)	(0.01)		(0.02)	(0.01)	(0.02)
Other		0.21	0.01	-0.00		0.21	-0.01	0.01
		(0.22)	(0.02)	(0.04)		(0.17)	(0.02)	(0.05)
Global Majors		-0.06**	-0.02*	-0.03		-0.02	-0.02	-0.00
		(0.03)	(0.01)	(0.02)		(0.03)	(0.01)	(0.02)
Other		0.04	-0.07	0.00		0.21**	-0.02	0.02
		(0.06)	(.)	(0.12)		(0.09)	(.)	(0.15)
Regional Major		-0.06**	-0.01	-0.02		-0.06	-0.01	-0.01
		(0.03)	(0.01)	(0.03)		(0.04)	(0.02)	(0.03)
Specialty		-0.07*	-0.01	-0.02		-0.07	-0.01	0.00
		(0.04)	(0.01)	(0.02)		(0.05)	(0.02)	(0.02)
2017Q2		-0.00	-0.00	0.01		0.02	0.01	0.02

Table D.9 – continued from previous page

	Dependent variable: $Return$				Dependent variable: CAR			
	Spec. 1	Spec. 2	Spec. 3	Spec. 4	Spec. 1	Spec. 2	Spec. 3	Spec. 4
		(0.03)	(0.01)	(0.02)		(0.03)	(0.01)	(0.03)
2017Q3		0.07	0.00	0.01		0.07*	0.01	0.04
		(0.04)	(0.01)	(0.02)		(0.04)	(0.01)	(0.03)
2017Q4		-0.01	-0.00	-0.01		-0.02	-0.01	-0.02
		(0.03)	(0.01)	(0.02)		(0.03)	(0.01)	(0.03)
2018Q1		-0.03	0.00	-0.00		-0.06	0.01	0.01
		(0.03)	(0.01)	(0.02)		(0.06)	(0.01)	(0.03)
2018Q2		0.06	0.01	0.02		0.06	0.02	0.03
		(0.05)	(0.01)	(0.02)		(0.04)	(0.02)	(0.03)
2018Q3		-0.03	-0.01*	-0.02		0.00	-0.00	0.02
		(0.03)	(0.01)	(0.02)		(0.04)	(0.01)	(0.03)
2018Q4		-0.07**	-0.02	-0.04*		-0.04	-0.00	-0.03
		(0.03)	(0.01)	(0.02)		(0.04)	(0.01)	(0.03)
Many Markets		-0.05**	-0.01*	-0.04**		-0.10**	-0.01	-0.05***
		(0.02)	(0.01)	(0.02)		(0.04)	(0.01)	(0.02)
Observations	510	510	510	510	510	510	510	510
Adjusted R^2	0.372	0.404		0.609	0.359	0.405		0.480
Control variables	No	Yes	Yes	Yes	No	Yes	Yes	Yes
Outlier correction	No	No	Yes	Yes	No	No	Yes	Yes

Table D.9 contains the regression results for model 3 concerning the effect of product success probability $Prob$ and product portfolio importance $Port$ on $Return$ (column 2-5) and on CAR (column 6-9). All control variables included. Standard errors in parenthesis. P-value $\leq 0.1 :^{*} \leq 0.05 :^{**} \leq 0.01 :^{***}$.

Table D.4: Average CAR for positive and negative news grouped by portfolio importance

Product Portfolio Importance		Positive	Negative
All		0.0986^{s}	-0.3358^{s}
Below median	$[0; 0.159)$	0.0285^{s}	-0.0649^{s}
Above median	$[0.159; 1]$	0.1093^{s}	-0.3682^{s}
Above median − Below median		$0.0808^{\circ\circ\circ,***}$	$-0.3033^{\circ\circ\circ,***}$
1st. quartile	$[0; 0.039)$	0.0151^{s}	-0.0322
2nd. quartile	$[0.039; 0.159)$	0.0437^{s}	-0.0836^{s}
3rd. quartile	$[0.159; 0.774)$	0.0927^{s}	-0.2748^{s}
4th. quartile	$[0.774; 1]$	0.1273	-0.4500^{s}

Table D.4 groups events by outcome and portfolio importance of their associated product and reports the average CAR for each group. One, two or three circles (stars) denote significance of the difference between two samples using the Wilcoxon rank sum test (two sample t-test) on the 10%, 5% or 1% significance level, respectively. 's' denotes that the average group CAR is significantly different from zero for both t-test and Corrados rank test at the 5% level. The test statistics are explained in Appendix D.3.

Table D.5: Average CAR for positive and negative news grouped by product success probability

Product Success Probability		Positive	Negative
All		0.0986^{s}	-0.3358^{s}
Below median	$[0; 0.774)$	0.1479^{s}	-0.3542^{s}
Above median	$[0.774; 1]$	0.0561^{s}	-0.3097^{s}
Above median − Below median		$0.0918^{\circ\circ\circ,***}$	-0.0445
1st. quartile	$[0; 0.351)$	0.1595^{s}	-0.3076^{s}
2nd. quartile	$[0.351; 0.774)$	0.1351^{s}	-0.3933^{s}
3rd. quartile	$[0.774; 0.911)$	0.0951^{s}	-0.3590^{s}
4th. quartile	$[0.911; 1]$	0.0247	-0.2190^{s}

Table D.5 groups events by outcome and success probability of their associated product and reports the average CAR for each group. One, two or three circles (stars) denote significance of the difference between two samples using the Wilcoxon rank sum test (two sample t-test) on the 10%, 5% or 1% significance level, respectively. 's' denotes that the average group CAR is significantly different from zero for both t-test and Corrados rank test at the 5% level. The test statistics are explained in Appendix D.3.

Table D.6: Model Selection - Inclusion of interaction effect between $Port$ and $Prob$ according to Bayesian information criterion

Regression model 3	Log likelihood	Degrees of freedom	BIC
No interaction; no control variables	23.43	6	-9.46
With interaction; no control variables	28.61	8	-7.35
No interaction; with control variables	56.64	34	98.69
With interaction; with control variables	61.33	36	101.77

Table D.6 shows the BIC of four different versions of model 3 including and excluding the interaction effect of $Port$ and $Prob$ both with and without including control variables.

Table D.10: Robustness check 3 - Effect of product success probability $HIST$ and product portfolio importance DI on $Return$ and CAR

	Dependent variable: $Return$				Dependent variable: CAR			
	Spec. 1	Spec. 2	Spec. 3	Spec. 4	Spec. 1	Spec. 2	Spec. 3	Spec. 4
$Positive$	0.60***	0.61***	0.43***	0.52***	0.61***	0.59***	0.39***	0.52***
	(0.05)	(0.06)	(0.05)	(0.03)	(0.04)	(0.06)	(0.06)	(0.03)
DI	0.05***	0.05***	0.05***	0.05***	0.07***	0.06***	0.06***	0.05***
	(0.01)	(0.01)	(0.02)	(0.01)	(0.01)	(0.02)	(0.02)	(0.01)
$DI \times Positive$	-0.07***	-0.07***	-0.05***	-0.07***	-0.08***	-0.08***	-0.06***	-0.07***
	(0.01)	(0.02)	(0.02)	(0.01)	(0.01)	(0.02)	(0.02)	(0.01)
$HIST$	0.06	0.11*	0.17**	0.11**	0.09	0.13	0.12	0.08
	(0.07)	(0.06)	(0.06)	(0.05)	(0.06)	(0.09)	(0.08)	(0.06)
$HIST \times Positive$	-0.24***	-0.22***	-0.21**	-0.18***	-0.20***	-0.17**	-0.16*	-0.15**
	(0.08)	(0.07)	(0.06)	(0.05)	(0.07)	(0.08)	(0.08)	(0.06)
Constant	-0.32***	-0.35***	-0.29***	-0.34***	-0.43***	-0.40***	-0.27***	-0.36***
	(0.04)	(0.06)	(0.05)	(0.04)	(0.03)	(0.06)	(0.06)	(0.05)
Observations	703	703	703	703	703	703	703	703
Adjusted R^2	0.291	0.305	-	0.505	0.360	0.377	-	0.422
Control variables	No	Yes	Yes	Yes	No	Yes	Yes	Yes
Outlier correction	No	No	Yes	Yes	No	No	Yes	Yes

Table D.10 contains robustness check 3 concerning the effect of product success probability $HIST$ and portfolio index DI on $Return$ (column 2-5) and on CAR (column 6-9). Standard errors in parenthesis. P-value $\leq 0.1 :^{*} \leq 0.05 :^{**} \leq 0.01 :^{***}$.

