博士論文

An Assessment Method of Academic Researchers' "Startup Readiness": Case Studies in the Biopharmaceutical Domain

（大学研究者の起業態勢の評価手法に関する研究
～バイオ医薬分野をケーススタディとして～）

郷治 友孝

Abstract

Given that attention in academic startups has become unprecedentedly eminent since the 2010s across countries/regions, stakeholders such as scientists, venture capitalists, business managers, university administrators and policymakers, are increasingly interested in opportunities of academic startups, and key success factors and determinants of their creation and success. To address these concerns, earlier studies collected past data of scientists in specified academic organizations and regions by using conventional methods such as personal interviews, reading published papers, and conducting field projects. This thesis, however, aims to propose an assessment method of startup readiness of academic researchers in the biopharmaceutical domain where abundant startups with intense scientific linkage have attracted venture capital financing and entrepreneurship for further R&D opportunities and commercialization, based on digital data sources that are publicly available or purchasable, independent of conventionally customized surveys and acquainted sources. This dissertation defines ***startup readiness*** as the concept describing the state when one is prepared to initiate startups and willing to do so with a hope to be successful. It is hypothesized that, long before their technology readiness matures, research topics and researchers in the biopharmaceutical domain, together with startup readiness, can be regarded as investment opportunities for venture capitalists and as career opportunities for managerial talent, which could produce academic startups by leveraging their scientific strengths.

Although this thesis follows the basic view of resource-based theory proposed by earlier studies, it presents a method using logistic regression modeling to assess and detect startup readiness of such academic researchers, at an earlier stage, in a more timely manner, in greater detail, on a larger scale, and in a scalable manner. This method first sorts specific industry segments by the financing activities that are active, and the related growing research topics that attract increased academic and industrial attention. The assessment model then attempts to compute relevant researchers' startup readiness in particular in terms of startup participation and exit such as IPO (initial public offering) and M&A (merger and acquisition), using data sources that are both real-time and computable, regarding startup finances, research papers, patents, academic organizations, and national socioeconomics. The model suggests explanatory variables to work on in order to improve or influence their startup readiness. The implication of this model is that it can help enable formulation and development of promising academic startups in the biopharmaceutical domain, in that it allows researchers to focus on enhancing their scientific prominence and innovation capability, letting business stakeholders exercise

their expertise such as financing, management and business, in a mutually complementary fashion.

Based on earlier literature studies and this thesis's focus on the biopharmaceutical domain that is a very intense science-based technology commercialization field, it is assumed that features of papers and patents, which show academic researchers' scientific prominence and innovation mindset, are more pivotal than that of other assets. Furthermore, authors' network centralities are explored as researchers' potential features of papers, to identify their emerging promising studies. Hot Topic Features are also developed, as features to measure how emerging their field/topic is. Then, the conceptual framework was built as a testable, practical model to assess, detect and explain startup readiness of academic researchers in the biopharmaceutical domain, using Individual Factors (composed of Paper-related Features and Patent-related Features), Hot Topic Features, and Ecosystem Factors (composed of Academic Organization-related Features and Nation-related Features). In order to build the proposed logistic regression assessment model, this thesis conducted stepwise selections of these factors/features to lower prediction error, created their Multivariable Fractional Polynomials (MFPs) when they are not linear with the logit of their target variables, created their Interaction Terms Factors to consider their externalities and spillovers, and addressed their multicollinearity to mitigate redundancy among them.

By implementing the logistic regression assessment model, it is found that the assessment model yields good assessment performance overall, and shows higher performance when the model assesses startup readiness regarding Exit compared to that of Participant. It is also suggested that, in the biopharmaceutical domain, the model shows excellent performance when assessing a genuinely scientific concept with high keyword growth compared to when assessing a topic that is already a technology tool in practical application, while being able to indicate good performance by combining a certain range of highest growth research topics. While Paper- and Patent-related Features belonging to Individual Factors are remarkably different between Participant/Exit researchers and non-Participant/non-Exit researchers, it is observed that Paper-related Features play the most pivotal role to assess startup readiness, especially when assessing academic researchers' potential of Exit both in a genuinely scientific topic and in a certain range of most growing topics. This thesis also shows that the proposed model is useful to identify the key factors/features that are important explanatory variables of startup readiness.

The logistic regression model based on this dissertation's selected and constructed explanatory variables will enable a wide range of stakeholders, including venture capitalists, managerial entrepreneurs, policymakers, university administrators, and

academic researchers themselves, to detect potential scientific founders to work with, to identify the determinants to implement policies with, and to recognize the variables to work on to improve their startup readiness. The data analysis method of this thesis can be implemented even by stakeholders with little or no expertise related to specific disciplines, industries and regions, since this method is structured in a way that does not need such expertise and uses accessible, real-time digital data that is purchasable or publicly available for anyone, which proposes an assessment method of startup readiness that is scalable without limitations of earlier research approaches.

Acknowledgements

Firstly, I would like to express my sincerest gratitude to my supervisor Prof. Ichiro Sakata for his continuous support to my research ever since I started to aspire to conduct data-scientific analysis on scientific founders of academic startups back in 2015. Without him, I would have never had the tenacity to tackle research questions as meaningful as the ones that lie at the core of my thesis. Equally importantly, I extend my deepest gratitude to the Board of Directors of The University of Tokyo Edge Capital Co., Ltd. (UTEC), who made it possible for me to pursue my Ph.D. research while holding the mantle of UTEC's presidency in parallel. Without the approval of the Board and the fervent support of its former Chairman Nobuya Minami, such research conducted by an incumbent venture capital managing director for practical application would never have been realized.

I would like to extend my thanks to Prof. Junichiro Mori, Dr. Kimitaka Asatani, Dr. Masanao Ochi, Kazuya Tanaka, Bohua Shao, Yoshiro Kondo and everyone at the Sakata & Mori Laboratory in the Department of Technology Management for Innovation (TMI) of The University of Tokyo's School of Engineering. Without their instructional, insightful advice and inputs, I could not have finished my thesis. I would also like to thank my co-authors at TMI: Takanari Matsuda, Yuki Hayashi and Hiroko Yamano, who guided me through an ocean of scientists' data, with inspiring inputs and valuable suggestions. I feel extremely lucky to have Yuki Hayashi and Kiran Mysore by my side, with whom I cultivated relationship during their master's programs at TMI, leading up to both of them joining UTEC, eventually. Their help and cooperation were essential to arrange the datasets and improve the composition of my thesis.

Co-workers at UTEC have supported me beyond reason. Especially, Dr. Atsushi Usami helped open my eyes to the biopharmaceutical domain as a scientific category to explore, since he conducted due diligence with me regarding a genome editing startup in 2015. I also would like to express my special thanks to Ms. Ayano Iijima, who has offered me invaluable help for every kind of logistics, to juggle UTEC, Japan Venture Capital Association and my research on everyday problems.

Insights from scholars around the world also refined my research. I would like to especially thank the following couple of researchers for their discussions with me in international conferences such as PICMET: Prof. Elicia Maine from Simon Fraser University, Canada, who suggested active roles of Principal Investigators (corresponding authors in most cases) in formulating startups, and Prof. Henry Chesbrough from UC Berkeley, U.S., who suggested impact of academic organizations, nations and patent-paper pairs on academic startups.

Although unusual in acknowledgements, I also would like to introduce Tadataka Ino (1745-1818) as a person I appreciate for inspiring my research, a historical figure known for completing the first map of Japan using the latest scientific techniques at the time that he studied in his middle age. His statue in my neighborhood has encouraged me to map out scientists.

Last but definitely not the least, I would like to thank my family and parents for all the support and understanding throughout my studies. Unconditional and never-ending tolerance and patience of my wife Motoko, in particular, allowed and enabled me to initiate, continue and eventually complete this doctoral research.

In Tokyo, Japan, November 29, 2019

Tomotaka Goji

Contents

Figures

Tables

Chapter 1. Introduction

Academic startups are increasingly attracting attention as key to transferring knowledge from academia to society, to develop innovative products and services thereby helping to create and develop new industries. Such attention has become especially eminent in the 2010s and citations regarding academic startups as research paper topics and titles are ever-increasing (Figure 1-1). This phenomenon has become prevalent across various countries and disciplines, regardless of a person being in a developed or developing country, or belonging to the field of social or natural sciences (Table 1.1 and Table 1.2). Therefore, among the stakeholders in this field, whether one is a scientist, investor, business person, university administrator or policymaker, his or her primary concern is to find opportunities of academic startups, and recognize key success factors of their creation and success while effectively controlling those determinants to achieve the goal no matter what. Scientists might want to promote public support to their research by showing their startups' commercial usefulness; investors might want to gain profitable returns on their investment in academic startups; businessmen might want to be entrepreneurs commercializing innovative research outcomes through academic startups; university administrators might want to create as many successful startups as possible out of their universities to secure external funding through licensing the intellectual property and exercising stock options in those startups; or policymakers might want to vitalize their countries or regions by creating and/or developing disruptive industries emanating from academic startups.

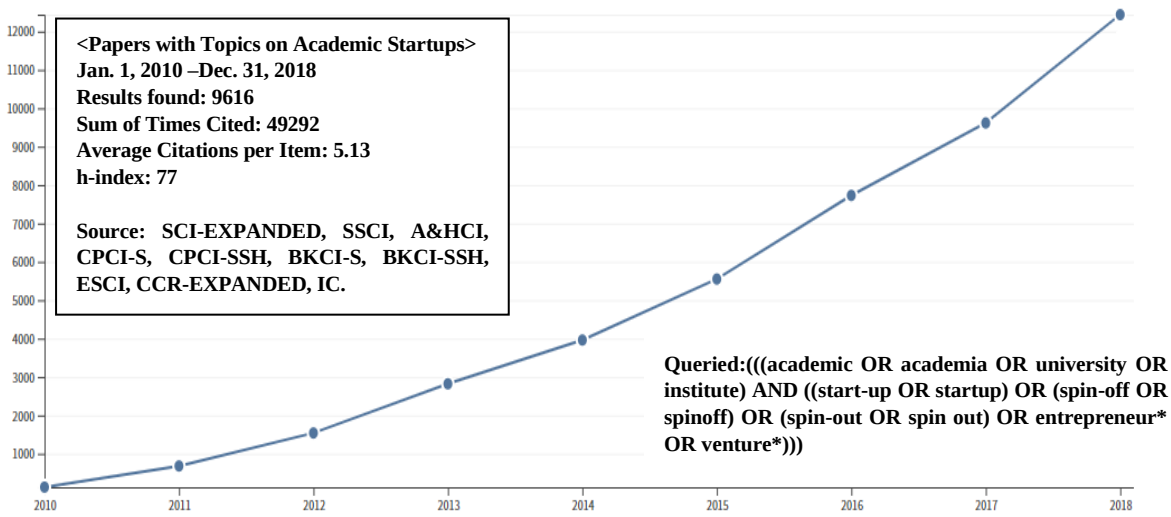


Figure 1-1 Citations on Academic Startup in Each Year (Web of Science (WoS) Core Collection)

Table 1.1 Top 25 Counties/Regions That Academic Startup Papers Belong to (WoS Core Collection)

Rank	Country/Region	Freq.	%/(All=9616)
1	USA	1878	19.53
2	PEOPLES R CHINA	1150	11.96
3	ENGLAND	780	8.11
4	SPAIN	721	7.50
5	ITALY	462	4.80
6	GERMANY	460	4.78
7	CANADA	283	2.94
8	RUSSIA	256	2.66
9	AUSTRALIA	251	2.61
10	FRANCE	250	2.60
11	SWEDEN	240	2.50
12	ROMANIA	237	2.47
13	NETHERLANDS	234	2.43
14	MALAYSIA	233	2.42
15	INDIA	199	2.07
16	PORTUGAL	197	2.05
17	BRAZIL	196	2.04
18	FINLAND	186	1.93
19	POLAND	168	1.75
20	BELGIUM	152	1.58
21	JAPAN	137	1.43
22	TAIWAN	135	1.40
23	SOUTH AFRICA	134	1.39
24	CZECH REPUBLIC	132	1.37
25	INDONESIA	132	1.37

Table 1.2 Top 25 Categories That Academic Startups Papers Belong to (WoS Core Collection)

Rank	Category	Freq.	%/(All=9616)
1	MANAGEMENT	2448	25.46
2	BUSINESS	2133	22.18
3	EDUCATION EDUCATIONAL RESEARCH	2049	21.31
4	ECONOMICS	930	9.67
5	SOCIAL SCIENCES INTERDISCIPLINARY	648	6.74
6	EDUCATION SCIENTIFIC DISCIPLINES	378	3.93
7	ENGINEERING INDUSTRIAL	371	3.86
8	ENGINEERING MULTIDISCIPLINARY	322	3.35
9	ENGINEERING ELECTRICAL ELECTRONIC	316	3.29
10	REGIONAL URBAN PLANNING	285	2.96
11	PHYSICS APPLIED	235	2.44
12	ENVIRONMENTAL STUDIES	226	2.35
13	OPERATIONS RESEARCH MANAGEMENT SCIENCE	217	2.26
14	COMPUTER SCIENCE INTERDISCIPLINARY APPLICATIONS	201	2.09
15	INFORMATION SCIENCE LIBRARY SCIENCE	201	2.09
16	BUSINESS FINANCE	186	1.93
17	COMPUTER SCIENCE THEORY METHODS	183	1.90
18	MULTIDISCIPLINARY SCIENCES	167	1.74
19	COMPUTER SCIENCE INFORMATION SYSTEMS	154	1.60
20	GEOGRAPHY	142	1.48
21	HUMANITIES MULTIDISCIPLINARY	127	1.32
22	GREEN SUSTAINABLE SCIENCE TECHNOLOGY	125	1.30
23	ENVIRONMENTAL SCIENCES	110	1.14
24	PUBLIC ADMINISTRATION	107	1.11
25	URBAN STUDIES	102	1.06

Conventionally, the most common approach to look for information regarding opportunities and determinants of academic startups' creation and success has been personal interviews, reading published papers, or conducting field projects in person.

However, none of these methods are instantaneous, comprehensive, and scalable. They also miss data, and the information surveyed by them could be skewed or obsolete, due to their fragmented, reactive and time-consuming way of survey. Since the attention and penetration of academic startups is rapidly rising worldwide in various areas, a new method to solve those difficulties is needed.

This thesis, as case studies in the biopharmaceutical domain, aims to propose an assessment method of academic researchers' promise regarding participation in startups and its financial exit such as IPO (initial public offering) and M&A (merger and acquisition). The proposed method of this thesis is useful in a real-time manner, covers greater detail, works on a larger scale, and is applicable at an earlier stage of the project.

1.1. Motivation and Scope

Conventionally, research examining factors contributing to the creation of scientific research-based startups and academic entrepreneurship has specifically examined factors other than scientific research itself. One example of a generally accepted notion about commercialization from advanced technology research is the technology readiness levels (TRL) concept. In the mid-1970s, NASA introduced TRL: a criterion to evaluate the maturity of technologies derived from science. It has been used to explain why some new technologies engender industrial transitions. However, it has been used from the perspective of project management ranging from performance to scheduling to budget, without addressing the emergence or profile of the research itself [1]. Essentially, TRL has been expected to facilitate transdisciplinary expertise between academia and practitioners by supporting the analysis and design of an industry's transition [2].

Startup readiness is a novel concept that is proposed in this dissertation as an earlier-applied criterion compared to TRL, as developed in my earlier studies with co-authors [3, 4]. Startup readiness is defined as the concept describing the state when one is prepared to initiate startups and willing to do so with a promise to be successful. While TRL is based on a scale from 1 to 9, with 1 being the most basic technology and with 9 being the most mature technology, startup readiness is assessment of such preparation and such promise, with a value between 0 and 1, determined as the probability of an academic researcher to be part of a startup participant class and a startup exit class. Given the growing attention and interest regarding academic startups among increasing stakeholders, academic startups have recently been gaining relatively easier access to venture capitalists and managerial talent. Long before their TRL matures, research topics and researchers together with enough startup readiness, which might later beget important firms by leveraging their scientific strengths, can be regarded as investment opportunities for venture capitalists and as career opportunities for

managerial talent. Moreover, startup readiness could be the focus of attention for university administrators and government policymakers too, some of whom recently tend to be keen to foster the creation of scientific startups.

When it comes to the scope of startup readiness to be addressed, this thesis emphasizes case studies in life sciences, specifically in the biopharmaceutical domain. The classic definition of ***biopharmaceutical***, both in science and industry, is pharmaceuticals (medicinal products, therapeutics, prophylactics and in vivo diagnostics) with active agents that are inherently biological in nature and which are manufactured using biotechnology (products manufactured by or from living organisms, usually involving bioprocessing). Additionally, a ***drug*** is defined as a pharmaceutical that is inherently chemical (not biological) in nature and which is manufactured using chemical methods. Biopharmaceuticals are distinct from drugs: drugs are small molecules or other synthetic chemical substances. The inherent differences between these two classes include product and active agent sources, identity, structure, composition, manufacturing methods and equipment, intellectual property, formulation, handling, dosing, regulation, and marketing [5].

Among the various life sciences domains, the biopharmaceutical domain was chosen for the case studies herein for the following several reasons.

(i) Much of the domain of life sciences such as biopharmaceutical research can be characterized as falling into the so-called Pasteur's quadrant, a classification of scientific research projects that are aimed at fundamental understanding of scientific problems and at providing immediate benefits to society. For that quadrant, many studies have found evidence of greater commercialization activities by academic entrepreneurship [6, 7, 8, 9, 10]. In fact, it is observed that, among 94,669 researchers related to most-emerging five biopharmaceutical research topics between 2014 and 2017, 3,156 researchers became participants of startups and 1,556 of them achieved exits (i.e., IPO's or M&A's) by 2018, which suggests that the biopharmaceutical domain could be a suitable field to explore from this perspective.

(ii) Since this thesis discusses academic researchers' startup readiness based on their research outcomes, domains which have intense science linkage with scientific founders are desirable. The biopharmaceutical field is appropriate because many founders themselves are from the academia. With growing attention and interest, the role of leading scientists who become entrepreneurial in this field is getting pivotal. Table 1.3 below lists examples of entrepreneurial scientists in emerging research topics of the biopharmaceutical domain. The academic startups they were engaged with have been successful either at raising venture capital, achieving an IPO, or being

acquired by big pharmaceutical companies according to a database named VentureSource. VentureSource, compiled by Dow Jones & Company, is a comprehensive global database of companies backed by venture capital and private equity in every region, industry, and stage of development. From the database, data of daily global startup investment deals can be extracted, with respect to each industry field with its specific industry code/subcode. Up-to-date information related to the amount of financing, the number of financing rounds, keywords, and participants of the startups are available.

Table 1.3 Startup Activities in Which Top Startup Participants Are Engaged Relative to Each Emerging Research Topic

Exosome				
Startup Participant	Role	Startup Company	Company Overview Brief Description	Most Recent Financing as of 03/01/2018 (MM)
Zhang, Bin	VP	Cisen Pharmaceutical Co. Ltd.	Manufacturer of chemical pharmaceutical agents and related products such as non- polyvinyl chloride (PVC) soft bag infusions, plastic bottle infusions, lyophilized powder injections, tablets, ointments, eye drops, and capsules	09/29/2017 IPO 1166.00RMB (Chinese Yuan)
Chen, Wei	Board Member, Outsider	Immunophotonics Inc.	Developer of a cancer treatment	09/04/2014 VC 1st 2.49 USD
Xu, Bin	EVP	Grandhope Biotech Co. Ltd.	Provider of medical materials and devices for the treatment of damaged tissue and organs	07/06/2011 IPO 278.4 RMB (Chinese Yuan)
Johansson, Henrik J.	Unknown Executive	Halo Genomics AB	Developer of targeted re-sequencing technology for DNA sequencing	12/01/2011 Acquired by Agilent

Microbiome				
Startup Participant	Role	Startup Company	Company Overview Brief Description	Most Recent Financing as of 03/01/2018 (MM)
Xavier, Ramnik J.	Cofounder	Jnana Therapeutics	Developer of drugs that target cellular proteins	12/14/2017 VC 1st 50.00 USD
de Vos, Willem M.	Chairman, Scientific Advisory Board	MicroDish BV	Developer of micro-engineered culture chips and nanoscale reagents aiming to improve microbial culture	03/31/2011 VC 1st N.A.
Mazmania, Sarkis	Director	Axial Biotherapeutics Inc.	Developer of biotherapeutics that target neurological diseases and disorders	06/22/2017 VC 1st 19.20 USD

(…Continued on Next Page)

(…Continued from Previous Page)

CRISPRR				
Startup Participant	Role	Startup Company	Company Overview Brief Description	Most Recent Financing as of 03/01/2018 (MM)
Zhang, Feng	Cofounder	Editas Medicine	Developer of human therapeutics based on genome editing technologies	02/03/2016 IPO 94.40 USD
Doudna, Jennifer A.	Cofounder	Editas Medicine	Same as above	Same as above
	Cofounder	Intellia Therapeutics Inc.	Provider of CRISPR-Cas9 focused biotechnology	05/06/2016 IPO 108.00USD
Joung, J. Keith	Cofounder	Editas Medicine	Developer of human therapeutics based on genome editing technologies	Same as above

Cas9				
Startup Participant	Role	Startup Company	Company Overview Brief Description	Most Recent Financing as of 03/01/2018 (MM)
Zhang, Feng	Cofounder	Editas Medicine	Developer of human therapeutics based on genome editing technologies	Same as above
Doudna, Jennifer A.	Cofounder	Editas Medicine	Same as above	Same as above
	Cofounder	Intellia Therapeutics Inc	Provider of CRISPR-Cas9 focused biotechnology	Same as above
Joung, J. Keith	Cofounder	Editas Medicine	Developer of human therapeutics based on genome editing technologies	Same as above

CAR-T				
Startup Participant	Role	Startup Company	Company Overview Brief Description	Most Recent Financing as of 03/01/2018 (MM)
June, Carl H.	Cofounder	Tmunity Therapeutics Inc.	Developer of T-cell immunotherapies	01/23/2018 VC 2nd 100.00USD
Sadelain, Michel	Cofounder	Juno Therapeutics Inc.	Developer of medicines to treat cancer	12/19/2014 IPO 264.55USD
Riviere, Isabelle	Cofounder	Juno Therapeutics Inc.	Same as above	Same as above

(iii) Table 1.4 presents three startups founded by academic researchers with strong academic capabilities in the biopharmaceutical field called CRISPR-Cas9 (See 0), all of which achieved IPO (initial public offering) in 2016 even way before the medical or commercial application of their research outcome. These startups were founded by researchers who have been ranked among the top 10 authors for one or more year(s) since 2012 based upon one or more of the five kinds of centralities among their paper

citation networks to be calculated in 3.1.1: betweenness centrality, closeness centrality, degree centrality, eigenvector centrality and PageRank, or number of citations. These startup examples initiated the origin of this thesis's hypothesis that, for academic startups with such intense science linkage, academic prominence expressed as paper bibliometrics of scientific founders could matter to assess their startup readiness.

Table 1.4 Chronological Comparison of Scientific Founders of High Network Centralities and Number of Citations to Their Startups' Founding & Funding Records

	Legend: Bw, betweenness centrality; **Cl**, closeness centrality; **Dg**, degree centrality; **Eg**, eigenvector centrality; **PR**, PageRank; and **Citation**, number of citations				
Editas Medicine					
	2012	**2013**	**2014**	**2015**	**2016**
History of Founding & Financing		Founded & VC-1st (Nov., $43.00 M)		VC-2nd (Aug., $120.0 M)	IPO (Feb., $94.40 M)
Zhang, Feng		6th in Bw and Dg	1st in Bw, Cl, Dg, Eg and PR	1st in Bw, Cl, Dg, Eg and PR	1st in Bw, Cl, Dg, Eg and PR
		2nd in Citation	1st in Citation	1st in Citation	2nd in Citation
Doudna, Jennifer	Tie for 7th in Bw, Cl, Dg, Eg and PR	11th in Bw and Dg	3rd in Bw	7th in Bw	6th in Bw
	Tie for 1st in Citation	11th in Citation	4th in Citation	2nd in Citation	1st in Citation
Church, George		1st in Bw, Cl, Dg, Eg and PR	2nd in Bw, Cl, Dg, Eg and PR	4th in Bw	5th in Bw
		1st in Citation	5th in Citation	5th in Citation	15th in Citation
Crispr Therapeutics					
	2012	**2013**	**2014**	**2015**	**2016**
History of Founding & Financing		Founded & VC-1st (Oct., $0.56 M)	VC-2nd (Apr., 10.83 M)	VC-3rd (Apr., $45.61 M) VC-4th (Apr., $28.00 M) Corporate (Oct, $105.00 M) Debt-Bridge/Convertible (Dec., $35.00 M)	VC-5th (June, $38.00 M) IPO (Oct., $56.00 M)
Charpentier, Emmanuelle	Tie for 7th in Bw, Cl, Dg, Eg and PR	22nd in Bw, Cl, Dg, Eg and PR	13th in Bw	20th in Bw	15th in Bw
	Tie for 1st in Citation	n/a	2nd in Citation	6th in Citation	12th in Citation
Intellia Therapeutics					
	2012	**2013**	**2014**	**2015**	**2016**
History of Founding & Financing			Founded (May) & VC-1st (June, $15.00M)	VC-2nd (Aug., $70.02M)	IPO (May, $108.00M)
Doudna, Jennifer	Tie for 7th in Bw, Cl, Dg, Eg and PR	11th in Bw and Dg	3rd in Bw	7th in Bw	6th in Bw
	Tie for 1st in Citation	11th in Citation	4th in Citation	2nd in Citation	1st in Citation
Barrangou, Rodolphe	Tie for 1st in Bw, Cl, Dg, Eg and PR	n/a	79th in Bw	88th in Bw	83rd in Eg
	Tie for 1st in Citation	n/a	Tie for 79thin Citation	13th in Citation	6th in Citation

(iv) The biopharmaceutical industry is known for its very high research and development spending. As reported in June 2018 by EvaluatePharma, a major provider of market intelligence and forecasts for the pharmaceutical industry, global R&D spending in 2017 surged by 3.9% to a record $165 USD billion compared to 2016, which also found that the R&D spending hit a new high in relative terms as percentage of sales 20.9%, substantially by virtue of the rise of global sales of

biopharmaceutical products. The report also indicates that overall R&D spending is expected to grow by 3% each year. [11]

(v) Biopharmaceutical science has expanded considerably in terms of both commercialization and entrepreneurship since the beginning of the 21st century. Of the 10 top-selling pharmaceutical products worldwide in 2018, 8 were biopharmaceutical; 6 had origins in startup companies (Table 1.5). By contrast, 2001 had only one biopharmaceutical drug without a startup origin [12].

Table 1.5 Top 10 Pharmaceutical Products by Global Sales in 2018 Compared to 2017 (USD)

18 Rank (17 Rank)	Product [a]	Therapeutic Subcategory	Mechanism of Action	Vendor Company	Originator [b]	2017 Sales ($m)	2018 Sales ($m)	Growth Per Year (%)
1 (1)	Humira [B]	Other anti-rheumatics	Tumor necrosis factor alpha (TNFa) antibody	AbbVie, Eisai	Knoll	18,923	20,472	8%
2 (3)	Revlimid	Other cytostatics	Interleukin-6 (IL-6) antagonist; Natural killer (NK) cell stimulant; Natural killer T-cell (NKT) stimulant; Tumor necrosis factor alpha (TNFa) inhibitor; Vascular endothelial growth factor (VEGF) inhibitor	Celgene, BeiGene	Celgene	8,211	9,809	19%
3 (9)	Opdivo [B]	Anti-neoplastic MAbs	Programmed cell death protein 1 (PD1) antibody	Bristol-Myers Squibb, Ono Pharmaceutical	Ono Pharmaceutical	5,761	7,565	31%
4 (2)	Enbrel [B]	Other anti-rheumatics	Tumor necrosis factor alpha (TNFa) inhibitor	Amgen, Pfizer, Takeda	Immunex, acquired by Amgen [S]	8,234	9,809	19%
5 (21)	Keytruda [B]	Anti-neoplastic Mabs	Programmed cell death protein 1 (PD1) antibody	Merck & Co., Otsuka Holdings	Organon BioSciences [S]	3,827	7,205	88%
6 (6)	Herceptin [B]	Anti-neoplastic MAbs	Epidermal growth factor receptor ErbB-2 (HER2) antibody	Roche	Genentech [S]	7,126	7,140	0%
7 (8)	Eylea [B]	Eye/Ophthalmic preparations	Vascular endothelial growth factor receptor (VEGFR) antagonist	Regeneron Pharmaceuticals, Bayer, Santen Pharmaceutical	Regeneron Pharmaceuticals [S]	6,291	7,070	12%
8 (7)	Avastin [B]	Anti-neoplastic MAbs	Vascular endothelial growth factor receptor (VEGFR) antibody	Roche	Genentech [S]	6,795	7,004	3%
9 (4)	Rituxan [B]	Anti-neoplastic MAbs	B-lymphocyte antigen CD20 antibody	Roche	IDEC Pharmaceuticals [S], merged with Biogen	7,528	6,925	-8%
10 (14)	Eliquis	Anti-coagulants	Coagulation factor Xa inhibitor	Bristol-Myers Squibb	DuPont Pharmaceuticals	4,872	6,438	32%

Source: Evaluate Ltd "Top 100 Products in 2024"

a B indicates that the product belongs among biopharmaceutical products.

b S indicates that the product has origins in startups

(vi) Life sciences, especially those in the biopharmaceutical domain, can be characterized by their discrete development and absence of network effects. Thanks to this sort of the nature of technology and network effects in market, innovators in this

domain do not need to fear other players, in terms of infringing intellectual property rights and having problems accessing complementary assets. Policies that stimulate spin-offs from universities by scientists cater to this "romantic" view of technological innovation, which could enable government and academic administrators to focus on stimulating individual academic entrepreneurs by creating circumstances that facilitate them [13]. In other words, the biopharmaceutical field is a "romantic" domain in which one can observe a positive spiral composed of abundant startups with intense scientific linkage and active R&D, both of which lead to proactive venture financing.