Table D.11: Robustness check 4 - Effect of product portfolio importance $Port$ on CAR using *Dow Jones Industrial Index* as reference market

	Dependent variable: CAR			
	Spec. 1	Spec. 2	Spec. 3	Spec. 4
$Port$	-0.44***	-0.47***	-0.35***	-0.30***
	(0.05)	(0.10)	(0.10)	(0.04)
$Positive$	0.07**	0.06*	0.03**	0.09***
	(0.04)	(0.03)	(0.01)	(0.03)
$Port \times Positive$	0.54***	0.55***	0.42***	0.39***
	(0.06)	(0.10)	(0.10)	(0.04)
Constant	-0.04	0.01	0.05	-0.05
	(0.03)	(0.06)	(0.04)	(0.04)
Observations	510	510	510	510
Adjusted R^2	0.354	0.408	-	0.475
Control variables	No	Yes	Yes	Yes
Outlier correction	No	No	Yes	Yes

Robustness check: Table D.11 contains four specifications for model 1 concerning the effect of product portfolio importance $Port$ on CAR using the Dow Jones industrial index during calculation of abnormal returns. Standard errors in parenthesis. P-value $\leq 0.1 : *$ $\leq 0.05 : **$ $\leq 0.01 :^{***}$.

Table D.12: Robustness check 5 - Effect of product success probability $Prob$ on CAR using *Dow Jones Industrial Index* as reference market

	Dependent variable: CAR			
	Spec. 1	Spec. 2	Spec. 3	Spec. 4
$Prob$	0.06	0.07	0.19*	0.12*
	(0.07)	(0.10)	(0.11)	(0.06)
$Positive$	0.58***	0.57***	0.41***	0.52***
	(0.05)	(0.07)	(0.09)	(0.04)
$Prob \times Positive$	-0.23***	-0.21**	-0.27**	-0.24***
	(0.08)	(0.10)	(0.11)	(0.06)
Constant	-0.37***	-0.36***	-0.27***	-0.34***
	(0.04)	(0.07)	(0.09)	(0.05)
Observations	703	703	703	703
Adjusted R^2	0.314	0.336		0.375
Control variables	No	Yes	Yes	Yes
Outlier correction	No	No	Yes	Yes

Robustness check: Table D.12 contains four specifications for model 2 concerning the effect of product success probability $Prob$ on CAR using the Dow Jones industrial index during calculation of abnormal returns. Standard errors in parenthesis. P-value $\leq 0.1 : *$ $\leq 0.05 : **$ $\leq 0.01 :^{***}$.

Table D.13: Robustness check 6 - Effect of product success probability $Prob$ and product portfolio importance $Port$ on CAR using *Dow Jones Industrial Index* as reference market

	Dependent variable: CAR			
	Spec. 1	Spec. 2	Spec. 3	Spec. 4
$Port$	-0.45***	-0.47***	-0.31***	-0.28***
	(0.05)	(0.10)	(0.07)	(0.04)
$Positive$	0.08	0.07	0.11***	0.18***
	(0.07)	(0.08)	(0.03)	(0.04)
$Port \times Positive$	0.53***	0.55***	0.37***	0.36***
	(0.06)	(0.10)	(0.07)	(0.04)
$Prob$	-0.10	-0.02	0.09*	0.09*
	(0.07)	(0.10)	(0.05)	(0.05)
$Prob \times Positive$	0.00	-0.00	-0.09**	-0.12**
	(0.08)	(0.11)	(0.04)	(0.05)
Constant	0.02	0.02	-0.02	-0.11**
	(0.06)	(0.09)	(0.04)	(0.05)
Observations	510	510	510	510
Adjusted R^2	0.360	0.406		0.479
Control variables	No	Yes	Yes	Yes
Outlier correction	No	No	Yes	Yes

Robustness check: Table D.13 contains four specifications for model 3 concerning the effect of product success probability $Prob$ and product portfolio importance $Port$ on CAR using the Dow Jones industrial index during calculation of abnormal returns. Standard errors in parenthesis. P-value $\leq 0.1 : *$ $\leq 0.05 : **$ $\leq 0.01 :^{***}$.

Table D.14: Robustness check 7 - Effect of product portfolio importance $Port$ on CAR using *MSCI World Index* as reference market

	Dependent variable: CAR			
	Spec. 1	Spec. 2	Spec. 3	Spec. 4
$Port$	-0.44***	-0.47***	-0.35***	-0.29***
	(0.05)	(0.10)	(0.10)	(0.04)
$Positive$	0.07**	0.06*	0.04***	0.09***
	(0.04)	(0.03)	(0.01)	(0.02)
$Port \times Positive$	0.54***	0.55***	0.41***	0.38***
	(0.06)	(0.10)	(0.10)	(0.04)
Constant	-0.04	0.01	0.05	-0.05
	(0.03)	(0.06)	(0.04)	(0.04)
Observations	510	510	510	510
Adjusted R^2	0.353	0.408		0.475
Control variables	No	Yes	Yes	Yes
Outlier correction	No	No	Yes	Yes

Robustness check: Table D.14 contains four specifications for model 1 concerning the effect of product portfolio importance $Port$ on CAR using the MSCI World index index during calculation of abnormal returns. Standard errors in parenthesis. P-value $\leq 0.1 : *$ $\leq 0.05 : **$ $\leq 0.01 :^{***}$.

Table D.15: Robustness check 8 - Effect of product success probability $Prob$ on CAR using *MSCI World Index* as reference market

	Dependent variable: CAR			
	Spec. 1	Spec. 2	Spec. 3	Spec. 4
$Prob$	0.06	0.07	0.19*	0.12*
	(0.07)	(0.10)	(0.11)	(0.06)
$Positive$	0.58***	0.57***	0.40***	0.52***
	(0.05)	(0.07)	(0.09)	(0.04)
$Prob \times Positive$	-0.23***	-0.21**	-0.26**	-0.24***
	(0.08)	(0.10)	(0.11)	(0.06)
Constant	-0.37***	-0.36***	-0.27***	-0.34***
	(0.04)	(0.07)	(0.09)	(0.05)
Observations	703	703	703	703
Adjusted R^2	0.313	0.335		0.373
Control variables	No	Yes	Yes	Yes
Outlier correction	No	No	Yes	Yes

Robustness check: Table D.15 contains four specifications for model 2 concerning the effect of product success probability $Prob$ on CAR using the MSCI World Index during calculation of abnormal returns. Standard errors in parenthesis. P-value $\leq 0.1 : *$ $\leq 0.05 : **$ $\leq 0.01 :^{***}$.

Table D.16: Robustness check 9 - Effect of product success probability $Prob$ and product portfolio importance $Port$ on CAR using *MSCI World index* as reference market

	Dependent variable: CAR			
	Spec. 1	Spec. 2	Spec. 3	Spec. 4
$Port$	-0.45***	-0.47***	-0.31***	-0.27***
	(0.05)	(0.10)	(0.07)	(0.04)
$Positive$	0.08	0.07	0.10***	0.18***
	(0.07)	(0.08)	(0.03)	(0.04)
$Port \times Positive$	0.53***	0.55***	0.37***	0.36***
	(0.06)	(0.10)	(0.08)	(0.04)
$Prob$	-0.10	-0.02	0.08*	0.09*
	(0.07)	(0.10)	(0.05)	(0.05)
$Prob \times Positive$	0.00	-0.00	-0.09*	-0.12**
	(0.08)	(0.11)	(0.05)	(0.05)
Constant	0.02	0.02	-0.01	-0.11*
	(0.06)	(0.09)	(0.05)	(0.05)
Observations	510	510	510	510
Adjusted R^2	0.360	0.405		0.479
Control variables	No	Yes	Yes	Yes
Outlier correction	No	No	Yes	Yes

Robustness check: Table D.16 contains four specifications for model 3 concerning the effect of product success probability $Prob$ and product portfolio importance $Port$ on CAR using the MSCI World index during calculation of abnormal returns. Standard errors in parenthesis. P-value $\leq 0.1 : *$ $\leq 0.05 : **$ $\leq 0.01 :^{***}$.