The motivation of this dissertation is to show an assessment method that can be built based on real-time or timely available digital data that is purchasable or publicly available, completely independent of personal interviews or customized surveys of scientists and other stakeholders related to their academic and entrepreneurial activities. Such method will enable us to collect data in a scalable manner, without limitation of our acquainted sources, whereas earlier studies survey past data of scientists in specified academic organizations or regions. This method can be used even by stakeholders with little or no expertise related to specific disciplines, industries and/or regions.

1.2. Literature Study

For factors contributing to new academic firm creation, earlier researchers have used surveys of scientists reached by their research protocols, and have assessed individual and non-individual determinants of academic entrepreneurship. For instance, Rothaermel, Agung, and Jiang (2007) report that university policy, faculty, technology transfer offices, investors, founding teams, networks in which a firm is embedded, and other external conditions affect new firm creation [14]. Bercovitz and Feldman (2008) examine individual backgrounds and work environments of faculty members and their subsequent engagement in academic entrepreneurship. They find that participation is more likely at institutions where they trained if they had accepted the new initiative and had been active in technology transfer [15]. Jain, George, and Maltarich (2009) investigate the sense-making process accompanying university scientist participation in academic entrepreneurship and potential modification of their role identities. They suggest that scientists participate to preserve their academic role identities [16]. Clarysse, Tartari, and Salter (2011) examine how academic professionals' opportunity recognition capacity and their prior entrepreneurial experience shape the likelihood of their involvement in starting up a new venture and shape the roles of university technology transfer offices and the social environment [17]. Abreu and Grinevich (2013) analyze the

determinants of academic engagement, varying from demographic factors such as seniority and gender, to the research type, to entrepreneurial experience and training, and to institutional support [18]. Aldridge, Audretsch, Desai, and Nadella (2014) examine the role of scientist characteristics including academic rank, experience, networks and industry ties, access to human and financial resources, and supportive university conditions, in driving the likelihood of scientists to start companies [19]. Furthermore, my earlier studies with co-authors (2017, 2018, 2019), in light of bibliometric approaches, examine a researcher's different measures of academic centrality, such as the degree of centrality as an author in co-authorship networks, and the frequency of an author being a corresponding author or a first author, to assess the researcher's startup readiness, or likelihood of being a founder or a participant of a startup [3, 4, 20].

Related to the determinants presented above, with respect to startups in the so-called biotech clusters, earlier research efforts have evaluated factors conducive to entrepreneurship near research institutes and pharmaceutical companies. For instance, Auerswald and Dani (2017) point out a transition of entrepreneurial activity in the region from a dynamic driven by federal research spillovers, to one increasingly driven by private sector actors [21]. Curran, van Egeraat, and O'Gorman (2016) emphasize that founders' pre-entry experience related to the private sector is important to attract venture capital [22]. Allen, Gloor, Colladon, Woerner and Raz (2016) argue that location per se does not influence innovation success, but that a dynamic communication style and more diverse social ties are beneficial for innovation [23]. All of these studies imply that socio-economical non-individual determinants, such as venture capital activity level, ease of startup creation, and professional voluntary turnover, in particular, can affect entrepreneurship in biotech clusters. They will be touch upon later as part of Ecosystem Factors.

Some earlier researchers even conducted regression analyses of individual and non-individual determinants. For instance, Landry, Amara, and Rherrad (2006) present a model showing that a complementary set of resources, including financial, intellectual, knowledge, social and personal assets, must be mobilized by researchers to launch startups, albeit with little emphasis on research papers in which "publication assets" were not found to have substantial impact on spin-off activities by researchers [24]. Krabel and Mueller (2009) present a model suggesting that close ties to industry established through joint research projects with private firms, patenting activity, and prior founding experience are the most important factors enhancing activities for starting a business, whereas work experience in the private sector seems to be unimportant [25]. Criaco, Minola, and Migliorini (2014) present a model showing that industry human capital negatively affects university startup survival, whereas university human capital and

entrepreneurship human capital enhance the likelihood of university startup survival [26]. Huynh, Paton, Arias-Aranda, and Molina-Fernandez (2017) present a model demonstrating that entrepreneurial capabilities of a founding team positively influence the performance of a spin-off during the growth phase [27].

However, earlier research on determinants for the creation of startups commercializing scientists' research, makes very little or no reference to any of the following: (i) selection of specific "hot" topics that have attracted attention of academic researchers as well as startup stakeholders like investors and managerial cofounders (although Abreu and Grinevich (2013) introduce life sciences as research fields with greater commercialization activity) [18]; (ii) startups' potential of financial exits such as IPO and M&A that are of much interest as financial success for stakeholders such as venture capitalists and managerial talents (although Aldridge, Audretsch, Desai, and Nadella (2014) analyze technology transfer offices' knowledge communication and commercialization between academia and industry, as a key success factor) [19]; (iii) bibliometric analyses of those entrepreneurial scientists specifically in terms of their research domain (although Rothaermel, Agung, and Jiang (2007) and Aldridge, Audretsch, Desai, and Nadella (2014) assess academic titles such as professor) [14, 19], and (iv) the interactions of the factors/features that belong to the same or different factor/feature category(ies).

Moreover, to explain the concept of startup readiness of researchers, this thesis draw on the resource-based view of firms (Barney, 1991; Kogut and Zander, 1992; Conner and Prahalad, 1996; Grant, 1996) and its extended literature related to academic startups (Landry, Amara and Rherrad, 2006; Rasmussen and Borch, 2010; Knockaert, Spithoven, and Clarysse, 2010; Huynh, Patton, Arias-Aranda, and Molina-Fernández, 2017; Corsi, Prencipe, and Jesus Rodriguez-Gulias, 2019) to assume that, like entrepreneurs, startup readiness by academic researchers will increase when either the resources or their coordination will be appropriate or sufficient [28, 29, 30, 31, 32, 27, 33, 24, 34].

This thesis fundamentally agrees with earlier literature on the commercialization of academic research, in that resources that enable startup creation include knowledge assets, intellectual property assets, financial assets, social capital assets, personal assets, and organizational assets. Except for organizational assets that are part of environmental factors, these assets belong to individual factors. However, as discussed before, when it comes to the biopharmaceutical domain that is a very intense science-based technology commercialization field, it is assumed that the criticality of their knowledge assets and intellectual property assets, which shows their scientific prominence and innovation mindset, is more pivotal than that of other individual assets.

Overall, earlier studies survey past data of scientists in specified academic organizations or regions. Those methods have collected data using conventional methods such as personal interviews, reading published papers, or conducting field projects in person, offering only limited instantaneity, comprehensiveness, and scalability.

Although supporting the basic view of resource-based theory, given the aforementioned shortcomings of earlier research and the scope of this study that are described in 1.1, this dissertation attempts to propose a conceptual framework to tackle or alleviate the weakness of prior research and specifically suit the purpose of the following research questions.

1.3. Research Questions

The main goal of this dissertation is to take several significant steps toward digital approach that could query data sources related to the domain as discussed, navigate through results, track determinants (explanatory variables), and assess and interpret academic researchers' startup readiness. To reach this goal, this thesis addresses the following research questions (RQs), with main focus on Primary RQ and adequate attention to two Secondary RQs that are contributory to Primary RQ.

Primary RQ: What are the implications of this empirical research using the logistic regression model to assess academic researchers' startup readiness based on the variables derived and constructed from the relevant digital data sources, related to the growing topics of interest in the biopharmaceutical domain?

Having data on target and explanatory variables associated with the logistic regression model to be hereinafter referred to, regarding academic researchers' startup readiness, it is the interest of this thesis to look for implications of the model by conducting empirical research in the biopharmaceutical domain. Desired characteristics of such model include (a) how well it fits a set of observations and how adequately its diagnostic ability performs as its discrimination threshold is varied, (b) how well it expresses and assesses each academic researcher's startup readiness, (c) how well it enables to interpret explanatory variables per each researcher group, and so on. (a) Measures of goodness of fit typically summarize the discrepancy between observed values and the values expected under the model in question. Such measures can be used in statistical hypothesis testing, to test whether outcome frequencies follow a specified distribution. Diagnostic ability can be measured by analyzing the relationship between true positive rate (sensitivity, recall, probability of detection) and true negative rate (specificity), at various threshold settings. (b) Startup readiness per each researcher can be computed and expressed between 0 and 1 for assessment at a given observation time.

(c) Interpretation of explanatory variables are conducted in terms of (i) each explanatory variable's mean, standard deviation (SD) and distribution, (ii) effects of explanatory variables across researcher groups' assessment models, that are measured as the number of times the odds of each author's (researcher's) startup readiness increases regarding their target variables, (iii) importance of each set of factors/features for assessment, that is measured by comparing different feature/factor sets, and (iv) influential values that could affect the model. These implications are important for the model's explainability and effectiveness for our decision-making and assessment purposes. This Primary RQ is addressed in Chapter 5 and further discussed in Chapter 6.

Secondary RQ1: What are the potentially essential factors/features that can be derived from relevant digital data sources, to assess startup readiness of academic researchers who have intense scientific linkage such as those in the biopharmaceutical domain?

Prior researchers suggest that commercialization of academic research will be enhanced when either the resources or their coordination are appropriate or sufficient, arguing that various individual assets and environmental assets matter. Although this thesis fundamentally agree with their resource-based theory in earlier literature (to be discussed in 1.2), when it comes to the biopharmaceutical domain that is a very intense science-based technology commercialization field, it is assumed that scientific prominence and innovation capability of academic researchers are critical, which was not fully explored in earlier literature. This thesis tries to delve into academic researchers' knowledge assets and intellectual property assets as paper- and patent-related features that could signal scientific prominence and innovation capability by tapping into digital data sources specifically pertaining to papers and patents. Features regarding the ecosystem essential to academic entrepreneurship related to this domain, such as data on academic organizations and nations, are explored by referring to several relevant digital data sources as well. This thesis also tries to extract features from relevant digital sources that could signal how attractive ("hot") research topics are, for business stakeholders such as venture capitalists and managerial talents with financial, social, and personal assets. On the other hand, academic researchers' financial assets, social capital assets and personal assets, all of which are also part of conventionally considered individual assets as knowledge assets and intellectual property assets, are not addressed in this thesis partly due to the lack of digital data sources. This RQ is addressed in Chapter 2.

Secondary RQ2: What are the appropriate methodologies to be deployed, in order to construct a logistic regression model to assess academic researchers' startup readiness,

with respect to preprocessing data, selecting and constructing variables, and, building and implementing the model?

One of the important aims of this thesis is to show a method to construct a logistic regression model to assess academic researchers' startup readiness in the biopharmaceutical domain, that allows us to preprocess digital data sources, select and develop explanatory variables relative to each researcher's startup readiness, and build and implement the assessment model.

As the first step in designing a set of explanatory variables to assess startup readiness of academic researchers, it is necessary to identify target variables, i.e., variables that show startups' creation and success which signal startup readiness. In comparison with creation of academic startups that can be measured by observing their founding with academic researchers typically as their cofounders, definition of their "success" needs further argument. For the practical purpose of considering as many stakeholders as possible who could be their equity holders, it is presumed that an academic startup achieves "success" when it accomplishes "exit," be it an IPO or M&A. These questions are addressed in Chapter 3 and Chapter 4, albeit mainly in the latter. Regarding explanatory variables, since just raw data extracted from various digital data sources as discussed above, is neither complete, nor easy to interpret, it is important to overcome or alleviate the limitations of raw data themselves and turn them into an interpretable complementary set of effective explanatory variables. These questions are addressed in Chapter 4. It is also necessary that, together with the above target and explanatory variables, the assessment model of each researcher's startup readiness can be constructed and implemented. These questions are answered in Chapter 4 and Chapter 5.

1.4. Thesis Contributions

The findings from this study make several contributions to the current literature. Firstly, to explain startup readiness of academic startups with higher scientific linkage as well as higher R&D funding need, such as those in the biopharmaceutical domain, this study adds critical determinants to the conventional resource-based theory of firms (discussed in 1.2). One type of the determinants added by this thesis is, network centralities calculated for different authors based on their paper networks, as developed in Chapter 3. Secondly, this thesis presents several challenges of raw data as well as a complementary set of constructed determinants to assess startup readiness of academic researchers in one big picture. Thirdly, this study shows an assessment model that yields good assessing and classifying performance regarding academic startups' startup readiness.

This thesis enhances our knowledge of such startup readiness to aid decision-making and suggest a new assessment method for such purposes. One of the findings is newly constructed factors called ***Interaction Terms Factors*** which are combinations of original factors or their transformed forms across and within each category and each sub-category. The result indicates that startup readiness depends not only on explanatory variables separately, but also on externalities and spillovers from various explanatory variables that combine with other variables.

A key strength of this study is its focus on digital data. All of the data is available publicly or for purchase on a real-time or timely basis, irrespective of whether we can reach academic researchers individually. By consolidating researchers' names with this digital data, individual or environmental, this method is scalable for application in various fields and is adaptable to situations in which the cycles of individual or environmental changes are short and thus information needs to be updated often. This allows us to make effective intelligent decision-making to match each stakeholder's need, because the method proposed by this thesis can be used even by people with little or no expertise of specific disciplines. Compared to the conventional research methods, the computational approach discussed in this thesis could provide global, comprehensive, yet convenient and real-time understanding of academic researchers' startup readiness as discussed.

It is expected that the approach proposed by this dissertation can contribute to a wide range of researchers, business practitioners, university administrators, and policymakers, all of whom attempt to foster creation of research-based startups at an earlier stage, in a more timely manner, on a larger scale, and in greater detail, compared to conventional methods such as personal interviews and customized surveys. It can allow business professionals such as venture capitalists and managerial entrepreneurs, to compute the startup readiness of researchers in emerging research topics, thereby enabling them to evaluate potential scientific founders to work with. Also, this method can allow university administrators and policymakers to implement pro-entrepreneurship policies more effectively than earlier approaches because it enables ready identification of the variables and interaction terms that are important for startup readiness. In addition, this method can enable researchers to understand the variables that can be improved to increase their startup readiness.

Expert Interview: Three professionals relevant to the research questions were interviewed regarding suggested approaches and findings of this thesis, to have a better understanding of potential and actual application of this research. Please refer to 6.3

1.5. Thesis Structure

This thesis starts with the introduction of the rise of academic startups and conventional access approaches to their available information, challenges of which are discussed in Chapter 2. The introduction also addresses the scope of this thesis, i.e., the biopharmaceutical domain.

Chapter 2 proposes a conceptual framework and data sources for assessing such academic researchers' startups readiness throughout this thesis.

Chapter 3 is a chapter specifically dedicated to explore distinctive features to assess startup readiness of academic researchers in the biopharmaceutical domain. This chapter explores such researchers' paper-related network centralities as features to express their academic prominence, and their hot topic features as alternatives to conventionally-regarded individual factors.

Chapter 4 focuses on designing the assessment model. In this chapter, target variables Participant and Exit, and potential explanatory variables, part of which are explored in Chapter 3, specifically for researchers in the relevant domain, are employed to construct the model.

Chapter 5 implements the assessment model, leading to the computation of each researcher's startup readiness and the assessment of each explanatory variable's importance. In the former part of the chapter, cases of academic researchers related to Cas9 and Microbiome, classic emerging biopharmaceutical research topics, are discussed. The latter part discusses the most emerging five biopharmaceutical research topics combined, including Cas9 and Microbiome.

Chapter 6 discusses evaluations of this model and implications of the results. This chapter interprets characteristics of variables/features in each set, effects of each variable/feature, and importance of each set of variables/features in this model.

Chapter 7 is dedicated to draw the main conclusions and elaborate perspectives including future steps of this research.

Chapter 2. Conceptual Framework and Data Sources

2.1. Overview of Proposed Conceptual Framework

Earlier literature on the commercialization by academic researchers argues that resources that enable startup creation include knowledge assets, intellectual property assets, organizational assets, financial assets, social capital assets, and personal assets [14, 15, 16, 18, 19] all of which assumedly can be categorized as either "individual" or "environmental" factors. Although this thesis basically agrees with the earlier literature's view called the resource-based theory [28, 29, 30, 31, 32, 33] in that these "individual" and "environmental" assets have effects on creation of academic startups as well, since the scope of this thesis (as discussed in 1.1) is specifically focused on academic startups with high scientific linkage and R&D funding needs, conventional views of these assets should need several modifications for the purpose of this thesis.

Firstly, in the domain that is a very intense science-based technology commercialization field as exemplified in the biopharmaceutical domain herein, among "individual factors" of academic researchers, this thesis assumes that the essentiality of their "knowledge assets" and "intellectual property assets," both of which show their scientific prominence and innovation capability, is much higher than other assets. This thesis presumes that scientific prominence can be measured by ***Paper-related Features*** such as counts of publications, frequency of citation of a researcher's papers, frequency of a researcher's corresponding authorship in papers, centralities of a researcher in the networks of author citation and/or co-authorship. Innovation capability is assumed to be measurable by ***Patent-related Features*** such as counts of the patents that the researcher is an inventor on, and frequency of citation of the author's invented patents.

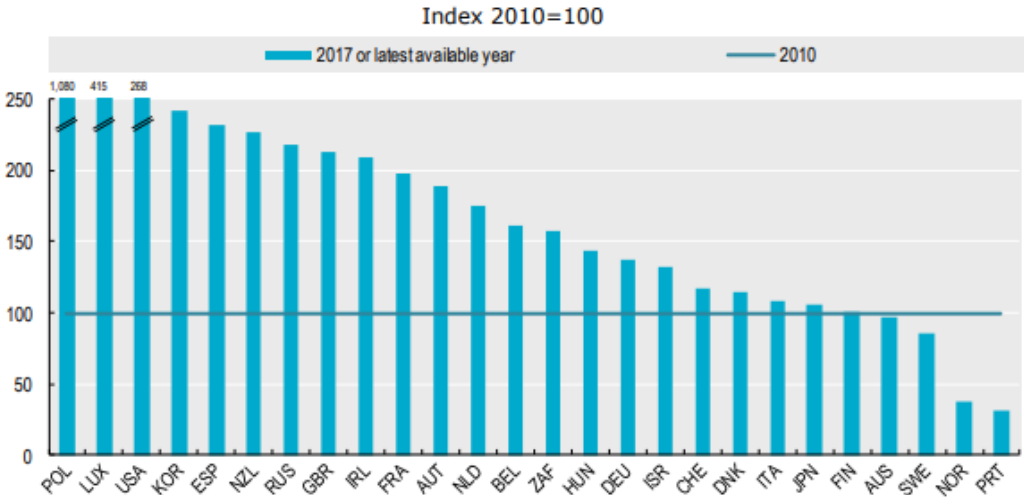


Figure 2-1 Trends in Venture Capital Investments
(OECD "Entrepreneurship at a Glance 2018")

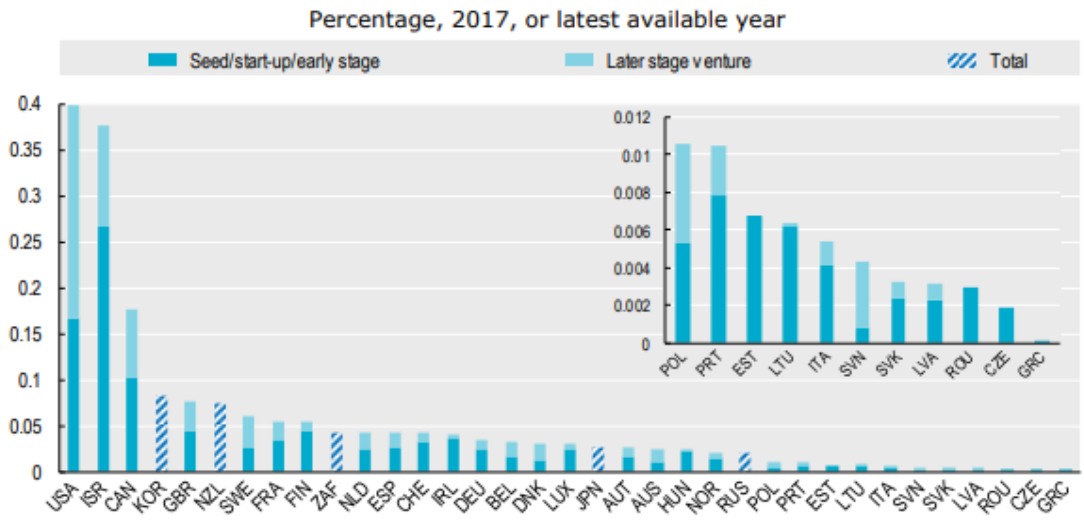


Figure 2-2 Venture Capital Investments as a Percentage of GDP
(OECD "Entrepreneurship at a Glance 2018")

Secondly, given the rapidly growing attention related to academic startups across experts in various fields as discussed in Chapter 1, this thesis presumes that academic researchers' "individual assets" other than "knowledge assets" and "intellectual property assets", such as "financial assets", "social capital assets", and "personal assets," can be complemented, even considerably replaced by competent venture capitalists and entrepreneurs, to the adequate extent to which these business partners are attracted to the startup opportunity. In the majority of OECD countries, venture capital investments are considerably growing, although they still constitute a small percentage of GDP, except for Israel and the United States (Figure 2-2 and Figure 2-1) [35]. The pool of managerial talent is also significantly increasing, as new enterprise creations hit record highs in around half OECD countries (Figure 2-3). With these recent trends, a growing number of competent venture capitalists and entrepreneurs are willing to work with academic researchers with strong "knowledge assets" and "intellectual property assets," to create and develop academic startups with "hot" research topics. To serve this purpose, ***Hot Topic Factors*** are incorporated into this thesis to measure the degree of social attention from financial, scientific and innovative perspectives to specified discipline or research topic in question. One could argue that these Hot Topic Factors could be categorized as part of the following ***Ecosystem Factors*** in a larger sense, but Hot Topic Factors are fractionated herein as more short-term and transient factors to measure how much the relevant discipline/topic is emerging, and intended to function as more of alternatives than compliments with respect to the aforementioned "individual assets", compared to the Ecosystem Factors as follows.

Thirdly, this thesis introduces Ecosystem Factors specifically composed of ***Academic Organization-related Features*** and ***Nation-related Features***, both of which

are profoundly relevant to academic startups' ecosystem. Academic Organization-related Features include features that suggest academic eminence such as research score of the academic organization to which academic researchers' corresponding authors belong. ***Nation-related Features*** include features that are favorable to academic startups of the country that those corresponding authors in question belong to, such as venture capital availability, ease of startup creation and human talent liquidity in life sciences industry. In selecting these features, this thesis referred not only to earlier studies described in 1.2, but also to the experience of The University of Tokyo Edge Capital Co., Ltd. (UTEC), a venture capital firm focusing on seed/early-stage academic startups, that I have led since its founding back in April, 2004 with my partners, as president and managing partner ever since February, 2006.

Figure 2-3 New Enterprise Creations, Selected Countries (OECD "Entrepreneurship at a Glance 2018")

Lastly, after adding appropriate selection and transformation on the above factors/features, their interaction terms are formed and introduced as ***Interaction Terms Factors***, to consider synergetic effects across potential variables. Azoulay, Ding, and Stuart (2009) propose that patenting has a positive effect on the rate of publications and that patenters may be shifting their research focus to questions of commercial interest [36]. Fehder, Murray and Stern (2014) argue that, for scientific discoveries with potential commercial applicability, researchers may seek to establish patents in addition to papers, which allows researchers to influence follow-on access to knowledge disclosed in a given scientific journal [37]. In lite of these earlier studies, this thesis presumes that combinations of some variables, including but not limited to paper-related and patent-related features, might work more (or less) effectively as determinants of startup readiness, compared to their solo variables.

In sum, this thesis proposes a conceptual framework to assess academic researchers' startup readiness, using essential individual factors (Paper-related Features and Patent-related Features), non-individual factors (Hot Topic Factors and Ecosystem factors) and their Interaction Terms Factors, as depicted in Figure 2-4 and will be described in 4.4 in detail. Given the resource-based view in prior studies and consideration of the characteristics of biopharmaceutical academic startups that have intense science-linkage, this proposed framework functions with (i) Essential Individual Factors, that are composed of Paper-related Features (specifically to be explored in Chapter 3) and Patent-related Features, both of which are critical factors for academic researchers, (ii) Hot Topic Factors that signals attractiveness for business partners who could complement academic researchers' lack of individual factors other than Essential Individual Factors, (iii) Ecosystem Factors composed of Academic Organization-related Features and Nation-related Features, and (iv) Interaction Terms Factors composed of combinations of the above factors with synergetic effects across them. This framework, incidentally, abbreviates several of conventionally perceived individual assets of researchers in earlier studies, such as financial assets, social capital assets, and personal assets in particular. This is partly because digital data sources that possess data on such conventionally perceived individual assets cannot be found as of today, and partly because this thesis presumes that, when creating and growing startups, researchers' such assets can be complemented or even replaced by sufficient such assets of business partners like venture capitalists and managerial talents, when relevant research topics are "hot" enough.

Details of explanatory variables and features of this framework will be discussed in 4.4.

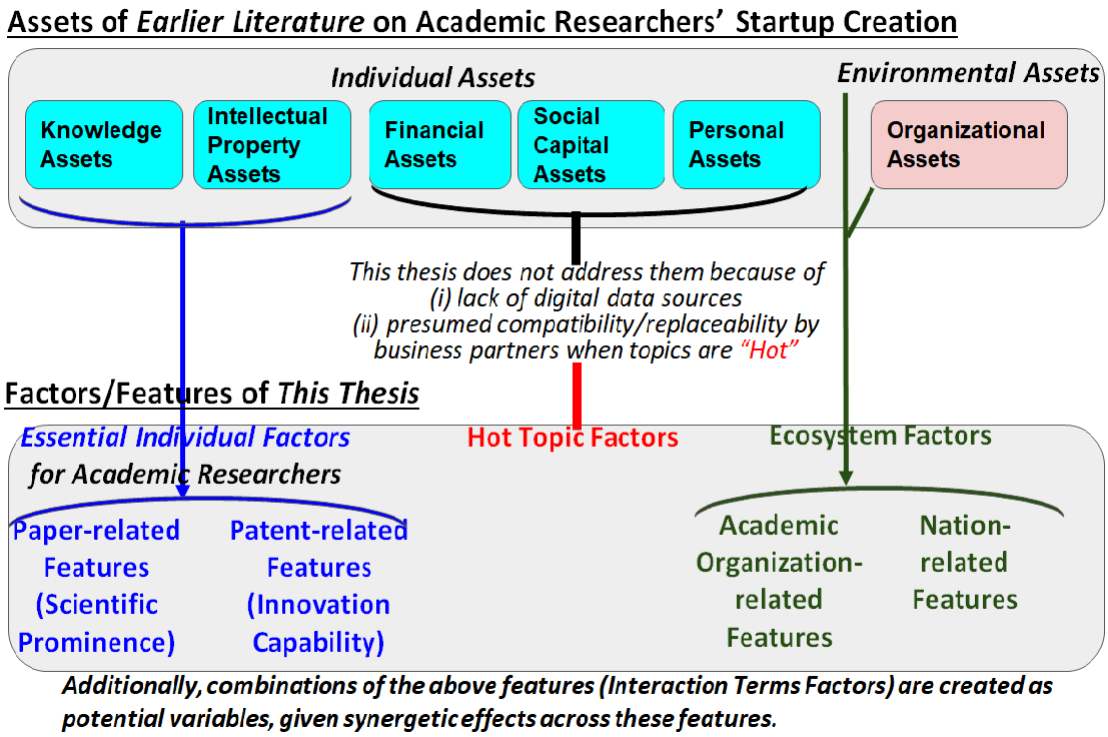


Figure 2-4 Conceptual Framework to Assess Academic Researchers' Startup Readiness

2.2. Challenges

It is acknowledged that information of various academic researchers' assets and factors, other than those employed and computed herein, can be used instead of or in addition to the ones this thesis has used. For example, it is plausible that researchers' financial, social capital, or personal assets or traits might affect their startup readiness in general. It could also be argued that researchers' qualitative assets or traits that are not computable, might influence their startup readiness.

However, such approaches have not been employed herein. Rather, the focus of this thesis is placed on collecting computable digital data that are promptly obtainable from openly available digital databases, regarding startup finances, papers, patents, academic organizations and national economics. At the same time, the thesis attempts to incorporate as many varieties of relevant data as reasonably possible regarding academic startups within the scope of this paper. Part of the reason for this is that the thesis tries to build an assessment method for a range of stakeholders: not only for academic researchers, but also for potential stakeholders presumably with little or no knowledge related to the concerned disciplines. Venture capitalists are a classic example. Even though it might be desirable to understand a whole set of assets and factors of academic researchers with whom they might invest, venture capitalists cannot afford to conduct

detailed interviews or personal surveys of researchers before seriously considering investment.