Table D.17: Linear prediction of CAR w.r.t *In-house*, *Late stage*, and *Positive*

Positive	In-house	Not In-house	
Early Stage	0.110	0.134	0.122
Late Stage	0.049	0.041	0.045
	0.065	0.067	

Negative	In-house	Not In-house	
Early Stage	-0.300	-0.107	-0.198
Late Stage	-0.329	-0.222	-0.273
	-0.321	-0.190	

Table D.17 contains the linear prediction of CAR with respect to the interaction of *In-house*, *Late stage*, and *Positive* derived from a linear regression model that includes $Prob$, $Port$, and all remaining control variables with standard errors clustered on company level. The difference between predictions is not significant for *In-house* and *Not In-house* given positive events. But the return difference is substantial for negative events (p-value=0.088).

D.3 Description of Parametric and Non-Parametric Tests

We employ two types of tests. The first type is used to test whether $\overline{CAR}$ for a group of N events is equal to zero ($H_0 : \overline{CAR} = 0$). We use the one-sample student t-test (t_1) derived from MacKinlay, 1997:

$$t_1 = \frac{\frac{1}{N}\sum_{i=1}^{N} CAR_i}{se(\overline{CAR})}$$

where $se(\overline{CAR})$ is the standard error of the mean cumulated abnormal returns in the group.

In addition we use the non-parametric Corrados rank test (c) that allows for missing returns (Corrado and Zivney, 1992):

$$c = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \big(rank(AR_{it})/(1 + M_i) - 0.5\big)/S(U)$$

where $rank(AR_{it})$ is the rank of the abnormal return of event i in its return series $t = -51... + 51$, M_i is the number of non-missing returns in the return window for event i and $S(U)$ is the standard deviation of missing return adjusted ranks given by:

$$S(U) = \sqrt{\frac{1}{103} \sum_{t=-51}^{+51} \Big(\frac{1}{\sqrt{N_t}} \sum_{i=1}^{N_t} \big(rank(AR_{it})/(1 + M_i) - 0.5\big)\Big)^2}$$

The second type of test is used to test whether the $\overline{CAR}$ of two groups A, B is equal ($H_0 : \overline{CAR}_A = \overline{CAR}_B$). We use Welch's t-test (t_2) as:

$$t_2 = \frac{\overline{CAR_A} - \overline{CAR_A}}{\sqrt{\frac{s_A}{N_A} + \frac{s_B}{N_B}}}$$

where s_A is the unbiased variance estimator of group A and N_A the number of events in group A. The degrees of freedom are approximated by

the Welch-Satterthwaite equation.

As non-parametric alternative, we use the Wilcoxon rank sum test. The test ranks the observations of CAR_A and CAR_B and then sums the ranks of each group R_i. The test statistic is then given by.

$$U_i = R_i - \frac{n_i(n_i + 1)}{2}$$

where n_i is the number of observations in group i. Usually significance tests are performed on the smaller value of U_i.

Bibliography

Abadie, Alberto et al. (2017). *When Should You Adjust Standard Errors for Clustering?* Tech. rep. National Bureau of Economic Research (cit. on p. 30).

Abbink, Klaus et al. (2010). "Intergroup conflict and intra-group punishment in an experimental contest game". In: *American Economic Review* 100.1, pp. 420–47 (cit. on p. 14).

Ahuja, Gautam and Elena Novelli (2017). "Activity overinvestment: The case of R&D". In: *Journal of Management* 43.8, pp. 2456–2468 (cit. on p. 36).

Allen, W David and Dorla A Evans (2005). "Bidding and overconfidence in experimental financial markets". In: *The journal of behavioral finance* 6.3, pp. 108–120 (cit. on p. 131).

Alós-Ferrer, Carlos and Alexander Ritschel (2018). "The reinforcement heuristic in normal form games". In: *Journal of Economic Behavior & Organization* 152, pp. 224–234 (cit. on pp. 10, 134).

Anderson, Lisa R and Beth A Freeborn (2010). "Varying the intensity of competition in a multiple prize rent seeking experiment". In: *Public Choice* 143.1-2, pp. 237–254 (cit. on p. 139).

Anderson, Lisa R and Sarah L Stafford (2003). "An experimental analysis of rent seeking under varying competitive conditions". In: *Public Choice* 115.1-2, pp. 199–216 (cit. on p. 20).

Apesteguia, Jose et al. (2007). "Imitation - theory and experimental evidence". In: *Journal of Economic Theory* 136.1, pp. 217–235 (cit. on p. 43).

Arora, Ashish et al. (2001). "Markets for technology and their implications for corporate strategy". In: *Industrial and corporate change* 10.2, pp. 419–451 (cit. on p. 129).

Ashburn, Ted T and Karl B Thor (2004). "Drug repositioning: identifying and developing new uses for existing drugs". In: *Nature reviews Drug discovery* 3.8, p. 673 (cit. on p. 82).
Baik, Kyung Hwan et al. (2015). "Group size and matching protocol in contests". In: *University of East Anglia CBESS Discussion Paper* 13-11R (cit. on p. 39).
Barberis, Nicholas (2013). "The psychology of tail events: Progress and challenges". In: *American Economic Review* 103.3, pp. 611–16 (cit. on p. 128).
Bardsley, Nicholas and Peter G Moffatt (2007). "The experimetrics of public goods: Inferring motivations from contributions". In: *Theory and Decision* 62.2, pp. 161–193 (cit. on p. 41).
Bell, Emma et al. (2018). *Business research methods*. Oxford university press (cit. on p. 119).
Bennett, Victor Manuel and Jason Snyder (2017). "The Empirics of Learning from Failure". In: *Strategy Science* 2.1, pp. 1–12 (cit. on p. 2).
Bigoni, Maria (2010). "What do you want to know? Information acquisition and learning in experimental Cournot games". In: *Research in Economics* 64.1, pp. 1–17 (cit. on p. 16).
Bigoni, Maria and Margherita Fort (2013). "Information and learning in oligopoly: An experiment". In: *Games and Economic Behavior* 81, pp. 192–214 (cit. on p. 15).
Black, Fischer et al. (1972). "The capital asset pricing model: Some empirical tests". In: *Studies in the theory of capital markets* 81.3, pp. 79–121 (cit. on p. 102).
Bleich, Justin et al. (2014). "Variable selection for BART: an application to gene regulation". In: *The Annals of Applied Statistics*, pp. 1750–1781 (cit. on p. 70).
Boosey, Luke et al. (2017). "Contests with group size uncertainty: Experimental evidence". In: *Games and Economic Behavior* 105, pp. 212–229 (cit. on p. 15).
Bordt, Sebastian and Helmut Farbmacher (2018). "Estimating individual heterogeneity in repeated public goods experiments". In: *mimeo* (cit. on p. 55).
Bosch-Domènech, Antoni et al. (2010). "A finite mixture analysis of beauty-contest data using generalized beta distributions". In: *Experimental economics* 13.4, pp. 461–475 (cit. on p. 41).
Bouton, Chad E et al. (2016). "Restoring cortical control of functional movement in a human with quadriplegia". In: *Nature* 533.7602, p. 247 (cit. on p. 3).