Also, most importantly, in the academic startup ecosystem that this thesis addresses, where intensive scientific linkage and high R&D funding coexist, abundant computable data are affordable which could be translated into an array of variables that this dissertation uses. Utilizing these data could be reasonable and feasible for most stakeholders, even those outside the academia, such as venture capitalists and managerial talents with a descent level of data science knowledge. Therefore, the proposed method, by merely using affordable digital data rather than published papers, interviews or surveys, is scalable for application to various disciplines in question and is also adaptable to situations in which the publication cycle is short and the publications are numerous.

2.3. Data Sources for Conceptual Framework

To build the conceptual framework shown in Figure 2-4, this thesis first cultivated the following two data sources: Web of Science Core Collection database and VentureSource database. Both databases are updated daily to include the latest information of scientific papers and startups respectively, and their data are available on subscription basis.

Web of Science (previously known as Web of Knowledge) is an online subscription-based scientific citation indexing service originally produced by the Institute for Scientific Information, later maintained by Clarivate Analytics (previously the Intellectual Property and Science business of Thomson Reuters), that provides a comprehensive citation search. It has a curated collection of over 20,000 peer-reviewed, high-quality scholarly journals published worldwide (including Open Access journals) in over 250 science, social sciences, and humanities disciplines. This thesis used The Web of Science Core Collection to extract data of papers that included the research topic in question, for calculation of each author's several types of citation centralities as well as co-authorship centralities in authors' citation networks and co-authorship networks respectively. This database also allows us to calculate authors' number of citations, first authorship and corresponding authorship. Obviously, this data source allows us to detect a various types of features representing academic researchers' scientific prominence and presence, as discussed later in 4.4.1.1.

VentureSource is compiled by Dow Jones, which is a comprehensive global database of companies backed by venture capital and private equity in every region, industry, and stage of development. We can extract data of daily global startup investment deals, with respect to each industry field with its specific industry code/subcode. Information related to the amount of financing, the number of financing

rounds, keywords, and participants of the startups are also available. This thesis used VentureSource for the purpose of seeking startups that authors have founded or participated in, and even searching startups that have financially exited, by executing queries of target authors' names.

As discussed in 2.1, in the domain that is a very intense science-based technology commercialization field as exemplified in the biopharmaceutical domain herein, it is presumed that "knowledge assets" such as ***Paper-related Features*** that show academic researchers' scientific prominence are not sufficient as academic researcher-specific ***Individual Factors***. Rather, academic researcher-specific Individual Factors should also include "intellectual property assets," which show their innovation capability too. Innovation capability is assumed to be measured by ***Patent-related Features***, such as counts of the patents that the researcher is an inventor on, how frequent the patents that the author is an inventor on were cited by other patents, to be discussed in 4.4.1.2. That is why this thesis additionally employed a database called Derwent Innovation, which is a patent database compiled by Clarivate Analytics.

Derwent Innovation - Derwent World Patents Index is a subscription-based database complied by Clarivate Analytics, containing patent applications and grants from 52 of the world's patent issuing authorities since 1963. Compiled in English by editorial staff, the database provides a short abstract detailing the nature and use of the invention described in a patent and is indexed into alphanumeric technology categories to allow retrieval of relevant patent documents by users, which enhances searchability and discoverability of patent data.

Moreover, in line with Individual Factors, this thesis introduces ***Hot Topic Factors***, as measures to allow stakeholders other than academic researchers, such as venture capitalists and managerial talents, to complement or even replace several of researchers' individual assets other than "knowledge assets" and "intellectual assets." It is assumed that Hot Topic Factors that attract such stakeholders would render researchers' lack of financial, social, personal assets and other relevant personal traits, being unimportant or even negligible, so that only the "knowledge assets" and "intellectual property assets" matter as academic researchers' (essential) Individual Factors.

To serve this purpose, Hot Topic Factors/Features are incorporated into this thesis, to measure the degree of social attention from financial, scientific and innovative perspectives to specified discipline or research topic in question. In other words, these factors/features can be indexes of how much emerging the discipline/topic is. To compute Hot Topic Factors, this thesis used Web of Science Core Collection, VentureSource, and Derwent Innovation Derwent World Patents Index, to measure how hot the research topics are in terms of paper, financing and patents.

Finally, it is presumed that in addition to Individual Factors in association with Hot Topic Factors, ***Ecosystem Factors*** also matter for academic researchers' startup readiness, regarding academic organizations and nations, to be discussed later in 4.4.2 and 4.4.3. As data sources to extract such Ecosystem Factors, this thesis incorporated (i) *The Times Higher Education - World University Rankings 2017* and (ii) *Reuters - The World's Most Innovative Universities 2017,* for several academic organization features; and (iii) (a) *OECD - Entrepreneurship at a Glance 2017*, (b) *World Bank - Doing Business 2017*, and (c) *Mercer - 2017 Workforce Turnover Around the World*, for several nation features.

The Times Higher Education - World University Rankings is the definitive list of the top universities globally, compiled by The Times Higher Education and independently audited by PricewaterhouseCoopers, which is freely available online. It includes more than 1,250 institutions across 86 countries in 2019. It is the only global university league table to judge research-intensive universities across each one of their core missions: teaching; research; international outlook; citations; industry income.

Reuter – The World's Most Innovative Universities, compiled by Reuters, is a list that measures the innovative capacity and achievement of universities. This ranking is composed of ten indicators: Patent Volume, Patent Success, Global Patents, Patent Citations, Patent Citation Impact, Percent of Patents Cited, Patent to Article Citation Impact, Industry Article Citation Impact, Percent of Industry Collaborative Articles, and Total Web of Science Papers. While data is published specifically for the top 100, the analysis covers 600 universities.

OECD - Entrepreneurship at a Glance, is a publication produced by the OECD-Eurostat Entrepreneurship Indicators Programme, based on official statistics. This includes a statistic named "VC Investments as a Percentage of GDP," which measures each country's activeness in venture capital financing.

World Bank - Doing Business, is a World Bank Group flagship publication, in a series of annual reports measuring the regulations that enhance business activity and those that constrain it. Doing Business presents quantitative indicators on business regulations and the protection of property rights that can be compared across 190 economies and over time. Doing Business measures regulations affecting 11 areas of the life of a business. This deals with rankings on the ease of doing business such as starting a business.

Mercer - *Workforce Turnover Around the World*, compiled by Mercer, is a purchasable report that explores information regarding voluntary and involuntary turnover for over 100 markets, across six career levels, by region, market, and industry, which includes life science industry specifically.

All of the above data sources are available publicly or for purchase on a real-time basis. Chapter 3 and Chapter 4 will discuss how to explore these data sources later.

Chapter 3. Exploring Distinctive Features to Assess Academic Researchers' Startup Readiness in Emerging Fields

Following the discussion of Chapter 2 regarding desirable factors for academic researchers in the biopharmaceutical domain to assess their startup readiness, this chapter explores their potential factors to develop, specifically with the aim to develop their Paper-related Features as the features to express their academic prominence, and their Hot Topic Features as alternatives to conventionally-regarded "individual factors" other than Paper-related Features and Patent-related Features. Although earlier studies have addressed various kinds of "individual" and "non-individual" factors as researchers' determinants for their creation of academic startups, few earlier studies have focused on potential factors that signal those researchers' emerging academic prominence as their important "individual factors." Moreover, few studies have researched factors to measure how much emerging the field/topic is, to allow for depreciating Individual Factors other than essential attributes as academic researchers.

Web of Science Core Collection, this thesis's data source to collect data on relevant research topics, is enriched in academic researchers' paper-related features, such as counts of publications, frequency of being cited by other papers as well as citing them, and frequency of being a corresponding author and a first author (to be described in 4.4.1.1), all of which can be attained simply by performing additions. However, as discussed in 1.1 (v), since this thesis addresses the biopharmaceutical domain that is an emerging research field characterized by their discrete scientific development and an positive spiral composed of academic entrepreneurs with intense scientific linkage and active R&D, this chapter tries to develop new features that surpass those features that were just summed up, in order to understand emerging academic capability among relevant academic researchers and assess startup readiness of the future core researchers.

Earlier studies on detecting emerging research fields that used bibliometric approaches include the works of Shibata, Kajikawa, Takeda and Matsushima (2008) that divided citation networks into clusters using the topological clustering method and tracked the positions of papers in each cluster [38], and Shibata, Kajikawa, Takeda, Sakata and Matsushima (2011) that calculated network centralities called betweenness centralities of papers with respect to regenerative medicine [39], and Sasaki, Hara and Sakata (2016) that calculated nine kinds of network centralities (degree centrality, betweenness centrality, closeness centrality, eigenvector centrality, network constraint, clustering coefficient, Page rank, hub score, and authority score) with respect to papers in solar cells field [40]. Although both studies did not address each researcher's preparedness to create startups, they attempted to propose a prediction model to identify

emerging promising studies that could attract many citations, to facilitate decision-making processes. This chapter tries to dilate the application of such network centrality, to explore and develop new Paper-related Features to assess academic researchers' startup readiness in the biopharmaceutical domain, as follows, by referring to my prior works with co-authors (2017, 2018) [3, 4]. First, as will be seen in 3.1, the research topic of CRISPR-Cas9 is addressed, a typical biopharmaceutical field rapidly emerging since 2012 that generated three IPOs in the U.S. in 2016. Results showed that, among the top 100 authors, authors with higher network centralities (betweenness centrality, closeness centrality, degree centrality, eigenvector centrality) in their author citation networks have higher rates of being founders, with potential to let their startups raise initial VC funding, similarly to the number of citations, a conventional bibliometric index. Furthermore, it became evident that theses centralities could serve as better features of scientists' potential to become founders, reflecting their startup readiness, because the centralities might encompass a wider range of potential founders. [3]

Secondly, in 3.2, by using VentureSource, "hot" industry codes/subcodes are sorted in that their financing activities are active, and subsequently the codes/subcodes that belong to the biopharmaceutical segment are detected as highlighted in yellow (See Table 3.1). Then, VentureSource is used again to extract key words of startups belonging to those codes/subcodes (See Table 3.7), then query "hot" research topics on Web of Science Core Collection (See Table 3.8), in that related growing research topics attract increased academic attention. These analyses of Hot Topic Features lead to expansion of our research topics to analyze six topics: Cas9, CRISPR, Exosome, Microbiome, CAR-T and Zika. Then, authors' degree centralities are calculated, which is historically first and conceptually simplest among various kinds of network centralities, among these six topics. Furthermore, co-authorship centrality is introduced as a potential new feature, likewise based on the data retrieved from Web of Science Core Collection. Results demonstrate that authors in the top 10% of both centrality rankings are far more likely to be startup participants than others across all six topics. This shows both the centralities' potential usefulness as features to assess researchers' startup readiness.

By performing exploratory research in this chapter as described, potential features to assess startup readiness, which could be distinctively useful to academic researchers in the emerging research field with intense science linkage, will be developed.

Table 3.1 Top 30 Most Actively Financing Industry Fields Among 281 VentureSource Industry Codes/Subcodes Based on 17,681 Financing Deals During Jan. 1, 2017 Through Dec. 31, 2017

Rank	Industry Segment	Industry Code/ Subcode	Average Finance Size	Order	# of Rounds	Order
1	Travel and Leisure	Transportation Services	148.25	4	224	18
2	Financial Institutions & Services	Lending	46.31	25	359	5
3	Consumer Information Services	Shopping Facilitators	34.23	36	921	1
4	Business Support Services	Facilities / Operations Management	32.35	43	272	13
5	Consumer Information Services	Email / Messaging	58.61	16	97	52
6	Financial Institutions & Services	Insurance	33.5	38	133	36
7	Financial Institutions & Services	Retail Investment Services / Brokerages	63.95	15	81	60
8	Wholesale Trade and Shipping	Logistics / Delivery Services	29.72	51	166	29
9	**Biopharmaceuticals**	**Pharmaceuticals**	**32.28**	**44**	**129**	**38**
10	Financial Institutions & Services	Payment / Transactional Processing	20.6	86	313	10
11	Business Support Services	Data Management Services	18.5	94	339	6
12	**Biopharmaceuticals**	**Biotechnology Therapeutics**	**23.87**	**72**	**154**	**32**
13	Financial Institutions & Services	Real Estate	23.46	73	140	35
14	Electronics & Computer Hardware	Consumer Electronics	17.71	99	316	9
15	**Biopharmaceuticals**	**Immunotherapy / Vaccines**	**29.05**	**54**	**93**	**55**
16	Medical Devices & Equipment	Medical Lab Instruments / Test Kits	34.51	35	65	74
17	Machinery & Industrial Goods	General Industrial Goods	35.91	34	60	77
18	Vehicles and Parts	Automotive Parts	50.62	19	49	92
19	Retailers	Food / Drug Retailers	22.73	76	118	41
20	**Biopharmaceuticals**	**Small Molecule Therapeutics**	**25.72**	**67**	**104**	**51**
21	**Biopharmaceuticals**	**Gene Therapy**	**30.58**	**47**	**68**	**73**
22	Software	Security	15.12	115	336	7
23	Retailers	Vehicle Parts Retailers / Vehicle Dealers	49.57	23	45	100
24	Travel and Leisure	Travel Arrangement / Tourism	17.53	100	179	25
25	Consumer Information Services	Entertainment	15.85	110	264	16
26	Financial Institutions & Services	Institutional Investment Services	19.79	91	129	37
27	Vehicles & Parts	Automobiles	157.79	2	30	130
28	Business Support Services	Procurement / Supply Chain	16.15	108	160	30
29	Media and Content	Broadcasting	39.56	28	37	112
30	Electronics & Computer Hardware	Electronic Components / Devices	15.08	117	194	23

Note: Rank here is based on the sum of both orders.

3.1. Exploring Network Centralities as Potential Features

3.1.1. Construction of Author Citation Networks and Introducing Network Centralities for Authors

This step of methodology has become a starting point and a basis of this doctoral research. For this section, papers including the terms "CRISPR Cas9" in the title, abstract, or keywords are extracted from the Web of Science Core Collection database. Papers in the CRISPR-Cas9 field are targeted as a dataset, from which names of all authors and paper citation-related information are extracted.

From the extracted data, lists of pairs (edge lists) of cited-and-citing papers are created. Using author lists for the respective papers, edge lists of cited-and-citing authors including all co-authors can be constructed, irrespective of order, for all pairs of cited-and-citing papers. For example, in a case where paper A with five authors is cited by paper B with three authors, we can create 15 (5 × 3) pairs as edge lists, to build up a whole new author citation network based on them. For duplicate authors both in cited and citing papers, the pairs of duplicates need to be eliminated in order to produce pairs comprising different authors only. The above method is applied to all pairs of cited-and-citing papers.

Based on edge lists of cited-and-citing authors, new author citation networks are created for each year during 2012–2016, with annual cumulative authors as nodes and with annual cumulative citation relations as edges (links) of these networks. From the created author citation networks, each author's centrality is calculated and arranged in descending order for each year (Appendix A-1). Network centrality represents how central each author is in terms of the position in the author citation network. The degree of centrality can be ascertained using several methods [41, 42, 43, 44, 45, 46]. The following are the five centralities used for this study: betweenness centrality, closeness centrality, degree centrality, eigenvector centrality and PageRank, all of which are non-directional centralities that are used commonly and frequently. They are mutually complementary in citation network analysis. For comparison, each year's lists of authors, with their number of citations to their published papers, are also created in descending order (Appendix A-2).