Bradley, Andrew P (1997). "The use of the area under the ROC curve in the evaluation of machine learning algorithms". In: *Pattern recognition* 30.7, pp. 1145–1159 (cit. on p. 72).
Breiman, Leo (2001). "Random forests". In: *Machine learning* 45.1, pp. 5–32 (cit. on p. 67).
Brocas, Isabelle et al. (2014). "Imperfect choice or imperfect attention? Understanding strategic thinking in private information games". In: *Review of Economic Studies* 81.3, pp. 944–970 (cit. on p. 41).
Brookins, Philip and Dmitry Ryvkin (2014). "An experimental study of bidding in contests of incomplete information". In: *Experimental Economics* 17.2, pp. 245–261 (cit. on p. 34).
Brown, George W (1951). "Iterative solution of games by fictitious play". In: *Activity analysis of production and allocation* 13.1, pp. 374–376 (cit. on p. 28).
Burbidge, Robert et al. (2001). "Drug design by machine learning: support vector machines for pharmaceutical data analysis". In: *Computers & chemistry* 26.1, pp. 5–14 (cit. on pp. 57, 58).
Buser, Thomas (2016). "The impact of losing in a competition on the willingness to seek further challenges". In: *Management Science* 62.12, pp. 3439–3449 (cit. on p. 34).
Camerer, Colin F et al. (2002). "Sophisticated experience-weighted attraction learning and strategic teaching in repeated games". In: *Journal of Economic theory* 104.1, pp. 137–188 (cit. on pp. 30, 150).
— (2004). "A cognitive hierarchy model of games". In: *The Quarterly Journal of Economics* 119.3, pp. 861–898 (cit. on pp. 40, 135).
Camerer, Colin and Teck H Ho (1999). "Experience-weighted attraction learning in normal form games". In: *Econometrica* 67.4, pp. 827–874 (cit. on pp. 5, 10, 25, 133).
Casari, Marco and Charles R Plott (2003). "Decentralized management of common property resources: experiments with a centuries-old institution". In: *Journal of Economic Behavior & Organization* 51.2, pp. 217–247 (cit. on p. 40).
Cason, Timothy N, William A Masters, et al. (2018). "Winner-take-all and proportional-prize contests: Theory and experimental results". In: *Journal of Economic Behavior and Organization* (cit. on p. 13).
Cason, Timothy N, Roman M Sheremeta, et al. (2012). "Communication and efficiency in competitive coordination games". In: *Games and Economic Behavior* 76.1, pp. 26–43 (cit. on pp. 9, 14).
Chipman, Hugh A et al. (2010). "BART: Bayesian additive regression trees". In: *The Annals of Applied Statistics* 4.1, pp. 266–298 (cit. on p. 67).

Chowdhury, Subhasish M, Anwesha Mukherjee, et al. (2017). "That's the ticket: Explicit lottery randomisation and learning in Tullock contests". In: *Available at SSRN 2937826* (cit. on p. 55).
Chowdhury, Subhasish M, Roman M Sheremeta, et al. (2014). "Overbidding and overspreading in rent-seeking experiments: Cost structure and prize allocation rules". In: *Games and Economic Behavior* 87, pp. 224–238 (cit. on pp. 11, 13, 43, 152).
Cohen, Fredric J (2005). "Macro trends in pharmaceutical innovation". In: *Nature Reviews Drug Discovery* 4.1, p. 78 (cit. on p. 105).
Conte, Anna et al. (2011). "Mixture models of choice under risk". In: *Journal of Econometrics* 162.1, pp. 79–88 (cit. on p. 41).
Corrado, Charles J and Terry L Zivney (1992). "The specification and power of the sign test in event study hypothesis tests using daily stock returns". In: *Journal of Financial and Quantitative analysis* 27.3, pp. 465–478 (cit. on p. 193).
Cosaert, Sam (2019). "What Types are There?" In: *Computational Economics* 53.2, pp. 533–554 (cit. on p. 40).
Crow, D. (2018). *Big Pharma efforts on Alzheimer's tested by Pfizer exit*. Last checked on 01 February, 2018. URL: https://www.ft.com/content/c4e1241e-f731-11e7-88f7-5465a6ce1a00 (cit. on p. 1).
Dawid, Herbert et al. (2019). "Manager remuneration, share buybacks, and firm performance". In: *Industrial and Corporate Change* 28.3, pp. 681–706 (cit. on p. 129).
De Carolis, Donna Marie et al. (2009). "Weathering the storm: the benefit of resources to high-technology ventures navigating adverse events". In: *Strategic Entrepreneurship Journal* 3.2, pp. 147–160 (cit. on p. 96).
Dechenaux, Emmanuel et al. (2015). "A survey of experimental research on contests, all-pay auctions and tournaments". In: *Experimental Economics* 18.4, pp. 609–669 (cit. on pp. 2, 9, 13, 42, 43, 51).
Dedman, Elisabeth et al. (2008). "Voluntary disclosure and its impact on share prices: Evidence from the UK biotechnology sector". In: *Journal of Accounting and Public Policy* 27.3, pp. 195–216 (cit. on pp. 94–96).
DeFusco, Richard A et al. (2015). *Quantitative investment analysis*. John Wiley & Sons (cit. on p. 97).
DeLong, Elizabeth R et al. (1988). "Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach". In: *Biometrics*, pp. 837–845 (cit. on p. 72).
Deng, Li and Xiao Li (2013). "Machine learning paradigms for speech recognition: An overview". In: *IEEE Transactions on Audio, Speech, and Language Processing* 21.5, pp. 1060–1089 (cit. on p. 3).

Denrell, Jerker (2003). "Vicarious learning, undersampling of failure, and the myths of management". In: *Organization Science* 14.3, pp. 227–243 (cit. on p. 2).
Dhaene, Geert and Koen Jochmans (2015). "Split-panel jackknife estimation of fixed-effect models". In: *The Review of Economic Studies* 82.3, pp. 991–1030 (cit. on p. 46).
Diaz-Uriarte, Ramon and Sara Alvarez de Andrés (2005). "Variable selection from random forests: application to gene expression data". In: *arXiv preprint q-bio/0503025* (cit. on p. 70).
DiMasi, JA et al. (2015). "A tool for predicting regulatory approval after phase II testing of new oncology compounds". In: *Clinical Pharmacology & Therapeutics* 98.5, pp. 506–513 (cit. on p. 57).
DiMasi, Joseph A (2001). "Risks in new drug development: approval success rates for investigational drugs". In: *Clinical Pharmacology & Therapeutics* 69.5, pp. 297–307 (cit. on pp. 56, 57).
DiMasi, Joseph A and Henry G Grabowski (2007). "The cost of biopharmaceutical R&D: is biotech different?" In: *Managerial and decision Economics* 28.4-5, pp. 469–479 (cit. on pp. 93, 105).
Dohmen, Thomas et al. (2011). "Individual risk attitudes: Measurement, determinants, and behavioral consequences". In: *Journal of the European Economic Association* 9.3, pp. 522–550 (cit. on pp. 18, 140).
Donovan, Julian (2018). *The Impact of Drug and Firm Attributes on Stock Price Movements Around FDA Drug Approval Decisions*. https://ir.lib.uwo.ca/undergradawards_2018/18/. Undergraduate Awards. 18 (cit. on p. 96).
Dosi, Giovanni et al. (2001). *Learning in evolutionary environments*. Tech. rep. LEM Working Paper Series (cit. on pp. 1, 8, 132).
Ely, Kirsten et al. (2003). "The usefulness of biotechnology firms' drug development status in the evaluation of research and development costs". In: *Journal of Accounting, Auditing & Finance* 18.1, pp. 163–196 (cit. on pp. 94–96).
Enache, Luminita et al. (2018). "The Disclosure of Good versus Bad News: Evidence from the Biotech Industry". In: *Available at SSRN 3290822* (cit. on p. 92).
Engel, Christoph (2007). "How much collusion? A meta-analysis of oligopoly experiments". In: *Journal of Competition Law & Economics* 3.4, pp. 491–549 (cit. on p. 15).
European Medicines Agency (2014). *European Medicines Agency policy on publication of cli nical data for medicinal products for human use*. "https://www.ema.europa.eu/en/documents/other/european-

medicines-agency-policy-publication-clinical-data-medicinal-products-human-use_en.pdf". Last checked on 17 April, 2019 (cit. on p. 92).
Evaluate Ltd. (2018). *EvaluatePharma Vision : Success Rates (PTRS) — New Molecular Entities (NMEs) — Overview — PTRS (Phase to Approval) & Phase Progression Probabilities*. URL: http://app.evaluate.com/ux/WebReport/TabbedSummaryPage.aspx?lType=mod%20Data&itemId=&tabId=1700&compId=0 (cit. on p. 80).
— (2019). *EvaluatePharma Vision Brochure Aug 2019*. URL: https://www.evaluate.com/sites/default/files/media/documents/EvaluatePharma%5C%20Vision%5C%20Brochure%5C%20Aug%5C%202019.pdf (cit. on pp. 109, 136).
Fallucchi, Francesco, R. Andrew Luccasen, et al. (Nov. 2018). "Identifying discrete behavioural types: a re-analysis of public goods game contributions by hierarchical clustering". In: *Journal of the Economic Science Association*. DOI: 10.1007/s40881-018-0060-7. URL: https://doi.org/10.1007/s40881-018-0060-7 (cit. on p. 55).
Fallucchi, Francesco and Elke Renner (2016). *Reputational concerns in repeated rent-seeking contests*. Tech. rep. CeDEx Discussion Paper Series (cit. on p. 39).
Fallucchi, Francesco, Elke Renner, and Martin Sefton (2013). "Information feedback and contest structure in rent-seeking games". In: *European Economic Review* 64, pp. 223–240 (cit. on pp. 11, 13, 16).
Fama, Eugene F and Kenneth R French (1993). "Common risk factors in the returns on stocks and bonds". In: *Journal of financial economics* 33.1, pp. 3–56 (cit. on p. 102).
Feijoo, Felipe et al. (2020). "Key indicators of phase transition for clinical trials through machine learning". In: *Drug Discovery Today* (cit. on pp. 57, 89, 127).
Feller, Avi et al. (2018). "Principal stratification in the Twilight Zone: Weakly separated components in finite mixture models". In: *arXiv preprint arXiv:1602.06595* (cit. on p. 42).
Filiz-Ozbay, Emel and Erkut Y Ozbay (2007). "Auctions with anticipated regret: Theory and experiment". In: *American Economic Review* 97.4, pp. 1407–1418 (cit. on p. 16).
Fischbacher, Urs (2007). "z-Tree: Zurich toolbox for ready-made economic experiments". In: *Experimental economics* 10.2, pp. 171–178 (cit. on p. 16).
Fischbacher, Urs et al. (2001). "Are people conditionally cooperative? Evidence from a public goods experiment". In: *Economics letters* 71.3, pp. 397–404 (cit. on pp. 38, 40, 55).

Fraley, Chris and Adrian E Raftery (2002). "Model-based clustering, discriminant analysis, and density estimation". In: *Journal of the American statistical Association* 97.458, pp. 611–631 (cit. on p. 41).
Friedman, Daniel et al. (2015). "From imitation to collusion: Long-run learning in a low-information environment". In: *Journal of Economic Theory* 155, pp. 185–205 (cit. on pp. 16, 17).
Friedman, Lawrence (1958). "Game-theory models in the allocation of advertising expenditures". In: *Operations research* 6.5, pp. 699–709 (cit. on pp. 9, 13).
Fudenberg, Drew et al. (1983). "Preemption, leapfrogging and competition in patent races". In: *European Economic Review* 22.1, pp. 3–31 (cit. on p. 13).
El-Gamal, Mahmoud A and David M Grether (1995). "Are people Bayesian? Uncovering behavioral strategies". In: *Journal of the American statistical Association* 90.432, pp. 1137–1145 (cit. on p. 41).
Garavaglia, Christian et al. (2013). "Technological regimes and demand structure in the evolution of the pharmaceutical industry". In: *Long Term Economic Development*. Springer, pp. 61–94 (cit. on p. 2).
Garvin, Susan and John H Kagel (1994). "Learning in common value auctions: Some initial observations". In: *Journal of Economic Behavior & Organization* 25.3, pp. 351–372 (cit. on p. 16).
Ghosh, Arpita and Patrick Hummel (2018). "Cardinal contests". In: *ACM Transactions on Economics and Computation (TEAC)* 6.2, p. 7 (cit. on p. 11).
Girotra, Karan et al. (2007). "Valuing R&D projects in a portfolio: Evidence from the pharmaceutical industry". In: *Management Science* 53.9, pp. 1452–1466 (cit. on pp. 76, 94, 96, 104).
Gnyawali, Devi R and Byung-Jin Robert Park (2011). "Co-opetition between giants: Collaboration with competitors for technological innovation". In: *Research policy* 40.5, pp. 650–663 (cit. on p. 131).
Gort, Michael and Steven Klepper (1982). "Time paths in the diffusion of product innovations". In: *The economic journal* 92.367, pp. 630–653 (cit. on p. 36).
Grabowski, Henry and John Vernon (1994). "Innovation and structural change in pharmaceuticals and biotechnology". In: *Industrial and Corporate Change* 3.2, pp. 435–449 (cit. on p. 91).
Greiner, Ben (2015). "Subject pool recruitment procedures: organizing experiments with ORSEE". In: *Journal of the Economic Science Association* 1.1, pp. 114–125 (cit. on p. 16).

Gu, Shihao et al. (2018). *Empirical asset pricing via machine learning*. Tech. rep. National Bureau of Economic Research (cit. on pp. 4, 127).
Gunnthorsdottir, Anna and Amnon Rapoport (2006). "Embedding social dilemmas in intergroup competition reduces free-riding". In: *Organizational Behavior and Human Decision Processes* 101.2, pp. 184–199 (cit. on p. 15).
Hao, Maggie et al. (2017). "Improvement in clinical trial disclosures and analysts' forecast accuracy: evidence from the pharmaceutical industry". In: *Review of Quantitative Finance and Accounting* 49.3, pp. 785–810 (cit. on pp. 94, 110).
Harris, Christopher and John Vickers (1985). "Perfect Equilibrium in a Model of a Race". In: *The Review of Economic Studies* 52.2, pp. 193–209 (cit. on p. 13).
Hay, Michael et al. (2014). "Clinical development success rates for investigational drugs". In: *Nature biotechnology* 32.1, p. 40 (cit. on pp. 57, 80).
Henderson Jr, Glenn V (1990). "Problems and solutions in conducting event studies". In: *Journal of Risk and Insurance*, pp. 282–306 (cit. on p. 118).
Henderson, Rebecca and Iain Cockburn (1994). "Measuring competence? Exploring firm effects in pharmaceutical research". In: *Strategic management journal* 15.S1, pp. 63–84 (cit. on p. 57).
Herrmann, Benedikt and Henrik Orzen (2008). "The appearance of homo rivalis: Social preferences and the nature of rent seeking". In: *CeDEx discussion paper series* (cit. on p. 39).
Ho, Teck H et al. (2008). "Individual differences in EWA learning with partial payoff information". In: *The Economic Journal* 118.525, pp. 37–59 (cit. on p. 30).
Ho, Teck-Hua et al. (2001). "Economic value of EWA lite: A functional theory of learning in games". In: (cit. on p. 28).
Hopkins, Michael M et al. (2007). "The myth of the biotech revolution: An assessment of technological, clinical and organisational change". In: *Research policy* 36.4, pp. 566–589 (cit. on p. 2).
Houser, Daniel et al. (2004). "Behavior in a dynamic decision problem: An analysis of experimental evidence using a Bayesian type classification algorithm". In: *Econometrica* 72.3, pp. 781–822 (cit. on pp. 38, 41).
Huck, Steffen, Hans-Theo Normann, and Jorg Oechssler (1999). "Learning in Cournot oligopoly–An experiment". In: *The Economic Journal* 109.454, pp. 80–95 (cit. on p. 16).