Therefore, papers that included the terms "CRISPR Cas9" in the title, abstract or keywords, were extracted from the Web of Science Core Collection database for (1) January 1, 2012 – December 31, 2012, (2) January 1, 2012 – December 31, 2013, (3) January 1, 2012 – December 31, 2014, (4) January 1, 2012 – December 31, 2015, and (5) January 1, 2012 – December 31, 2016. They produced Table 3.2 for the number of papers (facets), nodes (cited or citing papers), and edges (citations or links) for each

year's dataset. It is observed that the CRISPR-Cas9 research field rapidly emerged over the past few years.

From the extracted data, using author lists for each paper, edge lists of cited-and-citing authors are constructed, comprehensively including all co-authors, irrespective of order, for all pairs of cited-and-citing papers. Then new author citation networks were then extracted for each year for 2012–2016, with annual cumulative authors as network nodes and with annual cumulative citation relations as network edges (Table 3.3).

From the created author citation networks that included node and edge information for authors, a series of centrality calculation results are retrieved and arranged in descending order for each year. For this experiment, the five centralities are used, which are described previously to create lists of the top 100 authors for each centrality in descending order.

Table 3.4 shows lists of the top 10 authors for centralities for 2012–2016. They were created using this procedure. For comparison purposes, the lists of top 10 authors each year, with numerous citations of their published papers, are retrieved in descending order from the Web of Science Core Collection database, as shown in Table 3.5.

Table 3.2 Year-to-Year Paper Dataset and Citation Networks Queried using "CRISPR Cas9" (2012-16)

	2012	2013	2014	2015	2016
Facet (Paper) Count	4	89	485	1294	2775
Node Count	3	86	418	1080	2282
Edge Count	2	722	4939	13391	30274
Facet Growth	n/a	2225%	545%	267%	214%
Node Growth	n/a	2867%	486%	258%	211%
Edge Growth	n/a	36100%	684%	271%	226%

Table 3.3 Year-to-Year Author Dataset and Citation Networks Queried Using "CRISPR Cas9" (2012-16)

	2012	2013	2014	2015	2016
Facet (Author) Count	29	423	2449	7363	16250
Node Count	12	388	1908	5280	11155
Edge Count	50	22108	145190	418155	950838
Facet Growth	n/a	1459%	579%	301%	221%
Node Growth	n/a	3233%	492%	277%	211%
Edge Growth	n/a	44216%	657%	288%	227%

Table 3.4 Top 10 Authors, Ranked by Number of Five Centralities (2012-2016)

(2012)

Rank	Betweenness Centrality	Closeness Centrality	Degree Centrality	Eigenvector Centrality	Pagerank
1	Barrangou, Rodolphe	Barrangou, Rodolphe	Barrangou, Rodolphe	Barrangou, Rodolphe	Barrangou, Rodolphe
1	Gasiunas, Giedrius	Gasiunas, Giedrius	Gasiunas, Giedrius	Gasiunas, Giedrius	Gasiunas, Giedrius
1	Horvath, Philippe	Horvath, Philippe	Horvath, Philippe	Horvath, Philippe	Horvath, Philippe
1	Siksnys, Virginijus	Siksnys, Virginijus	Siksnys, Virginijus	Siksnys, Virginijus	Siksnys, Virginijus
5	Fremaux, Christophe	Fremaux, Christophe	Fremaux, Christophe	Fremaux, Christophe	Fremaux, Christophe
5	Sapranauskas, Rimantas	Sapranauskas, Rimantas	Sapranauskas, Rimantas	Sapranauskas, Rimantas	Sapranauskas, Rimantas
7	Charpentier, Emmanuelle	Charpentier, Emmanuelle	Charpentier, Emmanuelle	Charpentier, Emmanuelle	Charpentier, Emmanuelle
7	Chylinski, Krzysztof	Chylinski, Krzysztof	Chylinski, Krzysztof	Chylinski, Krzysztof	Chylinski, Krzysztof
7	Doudna, Jennifer A.	Doudna, Jennifer A.	Doudna, Jennifer A.	Doudna, Jennifer A.	Doudna, Jennifer A.
7	Fonfara, Ines	Fonfara, Ines	Fonfara, Ines	Fonfara, Ines	Fonfara, Ines
7	Hauer, Michael	Hauer, Michael	Hauer, Michael	Hauer, Michael	Hauer, Michael
7	Jinek, Martin	Jinek, Martin	Jinek, Martin	Jinek, Martin	Jinek, Martin

(2013)

Ranking	Betweenness Centrality	Closeness Centrality	Degree Centrality	Eigenvector Centrality	Pagerank
1	Zhang, Feng	Zhang, Feng	Zhang, Feng	Zhang, Feng	Zhang, Feng
2	Church, George M.	Church, George M.	Church, George M.	Church, George M.	Church, George M.
3	Doudna, Jennifer A.	Hsu, Patrick D.	Hsu, Patrick D.	Hsu, Patrick D.	Esvelt, Kevin M.
4	Esvelt, Kevin M.	Esvelt, Kevin M.	Esvelt, Kevin M.	Esvelt, Kevin M.	Hsu, Patrick D.
5	Hsu, Patrick D.	Ran, F. Ann	Ran, F. Ann	Ran, F. Ann	Ran, F. Ann
6	Ran, F. Ann	Mali, Prashant	Mali, Prashant	Mali, Prashant	Mali, Prashant
7	Mali, Prashant	Aach, John	Aach, John	Marraffini, Luciano A.	Aach, John
8	Aach, John	Marraffini, Luciano A.	Marraffini, Luciano A.	Aach, John	Marraffini, Luciano A.
9	Norville, Julie E.	Wu, Xuebing	Wu, Xuebing	Wu, Xuebing	Norville, Julie E.
10	Marraffini, Luciano A.	Norville, Julie E.	Norville, Julie E.	Norville, Julie E.	Wu, Xuebing

(2014)

Rank	Betweenness Centrality	Closeness Centrality	Degree Centrality	Eigenvector Centrality	Pagerank
1	Zhang, Feng	Zhang, Feng	Zhang, Feng	Zhang, Feng	Zhang, Feng
2	Hsu, Patrick D.	Hsu, Patrick D.	Hsu, Patrick D.	Hsu, Patrick D.	Hsu, Patrick D.
3	Ran, F. Ann	Ran, F. Ann	Ran, F. Ann	Ran, F. Ann	Ran, F. Ann
4	Church, George M.	Marraffini, Luciano A.	Marraffini, Luciano A.	Marraffini, Luciano A.	Marraffini, Luciano A.
5	Esvelt, Kevin M.	Wu, Xuebing	Wu, Xuebing	Wu, Xuebing	Church, George M.
6	Marraffini, Luciano A.	Church, George M.	Church, George M.	Church, George M.	Wu, Xuebing
7	Doudna, Jennifer A.	Jiang, Wenyan	Jiang, Wenyan	Jiang, Wenyan	Esvelt, Kevin M.
8	Wu, Xuebing	Esvelt, Kevin M.	Esvelt, Kevin M.	Esvelt, Kevin M.	Jiang, Wenyan
9	Scott, David A.	Cox, David	Cox, David	Cong, Le	Mali, Prashant
10	Mali, Prashant	Lin, Shuailiang	Lin, Shuailiang	Mali, Prashant	Cox, David

(2015)

Ranking	Betweenness Centrality	Closeness Centrality	Degree Centrality	Eigenvector Centrality	Pagerank
1	Church, George M.	Church, George M.	Church, George M.	Church, George M.	Church, George M.
2	Esvelt, Kevin M.	Esvelt, Kevin M.	Esvelt, Kevin M.	Esvelt, Kevin M.	Esvelt, Kevin M.
3	Jinek, Martin	Aach, John	Jinek, Martin	Aach, John	Aach, John
4	Aach, John	Norville, Julie E.	Aach, John	Norville, Julie E.	Norville, Julie E.
5	Norville, Julie E.	Mali, Prashant	Norville, Julie E.	Mali, Prashant	Mali, Prashant
6	Zhang, Feng	Yang, Luhan	Zhang, Feng	Yang, Luhan	Yang, Luhan
7	Mali, Prashant	Zhang, Feng	Mali, Prashant	Zhang, Feng	Zhang, Feng
8	Yang, Luhan	DiCarlo, James E.	Yang, Luhan	DiCarlo, James E.	DiCarlo, James E.
9	DiCarlo, James E.	Hsu, Patrick D.	DiCarlo, James E.	Hsu, Patrick D.	Hsu, Patrick D.
10	Guell, Marc	Jiang, Wenyan	Guell, Marc	Jiang, Wenyan	Jiang, Wenyan

(2016)

Rank	Betweenness Centrality	Closeness Centrality	Degree Centrality	Eigenvector Centrality	Pagerank
1	Zhang, Feng	Zhang, Feng	Zhang, Feng	Zhang, Feng	Zhang, Feng
2	Hsu, Patrick D.	Hsu, Patrick D.	Hsu, Patrick D.	Hsu, Patrick D.	Hsu, Patrick D.
3	Ran, F. Ann	Ran, F. Ann	Ran, F. Ann	Ran, F. Ann	Ran, F. Ann
4	Scott, David A.	Marraffini, Luciano A.	Marraffini, Luciano A.	Wu, Xuebing	Scott, David A.
5	Church, George M.	Wu, Xuebing	Wu, Xuebing	Marraffini, Luciano A.	Marraffini, Luciano A.
6	Doudna, Jennifer A.	Church, George M.	Church, George M.	Church, George M.	Wu, Xuebing
7	Esvelt, Kevin M.	Scott, David A.	Cong, Le	Cong, Le	Church, George M.
8	Marraffini, Luciano A.	Cong, Le	Scott, David A.	Jiang, Wenyan	Jiang, Wenyan
9	Wu, Xuebing	Jiang, Wenyan	Jiang, Wenyan	Habib, Naomi	Cong, Le
10	Mali, Prashant	Doudna, Jennifer A.	Doudna, Jennifer A.	Mali, Prashant	Esvelt, Kevin M.

Table 3.5 Top 10 Authors, Ranked by Number of Citations (2012-2016)

2012			2013			2014			2015			2016		
Ranking	Author	Number of Citations	Ranking	Author	Number of Citations	Ranking	Author	Number of Citations	Ranking	Author	Number of Citations	Ranking	Author	Number of Citations
1	Gasiunas, Giedrius	1	1	Church, George M.	76	1	Zhang, Feng	234	1	Zhang, Feng	454	1	Doudna, Jennifer A.	636
1	Barrangou, Rodolphe	1	2	Zhang, Feng	63	2	Charpentier, Emmanuelle	171	2	Doudna, Jennifer A.	307	2	Zhang, Feng	565
1	Horvath, Philippe	1	3	Esvelt, Kevin M.	49	3	Huang, Xingxu	168	3	Gersbach, Charles A.	304	3	Qi, Lei S.	554
1	Siksnys, Virginijus	1	4	Mali, Prashant	47	4	Doudna, Jennifer A.	160	4	Kim, Jin-Soo	251	4	Gersbach, Charles A.	535
1	Jinek, Martin	1	5	Hsu, Patrick D.	42	5	Church, George M.	156	5	Church, George M.	223	5	Kim, Jin-Soo	516
1	Chylinski, Krzysztof	1	5	Ran, F. Ann	42	6	O'Connor-Giles, Kate M.	141	6	Charpentier, Emmanuelle	217	6	Barrangou, Rodolphe	351
1	Fonfara, Ines	1	7	Scott, David A.	39	7	Wildonger, Jill	123	7	Huang, Xingxu	203	7	Joung, J. Keith	333
1	Hauer, Michael	1	8	Siksnys, Virginijus	38	7	Harrison, Melissa M.	123	8	Joung, J. Keith	199	8	Hsu, Patrick D.	276
1	Doudna, Jennifer A.	1	8	Gasiunas, Giedrius	38	9	Kim, Jin-Soo	120	9	Hsu, Patrick D.	164	8	Yamamoto, Takashi	276
1	Charpentier, Emmanuelle	1	10	Marraffini, Luciano A.	29	10	Joung, J. Keith	118	10	Sakuma, Tetsushi	161	10	Lim, Wendell A.	274
n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	10	Yamamoto, Takashi	161	n/a	n/a	n/a

3.1.2. Detection of Founders Among Authors with High Centralities

This second step also became a basic process throughout this doctoral research: surveying VentureSource to check status of authors or startups (See Figure 3-1). VentureSource is used to check whether the authors are founders of specific startups herein. As can be seen later, this methodology can be widely used to check various startup statuses ranging from academic researchers who are startup participants, to startups that achieved an IPO or an M&A. This database even includes information regarding dates of inception, financings, IPO's, M&A's, amount of raised capital and so on.

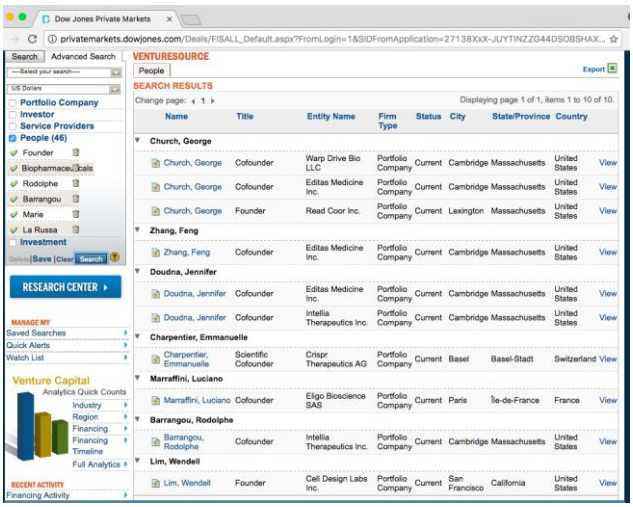


Figure 3-1 VentureSource That Querried Researchers Who Are Founders of Startups

From the year-to-year author lists with the five calculated centralities and the numbers of citations in descending order for 2012–2016 as described previously (Appendix A-1), queries of top authors are conducted in VentureSource to ascertain whether the authors are founders of their research-based startups. Then, after creating heat maps that display whether each author is a founder or not (Appendix A-1), this section observed how each author's centrality relates to the rate of

being a founder, and how widely these centralities cover those scientific founders. In addition to the author centrality ranking lists, for comparison purposes, lists of each year's top 100 cited authors are created in descending order, with similarly formatted heat maps (Appendix A-2).

Using the process described above, comparative studies are conducted related to the rate and the coverage of founders among the top authors based on the five centralities versus the number of citations. For this purpose, for 2012–2016, top 100 authors out of each centrality list were grouped into ten-rank orders, compared to the number of citations (Table 3.4). Then, the rates of founders per group are calculated. Furthermore, this section calculated the coverage of founders per unique author included in the top 100 authors in the combined centralities, compared to the coverage of the top 100 authors based on the number of citations (Table 3.6).

Using year-to-year author lists with calculated centralities in descending order for 2012–2016, along with the numbers of citations, top 100 authors of each of the centralities and the number of citations were grouped into ten-rank orders. Then, queries for them are executed in VentureSource to ascertain whether they are founders of startups in their respective research fields. The founder rate per group is also calculated. Moreover, another calculation conducted was the founder coverage for each unique author included in the top 100 authors in the combined centralities of this research, compared with the founder coverage of the top 100 cited authors.

Table 3.6 shows that, for the combined five centralities and the number of citations, higher ranking authors tended to have higher rates of founders, although the results from the number of citations tended to have higher founder rates in higher rankings in later years, than those of the combined centralities. Attached as Appendix A-1 and A-2 are the heat maps of top 100 authors highlighting founders, ranked by the five centralities combined and the number of citations respectively.

Results show that founders were more dispersedly detected in the top 100 authors in each of the combined centralities throughout 2013-2016, than those in the top 100 cited authors where founders existed only in top 30-40 in 2015 and 2016. More interestingly, results suggest that founder coverage of unique authors out of the top 100 authors in the combined centralities was much higher than those in the top 100 cited authors, except for the beginning of year 2012, with only few authors. The ratios between founder coverage of unique authors out of the top 100 authors in the five centralities combined and that of the top 100 cited authors were 1.468 in 2013, 1.140 in 2014, 1.563 in 2015 and 1.377 in 2016. It is noteworthy that, even in 2013 when the author citation network was just emerging, the combined centralities enabled us to identify a wider coverage of founders.

Table 3.6 Founder Rate and Coverage, Measured by Five Centralities Combined and Number of Citations (2013-2016)

	2013		2014		2015		2016	
	Five Centralities Combined	Number of Citations	Five Centralities Combined	Number of Citations	Five Centralities Combined	Number of Citations	Five Centralities Combined	Number of Citations
Founder Rate: 1–10	0.280	0.300	0.320	0.500	0.300	0.500	0.360	0.500
11–20	0.060	0.100	0.100	0.000	0.100	0.200	0.080	0.200
21–30	0.220	0.000	0.080	0.000	0.100	0.100	0.180	0.100
31–40	0.040	0.000	0.100	0.000	0.120	0.000	0.060	0.100
41–50	0.060	0.000	0.000	0.000	0.040	0.000	0.020	0.000
51–60	0.000	0.100	0.000	0.000	0.140	0.000	0.100	0.000
61–70	0.040	0.000	0.180	0.200	0.060	0.000	0.080	0.000
71–80	0.020	0.000	0.140	0.200	0.080	0.000	0.160	0.000
81–90	0.040	0.000	0.140	0.000	0.200	0.000	0.180	0.000
91–100	0.000	0.000	0.040	0.000	0.260	0.000	0.100	0.000
Total Number of Unique Founders (A): 1–100	8	5	12	9	15	8	15	9
Total Number of Unique Authors (B): 1–100	109	100	117	100	120	100	121	100
Founder Coverage (A/B)	0.073	0.050	0.103	0.090	0.125	0.080	0.124	0.090
Ratio regarding Total Number of Unique Authors (Five Centralities Combined / Number of Citations)	1.090		1.170		1.200		1.210	
Ratio regarding Founder Coverage (Five Centralities Combined / Number of Citations)	1.468		1.140		1.563		1.377	

3.2. Introducing Hot Topic Factors and Co-authorship Centrality as Potential Features

Figure 3-2 Methodology Proposed in 3.2

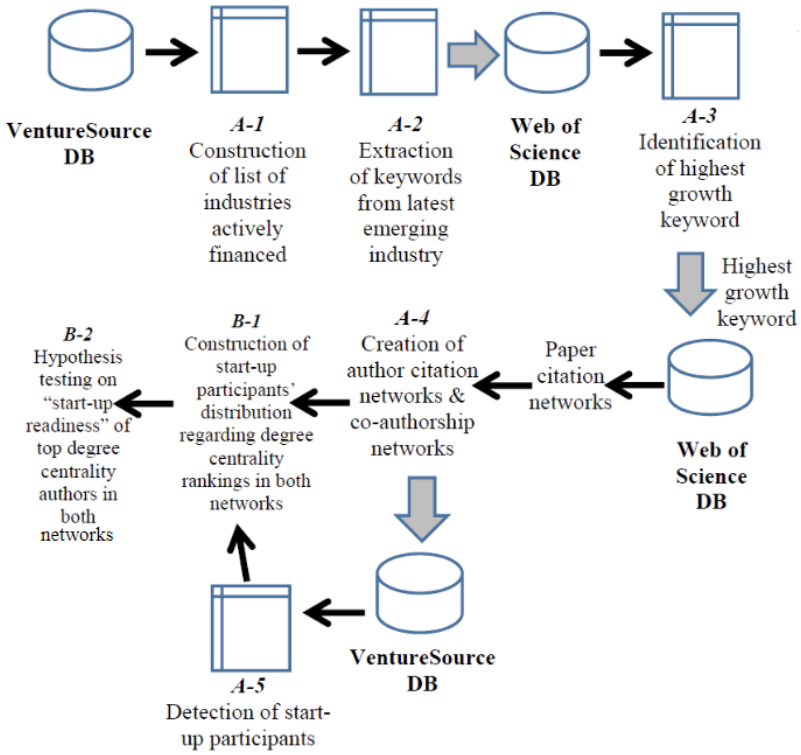


Figure 3-2 depicts the analytical scheme employed herein.

To begin with, using VentureSource, all financing deals between January 1, 2017 and December 31, 2017 are analyzed, to construct a list of industry segments that were most active in venture capital finance in 2017 (A-1). Using VentureSource, this thesis compiled a ranked list of top 30 most actively funded industry codes/subcodes among all 281 VentureSource industry codes/subcodes in 2017, based on the sum of the ranks of both the average financing size and the number of financing rounds (A-1). Then industry codes/subcodes belonging to VentureSource industry segment titled "Biopharmaceutical" is extracted.

For startups in the target codes/subcodes belonging to the biopharmaceutical segment on VentureSource, keywords that appear multiple times are surveyed (A-2). Then, differences in the keyword frequency are analyzed in research papers between 2014 and 2017 by searching the Web of Science Core Collection database, compiled by Clarivate Analytics. Through this process, keywords that have seen the highest growth (highest-growth keyword) in the above period for each target code/subcode, are identified, as emerging research topics (A-3).

Then, not only author citation networks but also co-authorship networks are created based on the research papers for each highest-growth keyword (A-4). Simultaneously, authors who are participants of startups relative to those keywords among the both networks are detected (A-5).

Finally, the distribution of the startup participant authors in the author citation networks as well as the co-authorship networks is analyzed in terms of degree centrality (B-1), partly as in 3.1.1, and then hypothesis testing of the top 10% degree centrality authors of both networks is conducted, relative to their startup readiness (B-2).

3.2.1. Introduction of Hot Topic Features

(i) Finding Most Actively Financed Industry Fields

Using the VentureSource database, lists of most actively financed industry fields were constructed based on the average financing size and the number of financing deals per field during the 365 days of 2017. According to VentureSource, 17,681 financing deals took place in 2017. Then all of their 281 industry codes/sub-codes were sorted in descending order both by the average financing deal size and the number of financing rounds. This thesis then constructed the top 30 (approximately top 10%) ranking of most actively financed industry fields, which were arranged based on the sum of the orders of both the average financing size and the number of financing rounds. The number of financing rounds in addition to the average financing size was examined, because this paper primarily addresses venture capital financing where the number of rounds of financing in seed or early stage companies is important, whereas private equity investment deals typically have a few large investment rounds in middle or later stage companies (Table 3.1).

From the top 30 industry fields above, industry codes/subcodes that belong to industry segment “Biopharmaceuticals” were extracted, as case studies to explore as a part of this survey. Particularly, five industry codes/subcodes as follows were selected: “Pharmaceuticals,” “Biotechnology Therapeutics,” “Immunotherapy / Vaccines,” “Small Molecule Therapeutics” and “Gene Therapy” for an additional survey shown below.

(ii) Identifying Highest Growth Keywords Related to Most Actively Financed Biopharmaceutical Industry Fields

In the VentureSource database, keywords representing company overviews are available. Using VentureSource, keywords were surveyed for the startups, which falls into the above five industry codes/subcodes. Table 3.7 below presents such keywords that appeared twice or more on each code/subcode.

Table 3.7 VentureSource Keywords That Appeared Twice or More for Startups Relative to Actively Financing Biopharmaceutical Industry Fields in 2017

(…Continued on Next Page)

Pharmaceuticals	
Appearance Frequency	Keywords (Appearance Frequency)
5 or Greater	drug*'(20), 'medic*'(12), 'cancer*'(12), 'therap*'(8), 'pharma*'(8), 'health*'(8), 'pharm*'(8), 'nan'(7), 'pain*'(5), 'vitamin*'(5), 'tumor*'(5), 'marijuana*'(5), 'cannabis*'(5), 'disease*'(5), 'pharmaceutic*'(5), 'protein*'(5)
3 to 4	vaccine*'(4), 'supplement*'(4), 'skin*'(4), 'oncolog*'(4), 'antibiotic*'(4), 'nutrition*'(4), 'capsule*'(4), 'intermediate*'(4), 'biotech*'(3), 'rare disease*'(3), 'bacteria*'(3), 'cardi*'(3), 'diabetes'(3), 'plant*'(3), 'treatment*'(3), 'oncology*'(3), 'therapeutic*'(3), 'child*'(3), 'cannabis'(3), 'medicine*'(3), 'treat*'(3), 'API*'(3), 'enzyme*'(3), 'immun*'(3), 'blood*'(3), 'tablet*'(3)
2	osteo*', 'OA', 'herpes', 'probiotic*', 'intestin*', 'digest*', 'molecule*', 'small molecul*', 'chemother*', 'immunoth*', 'opioid*', 'analges*', 'neuro*', 'molecul*', 'brain*', '*gum*', '*candy*', '*infect*', 'gynecolog*', 'clinic*', 'antimicrob*', 'bacteri*', 'broad-spectrum', 'cardio*', 'develop*', 'allerg*', 'neurolog*', 'cancer', 'respiratory*', 'asthma*', 'immune*', 'antibod*', 'obes*', 'compound*', 'antibiot*', 'bacteria', 'super*bug*', 'injection*', 'prescription', 'health', 'autoimmune*', 'Chinese medicine', 'herb*', 'diagnos*', 'exosome*', 'cosme*', 'tissue*', 'wrinkle*', 'API', 'glaucoma*', 'infect*', 'nutraceutic*'

Small Molecule Therapeutics	
Appearance Frequency	Keywords (Appearance Frequency)
5 or Greater	drug*'(23), 'cancer*'(22), 'therap*'(20), 'molecule*'(15), 'disease*'(9), 'immun*'(8), 'neuro*'(8), 'health*'(8), 'cancer'(8), 'small molecule*'(7), 'tumor*'(6), 'treat*'(6), 'protein*'(6), 'antibod*'(6), 'cell*'(5), 'infect*'(5), 'molecul*'(5), 'inflam*'(5), 'treatment*'(5), 'cardio*'(5), 'small molecule'(5), 'oncolog*'(5)
3 to 4	therapeutic*'(4), 'medic*'(4), 'pathogen*'(4), 'chronic*'(4), 'Alzheimer*'(4), 'biotech*'(3), 'immuno*'(3), 'food'(3), 'ion channel'(3), 'inhibit*'(3), 'nano*'(3), 'immune*'(3), 'C5a'(3), 'terminal*'(3), '*infect*'(3), 'therapeut*'(3), 'bio*'(3), 'kinase*'(3), 'tumor'(3), '*onco*'(3), 'diagnos*'(3), 'PSVT'(3), 'calcium*'(3), '*ventricul*'(3), 'brain*'(3), 'neurodegenerat*'(3)
2	biopharma*', 'affinity purification', 'bioprocess*', 'onco*', 'AI', 'substitut*', 'plant*', 'milk', 'cheese*', 'mayonnaise*', 'yogurt*', 'emul*', 'mitochondrial', 'eosinophil', 'leukocyte', '*immun*', 'skin*', 'spray*', 'resist*', 'bacteria*', 'antibiotic', 'AF', 'heart*', 'peptide*', '*cell*', 'derma*', 'nerv*', 'molecule', 'oncolo*', 'rare disease*', 'ossific*', 'fibrodysplasia', '(RAR)', 'integrin', 'antibiotic*', 'breast*', '*cancer', 'treatment', 'neurolog*', 'protein', 'lung', 'apoptosis', 'respiratory*', 'therapeutic', '*cancer*', 'fibrosis', 'GSK-3ß', 'Glycogen Synthase Kinase', 'bipolar disorder', 'diabetes'

Biotechnology Therapeutics	
Appearance Frequency	Keywords (Appearance Frequency)
5 or Greater	cancer*'(28), 'therap*'(20), 'drug*'(16), 'disease*'(12), 'biotech*'(11), 'cell*'(9), 'health*'(8), 'diseas*'(7), 'disorder*'(7), 'diagnos*'(7), 'gene*'(7), 'tumor*'(7), 'medic*'(6), 'bacteria*'(6), 'neuro*'(6), 'DNA'(6), 'oncolog*'(6), 'Alzheimer*'(5), 'cleantech'(5), 'treat*'(5), 'cancer'(5), 'tissue*'(5)
3 to 4	molecule*'(4), 'immune*'(4), 'therapeut*'(4), 'antibod*'(4), 'therapeutic*'(4), 'industry focused products and services'(4), 'infect*'(4), 'regenerat*'(4), 'protein*'(4), 'microbiome'(4), 'treatment*'(4), 'immun*'(4), 'neurolog*'(3), 'Alzheimer's*'(3), 'spinal cord*'(3), 'injur*'(3), '*amyloid*'(3), 'receptor'(3), 'central nervous system*'(3), 'CNS'(3), 'antibiotic*'(3), 'compound*'(3), 'onco*'(3), 'CMBC'(3), 'pharma*'(3), 'skin*'(3), 'molecule'(3), 'vaccine*'(3), 'inhibitor*'(3), 'inflam*'(3), 'respirat*'(3), '*cancer*'(3), 'bone*'(3), 'pharm*'(3)
2	blood*', 'platelet*', 'manufact*', 'contract*', 'life scienc*,', 'metaboli*', 'inflammat*', 'microbe*', 'inhal*', 'cardiovascular', 'PAF', 'biotech* atria*', 'arrhythmia', 'paediatr*', 'nephrolog*', 'renal*',

(…Continued from Previous Page)