Huck, Steffen, Hans-Theo Normann, and Jörg Oechssler (2004). "Two are few and four are many: number effects in experimental oligopolies". In: *Journal of Economic Behavior & Organization* 53.4, pp. 435–446 (cit. on p. 10).

Hwang, Thomas J (2013). "Stock market returns and clinical trial results of investigational compounds: an event study analysis of large biopharmaceutical companies". In: *PloS one* 8.8, e71966 (cit. on pp. 95, 96).

Jiao, Peiran and Heinrich Nax (2017). "Adoption and Abandonment of Decision-Making Principles: Evidence from Cournot Experiments". In: (cit. on p. 16).

Jones, R and JR Wilsdon (2018). "The Biomedical Bubble: Why UK research and innovation needs a greater diversity of priorities, politics, places and people". In: (cit. on p. 2).

Joos, Philip (2003). "Discussion - The Usefulness of Biotechnology Firms' Drug Development Status in the Evaluation of Research and Development Costs". In: *Journal of Accounting, Auditing & Finance* 18.1, pp. 197–206 (cit. on p. 97).

Joppi, Roberta, Silvio Garattini, et al. (2013). "Orphan drugs, orphan diseases. The first decade of orphan drug legislation in the EU". In: *European Journal of clinical pharmacology* 69.4, pp. 1009–1024 (cit. on p. 105).

Jovanovic, Boyan (1982). "Selection and the Evolution of Industry". In: *Econometrica: Journal of the Econometric Society*, pp. 649–670 (cit. on p. 36).

Jovanovic, Boyan and Yaw Nyarko (1995). "A Bayesian learning model fitted to a variety of empirical learning curves". In: *Brookings Papers on Economic Activity. Microeconomics* 1995, pp. 247–305 (cit. on p. 36).

Kaitin, Kenneth I and Joseph A DiMasi (2011). "Pharmaceutical innovation in the 21st century: new drug approvals in the first decade, 2000–2009". In: *Clinical pharmacology & therapeutics* 89.2, pp. 183–188 (cit. on p. 61).

Kapelner, Adam and Justin Bleich (2013). "bartMachine: Machine learning with Bayesian additive regression trees". In: *arXiv preprint arXiv: 1312.2171* (cit. on pp. 6, 67, 107, 136, 172).

Kellogg, David and John M Charnes (2000). "Real-options valuation for a biotechnology company". In: *Financial analysts journal* 56.3, pp. 76–84 (cit. on pp. 94, 104, 109).

Klepper, Steven (1996). "Entry, exit, growth, and innovation over the product life cycle". In: *The American economic review*, pp. 562–583 (cit. on p. 36).

Klepper, Steven (1997). "Industry life cycles". In: *Industrial and corporate change* 6.1, pp. 145–182 (cit. on p. 36).
Konrad, Kai A (2009). *Strategy and dynamics in contests*. Oxford University Press (cit. on pp. 12, 42).
Konrad, Kai A and Harris Schlesinger (1997). "Risk Aversion in Rent-Seeking and Rent-Augmenting Games". In: *The Economic Journal* 107.445, pp. 1671–1683 (cit. on pp. 18, 139).
Kourou, Konstantina et al. (2015). "Machine learning applications in cancer prognosis and prediction". In: *Computational and structural biotechnology journal* 13, pp. 8–17 (cit. on p. 3).
Krueger, Anne O (1974). "The political economy of the rent-seeking society". In: *The American economic review* 64.3, pp. 291–303 (cit. on p. 13).
Kruse, Douglas L (1992). "Profit sharing and productivity: Microeconomic evidence from the United States". In: *The Economic Journal* 102.410, pp. 24–36 (cit. on p. 13).
Kuhn, Max and Kjell Johnson (2013). *Applied predictive modeling*. Vol. 26. Springer (cit. on p. 67).
Kurzban, Robert and Daniel Houser (2005). "Experiments investigating cooperative types in humans: A complement to evolutionary theory and simulations". In: *Proceedings of the National Academy of Sciences* 102.5, pp. 1803–1807 (cit. on p. 40).
Kyle, Margaret K (2006). "The role of firm characteristics in pharmaceutical product launches". In: *The RAND Journal of Economics* 37.3, pp. 602–618 (cit. on pp. 57, 60).
Lechner, Michael (1995). "Some specification tests for probit models estimated on panel data". In: *Journal of Business & Economic Statistics* 13.4, pp. 475–488 (cit. on p. 46).
Lehmann, Erich L and George Casella (2006). *Theory of point estimation*. Springer Science & Business Media (cit. on p. 42).
Lendrem, Dennis et al. (2015). "R&D productivity rides again?" In: *Pharmaceutical statistics* 14.1, pp. 1–3 (cit. on p. 2).
León, Ángel, Diego Piñeiro, et al. (2004). *Valuation of a biotech company: A real options approach*. CEMFI, Centro de Estudios Monetarios y Financieros (cit. on p. 109).
Levin, Andrew et al. (2002). "Unit root tests in panel data: asymptotic and finite-sample properties". In: *Journal of econometrics* 108.1, pp. 1–24 (cit. on p. 33).
Lim, Wooyoung et al. (2014). "Bounded rationality and group size in Tullock contests: Experimental evidence". In: *Journal of Economic Behavior & Organization* 99, pp. 155–167 (cit. on pp. 11, 14, 22, 55, 150–152).

Lo, Andrew W et al. (2019). "Machine learning with statistical imputation for predicting drug approvals". In: *Harvard Data Science Review* (cit. on pp. 57, 68).
Lu, Xun and Liangjun Su (2017). "Determining the number of groups in latent panel structures with an application to income and democracy". In: *Quantitative Economics* 8.3, pp. 729–760 (cit. on p. 41).
MacKinlay, A Craig (1997). "Event studies in economics and finance". In: *Journal of economic literature* 35.1, pp. 13–39 (cit. on pp. 102, 103, 193).
Mago, Shakun D, Anya C Samak, et al. (2016). "Facing your opponents: Social identification and information feedback in contests". In: *Journal of Conflict Resolution* 60.3, pp. 459–481 (cit. on pp. 13, 14, 42).
Mago, Shakun D, Roman M Sheremeta, et al. (2013). "Best-of-three contest experiments: Strategic versus psychological momentum". In: *International Journal of Industrial Organization* 31.3, pp. 287–296 (cit. on pp. 18, 139).
Makin, S. (2018). *The amyloid hypothesis on trial*. Last checked on 06 February, 2020. URL: https://www.nature.com/articles/d41586-018-05719-4 (cit. on p. 134).
Malerba, Franco and Luigi Orsenigo (2002). "Innovation and market structure in the dynamics of the pharmaceutical industry and biotechnology: towards a history-friendly model". In: *Industrial and corporate change* 11.4, pp. 667–703 (cit. on p. 2).
Malik, Anes and Urquhart Lisa (2018). "World Preview 2018, Outlook to 2024". In: Last checked on 12 March, 2019. URL: http://info.evaluategroup.com/rs/607-YGS-364/images/WP2018.pdf (cit. on p. 56).
Masiliūnas, Aidas (2019). "Learning in Contests with Payoff Risk and Foregone Payoff Information". In: *Available at SSRN 3393995* (cit. on p. 15).
Mazzucato, Mariana and Massimiliano Tancioni (2008). "Innovation and idiosyncratic risk: an industry-and firm-level analysis". In: *Industrial and Corporate Change* 17.4, pp. 779–811 (cit. on p. 90).
McKelvey, Richard D and Thomas R Palfrey (1995). "Quantal response equilibria for normal form games". In: *Games and economic behavior* 10.1, pp. 6–38 (cit. on pp. 30, 150).
McLachlan, Geoffrey and David Peel (2004). *Finite mixture models*. John Wiley & Sons (cit. on p. 42).
Mercatanti, Andrea (2013). "A Likelihood-Based Analysis for Relaxing the Exclusion Restriction in Randomized Experiments with Noncom-

pliance". In: *Australian & New Zealand Journal of Statistics* 55.2, pp. 129–153 (cit. on p. 42).
Millner, Edward L and Michael D Pratt (1989). "An experimental investigation of efficient rent-seeking". In: *Public Choice* 62.2, pp. 139–151 (cit. on p. 13).
— (1991). "Risk aversion and rent-seeking: An extension and some experimental evidence". In: *Public Choice* 69.1, pp. 81–92 (cit. on p. 139).
Moffatt, Peter G (2015). *Experimetrics: Econometrics for experimental economics*. Palgrave Macmillan (cit. on p. 33).
Moore, Thomas J et al. (2018). "Estimated costs of pivotal trials for novel therapeutic agents approved by the US Food and Drug Administration, 2015-2016". In: *JAMA internal medicine* 178.11, pp. 1451–1457 (cit. on p. 93).
Morgan, John et al. (2012). "Endogenous entry in contests". In: *Economic Theory* 51.2, p. 435 (cit. on p. 22).
Neugebauer, Tibor and Reinhard Selten (2006). "Individual behavior of first-price auctions: The importance of information feedback in computerized experimental markets". In: *Games and Economic Behavior* 54.1, pp. 183–204 (cit. on p. 16).
Nosenzo, Daniele et al. (2015). "Cooperation in small groups: The effect of group size". In: *Experimental Economics* 18.1, pp. 4–14 (cit. on p. 55).
Oechssler, Jörg et al. (2016). "From imitation to collusion: a replication". In: *Journal of the Economic Science Association* 2.1, pp. 13–21 (cit. on pp. 16, 17).
Olsen, Randall J (1978). "Note on the uniqueness of the maximum likelihood estimator for the Tobit model". In: *Econometrica: Journal of the Econometric Society*, pp. 1211–1215 (cit. on p. 45).
Pammolli, Fabio et al. (2011). "The productivity crisis in pharmaceutical R&D". In: *Nature reviews Drug discovery* 10.6, pp. 428–438 (cit. on pp. 2, 61, 117).
Panattoni, Laura E (2011). "The effect of Paragraph IV decisions and generic entry before patent expiration on brand pharmaceutical firms". In: *Journal of health economics* 30.1, pp. 126–145 (cit. on p. 96).
Pangallo, Marco, Torsten Heinrich, et al. (2019). "Best reply structure and equilibrium convergence in generic games". In: *Science Advances* 5.2 (cit. on p. 11).
Pangallo, Marco, James Sanders, et al. (2017). "A taxonomy of learning dynamics in 2 x 2 games". In: *arXiv preprint arXiv:1701.09043* (cit. on pp. 11, 30).

Parco, James E et al. (2005). "Two-stage contests with budget constraints: An experimental study". In: *Journal of Mathematical Psychology* 49.4, pp. 320–338 (cit. on pp. 11, 28).
Paul, Steven M et al. (2010). "How to improve R&D productivity: the pharmaceutical industry's grand challenge". In: *Nature reviews Drug discovery* 9.3, p. 203 (cit. on pp. 82, 110, 166).
Pérez-Rodriguez, Jorge V and Beatriz González López-Valcárcel (2012). "Does innovation in obesity drugs affect stock markets? An event study analysis". In: *Gaceta sanitaria* 26.4, pp. 352–359 (cit. on p. 96).
Peters, Simon (2000). "On the use of the RESET test in microeconometric models". In: *Applied Economics Letters* 7.6, pp. 361–365 (cit. on p. 46).
Phillips, Nathaniel D et al. (2017). "FFTrees: A toolbox to create, visualize, and evaluate fast-and-frugal decision trees". In: *Judgment and Decision Making* 12.4, pp. 344–368 (cit. on pp. 86, 106, 171).
Polonio, Luca et al. (2015). "Strategic sophistication and attention in games: An eye-tracking study". In: *Games and Economic Behavior* 94, pp. 80–96 (cit. on p. 41).
Potters, Jan et al. (1998). "An experimental examination of rational rent-seeking". In: *European Journal of Political Economy* 14.4, pp. 783–800 (cit. on pp. 38, 46, 50, 52).
Price, Curtis R and Roman M Sheremeta (2011). "Endowment effects in contests". In: *Economics Letters* 111.3, pp. 217–219 (cit. on pp. 14, 43).
— (2015). "Endowment origin, demographic effects, and individual preferences in contests". In: *Journal of Economics & Management Strategy* 24.3, pp. 597–619 (cit. on p. 34).
Qureshi, Zaina P et al. (2011). "Market withdrawal of new molecular entities approved in the United States from 1980 to 2009". In: *Pharmacoepidemiology and drug safety* 20.7, pp. 772–777 (cit. on p. 166).
Ramalho, Esmeralda A and Joaquim JS Ramalho (2012). "Alternative versions of the RESET test for binary response index models: a comparative study". In: *Oxford bulletin of economics and statistics* 74.1, pp. 107–130 (cit. on p. 46).
Ramsey, James Bernard (1969). "Tests for specification errors in classical linear least-squares regression analysis". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 31.2, pp. 350–371 (cit. on p. 46).
Rassenti, Stephen et al. (2000). "Adaptation and convergence of behavior in repeated experimental Cournot games". In: *Journal of Economic Behavior & Organization* 41.2, pp. 117–146 (cit. on p. 25).
Regnstrom, Jan et al. (2010). "Factors associated with success of market authorisation applications for pharmaceutical drugs submitted to the

European Medicines Agency". In: *European journal of clinical pharmacology* 66.1, p. 39 (cit. on pp. 57, 61, 105).
Rockenbach, Bettina and Marcin Waligora (2016). "Desire to Win Drives Overbidding in Tullock Contests". In: *Available at SSRN 2727153* (cit. on p. 24).
Roth, Alvin E and Ido Erev (1995). "Learning in extensive-form games: Experimental data and simple dynamic models in the intermediate term". In: *Games and economic behavior* 8.1, pp. 164–212 (cit. on pp. 5, 10, 15, 28).
Rothenstein, Jeffrey M et al. (2011). "Company stock prices before and after public announcements related to oncology drugs". In: *Journal of the National Cancer Institute* 103.20, pp. 1507–1512 (cit. on pp. 95, 96, 102).
Rouet-Leduc, Bertrand et al. (2017). "Machine learning predicts laboratory earthquakes". In: *Geophysical Research Letters* 44.18, pp. 9276–9282 (cit. on p. 3).
Rubinstein, Ariel (2016). "A typology of players: Between instinctive and contemplative". In: *The Quarterly Journal of Economics* 131.2, pp. 859–890 (cit. on p. 55).
Sammut, Claude and Geoffrey I Webb (2017). *Encyclopedia of machine learning and data mining*. Springer Publishing Company, Incorporated (cit. on p. 59).
Sarkar, Salil K and Pieter J de Jong (2006). "Market response to FDA announcements". In: *The Quarterly Review of Economics and Finance* 46.4, pp. 586–597 (cit. on pp. 95, 96, 99).
Savikhin, Anya C and Roman M Sheremeta (2013). "Simultaneous decision-making in competitive and cooperative environments". In: *Economic Inquiry* 51.2, pp. 1311–1323 (cit. on p. 42).
Schmidt, David et al. (2001). "Dilemma games: game parameters and matching protocols". In: *Journal of Economic Behavior & Organization* 46.4, pp. 357–377 (cit. on p. 39).
Schmitt, Pamela et al. (2004). "Multi-period rent-seeking contests with carryover: Theory and experimental evidence". In: *Economics of Governance* 5.3, pp. 187–211 (cit. on p. 14).
Schram, Arthur and Joep Sonnemans (1996). "Voter turnout as a participation game: An experimental investigation". In: *International Journal of Game Theory* 25.3, pp. 385–406 (cit. on p. 13).
Selten, Reinhard (1965). "Die Strategiemethode zur Erforschung des eingeschränkt rationalen Verhaltens im Rahmen eines Oligopolexperi-

mentes". In: Seminar für Mathemat. Wirtschaftsforschung u. Ökonometrie (cit. on p. 38).
Selten, Reinhard and Rolf Stoecker (1986). "End behavior in sequences of finite Prisoner's Dilemma supergames A learning theory approach". In: *Journal of Economic Behavior & Organization* 7.1, pp. 47–70 (cit. on p. 15).
Seoane-Vazquez, Enrique et al. (2008). "Incentives for orphan drug research and development in the United States". In: *Orphanet journal of rare diseases* 3.1, p. 33 (cit. on p. 134).
Shachat, Jason and Lijia Wei (2012). "Procuring commodities: first-price sealed-bid or English auctions?" In: *Marketing Science* 31.2, pp. 317–333 (cit. on p. 41).
Sharma, Anurag and Nelson Lacey (2004). "Linking product development outcomes to market valuation of the firm: The case of the US pharmaceutical industry". In: *Journal of Product Innovation Management* 21.5, pp. 297–308 (cit. on pp. 95, 96, 131).
Sheremeta, Roman M (2010). "Experimental comparison of multi-stage and one-stage contests". In: *Games and Economic Behavior* 68.2, pp. 731–747 (cit. on pp. 14, 42).
— (2011). "Contest design: An experimental investigation". In: *Economic Inquiry* 49.2, pp. 573–590 (cit. on pp. 14, 139).
— (2013). "Overbidding and heterogeneous behavior in contest experiments". In: *Journal of Economic Surveys* 27.3, pp. 491–514 (cit. on p. 13).
— (2018). "Impulsive behavior in competition: Testing theories of overbidding in rent-seeking contests". In: *Available at SSRN 2676419* (cit. on pp. 9, 55).
Sheremeta, Roman M and Jingjing Zhang (2010). "Can groups solve the problem of over-bidding in contests?" In: *Social Choice and Welfare* 35.2, pp. 175–197 (cit. on pp. 13, 14, 42).
Shortridge, Rebecca Toppe (2004). "Market valuation of successful versus non-successful R&D efforts in the pharmaceutical industry". In: *Journal of Business Finance & Accounting* 31.9-10, pp. 1301–1325 (cit. on p. 94).
Shupp, Robert et al. (2013). "Resource allocation contests: Experimental evidence". In: *Journal of Economic Psychology* 39, pp. 257–267 (cit. on pp. 14, 18, 139).
Skaperdas, Stergios and Li Gan (1995). "Risk aversion in contests". In: *The Economic Journal*, pp. 951–962 (cit. on p. 18).

Snyder, James M (1989). "Election goals and the allocation of campaign resources". In: *Econometrica: Journal of the Econometric Society*, pp. 637–660 (cit. on p. 13).
Spiliopoulos, Leonidas et al. (2018). "Complexity, attention, and choice in games under time constraints: A process analysis." In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 44.10, p. 1609 (cit. on pp. 41, 42).
Stahl II, Dale O and Paul W Wilson (1994). "Experimental evidence on players' models of other players". In: *Journal of economic behavior & organization* 25.3, pp. 309–327 (cit. on pp. 40, 135).
Stahl, Dale O (1993). "Evolution of smartn players". In: *Games and Economic Behavior* 5.4, pp. 604–617 (cit. on pp. 40, 135).
Su, Liangjun et al. (2016). "Identifying latent structures in panel data". In: *Econometrica* 84.6, pp. 2215–2264 (cit. on pp. 5, 39, 41, 46, 134, 135).
Sutton, John (1998). *Technology and Market Structure - Theory and History*. MIT Press, pp. 10, 71, 170 (cit. on pp. 2, 37, 134).
Szymanski, Stefan (2003). "The assessment: the economics of sport". In: *Oxford Review of Economic Policy* 19.4, pp. 467–477 (cit. on p. 13).
Tanzi, Rudolph E and Ann B Parson (2001). *Decoding darkness: the search for the genetic causes of Alzheimer's disease*. Merloyd Lawrence Books (cit. on p. 1).
Thomas, David W et al. (2016). *Clinical development success rates 2006 2015*. Tech. rep. San Diego: Biomedtracker/Washington, DC: BIO/Bend: Amplion (cit. on p. 80).
Tibshirani, Robert (1996). "Regression shrinkage and selection via the lasso". In: *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288 (cit. on p. 70).
Topol, Eric J (2019). "High-performance medicine: the convergence of human and artificial intelligence." In: *Nature medicine* 25.1, pp. 44–56. ISSN: 1546-170X. DOI: 10.1038/s41591-018-0300-7. URL: http://www.ncbi.nlm.nih.gov/pubmed/30617339 (cit. on pp. 57, 58).
Treich, Nicolas (2010). "Risk-aversion and prudence in rent-seeking games". In: *Public Choice* 145.3-4, pp. 339–349 (cit. on p. 139).
Tullock, G (1980). *Efficient rent seeking*. College Station: Texas A&M University Press, pp. 97–112 (cit. on pp. 2, 8, 12, 13).
Tullock, Gordon (1967). "The welfare costs of tariffs, monopolies, and theft". In: *Economic Inquiry* 5.3, pp. 224–232 (cit. on p. 38).
Urbig, Diemo et al. (2013). "Investor reactions to new product development failures: The moderating role of product development stage".

In: *Journal of Management* 39.4, pp. 985–1015 (cit. on pp. 4, 95, 96, 99, 117).

US Food and Drug Agency (2007). *Section 801 of the Food and Drug Administration Amendments Act of 2007*. Last checked on 17 April, 2019. URL: https://www.govinfo.gov/content/pkg/PLAW-110publ85/pdf/PLAW-110publ85.pdf%5C#page=82 (cit. on p. 92).

— (2019). *The Drug Development Process*. "https://www.fda.gov/ForPatients/Approvals/Drugs/default.htm". Last checked on 17 April, 2019. URL: https://www.fda.gov/ForPatients/Approvals/Drugs/default.htm (cit. on p. 166).

Vega-Redondo, Fernando (1997). "The evolution of Walrasian behavior". In: *Econometrica: Journal of the Econometric Society*, pp. 375–384 (cit. on p. 16).

Wang, Wuyi et al. (2019). "The heterogeneous effects of the minimum wage on employment across states". In: *Economics Letters* 174, pp. 179–185 (cit. on p. 41).

Wong, Chi Heem et al. (2018). "Corrigendum: Estimation of clinical trial success rates and related parameters". In: *Biostatistics* 00, pp. 1–14. ISSN: 1465-4644. DOI: 10.1093/biostatistics/kxy072. URL: https://academic.oup.com/biostatistics/advance-article-abstract/doi/10.1093/biostatistics/kxx069/4817524 (cit. on pp. 57, 80, 93, 105).

— (2019). "Estimation of clinical trial success rates and related parameters". In: *Biostatistics* 20.2, pp. 273–286 (cit. on p. 57).

Xu, Bixia (2006). "R&D strategy and stock price volatility in the biotechnology industry". In: *Review of Accounting and Finance* 5.1, pp. 59–71 (cit. on p. 96).

Yang, Ming-Hsuan et al. (2002). "Detecting faces in images: A survey". In: *IEEE Transactions on pattern analysis and machine intelligence* 24.1, pp. 34–58 (cit. on p. 3).

Zarin, Deborah A. et al. (2016). "Trial Reporting in ClinicalTrials.gov — The Final Rule". In: *New England Journal of Medicine* 375.20, pp. 1998–2004. ISSN: 0028-4793. DOI: 10.1056/nejmsr1611785 (cit. on pp. 61, 82, 92).

Zhang, Lu et al. (2017). "From machine learning to deep learning: progress in machine intelligence for rational drug discovery". In: *Drug discovery today* 22.11, pp. 1680–1685 (cit. on pp. 3, 57, 58).

Ziad Obermeyer, EJE (2016). "Predicting the Future—Big Data". In: *Machine Learning, and Clinical Medicine.. N Engl J Med* 375.13, pp. 1216–1219 (cit. on p. 3).