	'neurologic*', 'orphan*', 'anaesthes*', 'tubulo*', 'oil', 'protein', 'insect*', 'agriculture', 'lysom*', 'stor*', 'HSP', 'misfold*', 'degenerat*', 'hear*', 'cochlear*', 'ear*', 'noise', 'restor*', 'cardiovascular*', 'myelofibrosis', 'MF', 'JAK2', 'ocular*', 'probiotic*', 'research*', 'clinical trial*', 'osteoporos*', 'hypoparathy*', 'hypoparathyroidism*', 'biopharm*', 'drug discover*', 'genomic*', '*microbiome*', 'th+F5eepeutic*', 'urea cycle disorder*', 'pathogenic*', '*bacteria*', '*health*', 'fish*', '*inflamation+F23*', 'detect*', 'vir*', 'T-cell*', '*immune*', 'dermatolog*', 'aesthetics', 'plasmotic', 'acne', 'hair removal', 'topical', 'vascular', 'COPD', 'addict*', 'opiate*', 'alcohol', 'abuse*', 'silk*', 'aesthetic*', 'defect*', 'diabetes', 'obes*', 'molecular*', 'rare disease*', '*skelet*', 'drug', 'molecul*', 'metabolic', 'intestin*', 'antibody'

Gene Therapy	
Appearance Frequency	Keywords (Appearance Frequency)
5 or Greater	gene*'(29), 'therap*'(19), 'cancer*'(16), 'DNA'(10), 'drug*'(7), 'genomic*'(5), 'cancer'(5), 'health*'(5), 'oncolog*'(5), 'cell*'(5)
3 to 4	DNA*'(4), 'genom*'(4), 'patient*'(4), 'immun*'(4), 'CRISPR'(4), 'gene therap*'(4), 'cure'(4), 'therapy'(4), 'gen*'(4), 'disease*'(4), 'biotech*'(4), 'virus'(3), 'gene'(3), 'AAV'(3), 'virus*'(3), 'diseas*'(3), 'immune*'(3), 'brain'(3)
2	Parkinson*', 'DTC genetic test*', 'therapeutics*', 'genetic research*', 'medic*', 'adeno*', 'treat*', 'Cas9', 'duchenne', 'dystroph*', '8muscular disease*', 'viral*', 'tumor*', 'research*', 'animal*', 'livestock', 'agricultur*', 'breed*', '*medic*', 'retina*', 'dystrop*', 'choroid*', 'degenerat*', 'CHM', 'REP-1', 'protein*', 'rare disorder*', 'treatment*', 'adeno', 'medical', 'treatment', 'health', 'stem cell*', 'CAR-T', 'HIV', '*gene*', 'chimer*', 'receptor*', '*cell*', 'CAR', 'life scien*', 'genetic engineer*', 'personal* medic*', 'molec* bio*', 'HCP', 'cell protein', 'genome*', 'glioblastoma'

Immunotherapy / Vaccines	
Appearance Frequency	Keywords (Appearance Frequency)
5 or Greater	'cancer*'(31), 'immun*'(22), 'tumor*'(15), 'therap*'(14), 'antibod*'(13), 'disease*'(13), 'vaccine*'(12), 'infect*'(12), 'cancer'(11), 'drug*'(11), 'vaccin*'(10), 'virus*'(8), '*immun*'(8), 'antigen*'(7), 'T cell*'(6), 'cell*'(6), '*therap*'(5), 'protein*'(5)
3 to 4	'oncolog*'(4), 'allerg*'(4), '*cancer*'(4), 'immuno*'(4), 'disease'(3), 'biotech*'(3), 'inflammat*'(3), 'viral*'(3), '*tumor*'(3), 'antibody*'(3), 'ADC'(3), 'medicine*'(3), 'ag*'(3), 'immunosenescence'(3), '*infect*'(3), 'antibiotic*'(3), 'inflam*'(3), 'medic*'(3), 'prevent*'(3), 'RSV'(3), 'target*'(3), 'T-cell'(3)
2	'therapeutic*', 'oncology', 'RNA', 'DNA', 'HIV', 'TME', 'patient', 'pathogen*', 'bacteria*', 'monoclonal', 'complement system', 'COPD', 'AMD', 'PNH', 'virus', 'treatment', 'patho*', '*oncolog*', 'respirat*', 'licens*', 'rhinovirus', 'cold*', 'asthma*', '*vir*', 'mosquito*', 'Zika', 'Dengue*', '*fever*', 'Hepati*', 'nanomedicine*', 'colorectal*', 'purif*', 'oncol*', 'viral', 'dendritic', 'biopharma*', 'diseas*', 'oncobio*', 'microb*', 'tumor', 'antigen'

Then, queries of all keywords in Table 3.7 are made on the Web of Science Core Collection database to analyze the growth in the frequency of those keywords that appeared in the title, keywords, or Keyword Plus of the research papers during 2014–2017. Rankings of keywords showing the most growth in incidence for each industry code/subcode were constructed based on growth multiples (Table 3.8).

Table 3.8 Keyword Frequency Growth Multiple in Web of Science Core Collection Related to Active Financing in Biopharmaceutical Industry Fields during 2014 through 2017

Pharmaceuticals

Rank	Keyword	Start Year Count	End Year Count	Growth (×)
1	therap	1	6	6
2	cardi	2	7	3.5
3	super*bug	9	29	3.22
4	exosome	400	880	2.2
5	cosme	4	8	2
6	immun	7	14	2
7	allerg	1	2	2
8	antibod	1	2	2
9	obes	23	44	1.91
10	marijuana	993	1431	1.44
11	opioid	3504	4910	1.4
12	candy	136	189	1.39
13	rare disease	8249	11439	1.39
14	cardio	1022	1409	1.38
15	probiotic	1551	2129	1.37
16	cannabis	1364	1866	1.37
17	antibiotic	15218	20716	1.36
18	health	124136	166047	1.34
19	medicine	31306	41035	1.31
20	clinic	13857	18097	1.31

Gene Therapy

Rank	Keyword	Start Year Count	End Year Count	Growth (×)
1	therap	1	6	6
2	Cas9	448	2384	5.32
3	CAR-T	97	485	5
4	CRISPR	694	3138	4.52
5	DTC genetic test	13	29	2.23
6	immun	7	14	2
7	diseas	4	8	2
8	CHM	104	152	1.46
9	AAV	566	813	1.44
10	molec* bio	1335	1829	1.37
11	rare disorder	3304	4444	1.35
12	health	124136	166231	1.34
13	medical	53946	68133	1.26
14	therapeutics	5456	6848	1.26
15	livestock	3309	4147	1.25
16	HCP	687	848	1.23
17	research	195874	241726	1.23
18	glioblastoma	3961	4858	1.23
19	treatment	227193	278033	1.22
20	therapy	117393	142505	1.21

Biotechnology Therapeutics

Rank	Keyword	Start Year Count	End Year Count	Growth (×)
1	therap	1	6	6
2	inflam	1	3	3
3	microbiome	1732	4403	2.54
4	immun	7	14	2
5	injur	1	2	2
6	inflammat	1	2	2
7	diseas	4	8	2
8	obes	23	44	1.91
9	restor	23	35	1.52
10	paediatr	8	12	1.5
11	aesthetic	3128	4648	1.49
12	tubulo	37	54	1.46
13	onco	126	178	1.41
14	drug discover	1177	1650	1.4
15	aesthetics	1761	2461	1.4
16	microbe	3381	4719	1.4
17	agriculture	8226	11458	1.39
18	rare disease	8249	11439	1.39
19	MF	1323	1819	1.37
20	probiotic	1551	2129	1.37

Small Molecule Therapeutics

Rank	Keyword	Start Year Count	End Year Count	Growth (×)
1	therap	1	6	6
2	inflam	1	3	3
3	immun	7	14	2
4	ventricul	1	2	2
5	PSVT	9	14	1.56
6	mayonnaise	41	60	1.46
7	yogurt	396	576	1.45
8	ossific	7	10	1.43
9	onco	126	178	1.41
10	rare disease	8249	11429	1.39
11	cardio	1022	1409	1.38
12	antibiotic	15218	20690	1.36
13	health	124135	165905	1.34
14	immuno	1186	1572	1.33
15	bio	9870	12901	1.31
16	derma	20	26	1.3
17	AI	1992	2570	1.29
18	food	42265	54414	1.29
19	therapeutic	51320	65065	1.27
20	nano	16638	21015	1.26

Immunotherapy / Vaccines

Rank	Keyword	Start Year Count	End Year Count	Growth (×)
1	Zika	26	2310	88.85
2	therap	1	6	6
3	inflam	1	3	3
4	allerg	1	2	2
5	immun	7	14	2
6	antibod	1	2	2
7	diseas	4	8	2
8	inflammat	1	2	2
9	TME	192	365	1.9
10	Dengue	1775	2729	1.54
11	nanomedicine	1012	1474	1.46
12	antibiotic	15218	20716	1.36
13	vaccin	3	4	1.33
14	immuno	1186	1572	1.33
15	mosquito	2481	3254	1.31
16	medicine	31306	41035	1.31
17	oncology	8337	10862	1.3
18	therapeutic	51320	65117	1.27
19	treatment	227193	277723	1.22
20	colorectal	14257	17225	1.21

From the rankings presented above, for an additional survey, this thesis extracted keywords with incidence that more than doubled between 2014 (Start Year) and 2017 (End Year) and that appeared more than 100 times in 2017 (End Year). Keywords that

met this criterion from each industry code/subcode were the following five keywords, all of which are emerging research topics that have attracted academic attention to a rapidly increasing degree.

Pharmaceuticals: *Exosome*

Exosomes are small microvesicles that are released from late endosomal compartments of cultured cells, in many and perhaps all eukaryotic fluids, including blood and urine [47] [48]. Exosomes are either released from the cell when multivesicular bodies fuse with the plasma membrane or are released directly from the plasma membrane [49]. Exosomes have specialized functions and they play an important role in processes such as coagulation, intercellular signaling, and waste management [47]. Consequently, growing interest has arisen in their clinical applications. Exosomes might be used for therapy and prognosis, or as biomarkers for health and disease.

Biotechnology Therapeutics: *Microbiome*

Microbiome refers to ecological communities of commensal, symbiotic and pathogenic microorganisms [50] [51] found in and on all multicellular organisms from plants to animals. It describes either the collective genomes of the microorganisms that reside in an environmental niche or the microorganisms themselves [52] [53] [54]. The microbiome can promote or disrupt human health by influencing both adaptive and innate immune functions [55].

Gene Therapy: *CRISPR, Cas9 and CAR-T*

In the technology designated as "clustered, regularly interspaced, short palindromic repeats" (CRISPR) and the CRISPR-associated protein 9 (Cas9), the Cas9 enzyme functions as a fundamental part of the larger construct in which an RNA molecule guides the targeting of any possible matching DNA sequence. It is actually used to specify the critical site of cleavage. Since CRISPR–Cas has emerged as a highly flexible research tool for genome editing that has potential to enable researchers to manipulate the genome precisely, including the medical use of the system for directly treating genetic disorders, it has been widely publicized over the fundamental parts of the CRISPR–Cas9 system [56] [57] [58].

The combination of chimeric antigen receptors (CARs) and artificial T cell receptors, CAR-T, are engineered receptors which graft an arbitrary specificity onto an immune effector cell (T cell). Typically, these receptors are used to graft the specificity of a monoclonal antibody onto a T cell, with transfer of their coding sequence facilitated by retroviral vectors. These receptors are chimeric because they comprise parts from different sources. The general premise of CAR T-Cells is rapid generation of T-Cells

targeted to specific tumor cells. Once the T-Cell has been engineered to become a CAR T-Cell, it acquires supraphysiologic properties and develops the capability to act as a ‘Living Drug’ [59, 60, 61].

Immunotherapy / Vaccines: *Zika*

Zika fever, also known as Zika virus disease or simply Zika, is an infectious disease caused by the Zika virus [62]. Symptoms include red eyes, joint pain, headache, fever, and a maculopapular rash [63] [64]. Although it has caused no associated fatalities [65], mother-to-child transmission during pregnancy can cause microcephaly and other brain malformations in babies [66]. An outbreak that started in Brazil in 2015 spread to the Americas, Pacific, Asia, and Africa. This eventuality led to the World Health Organization‘s declaration of Zika as a Public Health Emergency of International Concern in February 2016 [62, 67]. Zika virus was rarely studied until the major outbreak. No specific antiviral treatment is available today [68].

3.2.2. Construction of Author Citation Networks and Co-authorship Networks

Papers published during 2013–2017 that include the aforementioned highest-growth keywords relative to actively financed biopharmaceutical industry fields: “Exosome," “Microbiome,” “CRISPR,” “Cas9,” “CAR-T,” or “Zika" in the title, abstract, or keywords were extracted from the Web of Science. Those papers are targeted as datasets to extract names of all authors and paper citation-related information to create author citation networks, as demonstrated in 3.1.1. Additionally, co-authorship networks were constructed from the papers above. The co-authorship network is a social network in which the authors, through participation in one or more publication through an indirect path, have linked mutually, whereas author citation networks are based on direct citation relation among the authors. Therefore, it is inferred that we might observe different characteristics related to how central the startup participant authors are and how they are distributed, between author citation networks and co-authorship networks (Table 3.9).

Table 3.9 Comparison Among Research Paper Citation Networks, Author Citation Networks, and Co-authorship Networks Relative To Growing Keywords in Actively Financed Biopharmaceutical Industry Fields in 2014–2017

Exosome	Paper Citation Network	Author Citation Network	Co-authorship Network
Node Count	1,941	11,084	11,059
Edge Count	7,625	379,180	57,697

Cas9	Paper Citation Network	Author Citation Network	Co-authorship Network
Node Count	3,974	19,893	19,808
Edge Count	38,856	1,415,583	133,925

Microbiome	Paper Citation Network	Author Citation Network	Co-authorship Network
Node Count	8,814	37,116	36,877
Edge Count	38,134	1,694,176	233,184

CAR-T	Paper Citation Network	Author Citation Network	Co-authorship Network
Node Count	685	3,302	3,281
Edge Count	4,377	277,377	25,106

CRISPRR	Paper Citation Network	Author Citation Network	Co-authorship Network
Node Count	5,451	25,411	25,251
Edge Count	52,945	1,742,614	171,281

Zika	Paper Citation Network	Author Citation Network	Co-authorship Network
Node Count	2,987	13,137	12,943
Edge Count	29,196	1,864,665	102,358

3.2.3. Detection and Visualization of Startup Participants Among Authors in Author Citation Networks and Co-authorship Networks

Using the VentureSource database again, all the names of the nodes (authors) are queried in the author citation networks and co-authorship networks to detect startup participants relative to emerging research topics represented by highest-growth keywords. Then, the rankings of the startup participant authors are constructed from their degree centralities both in their author citation networks and co-authorship networks for each keyword showing increasing frequency, based on the sum of both regularized orders squared. Rankings of startup participant authors for each highest-growth keyword are shown in Appendix B.

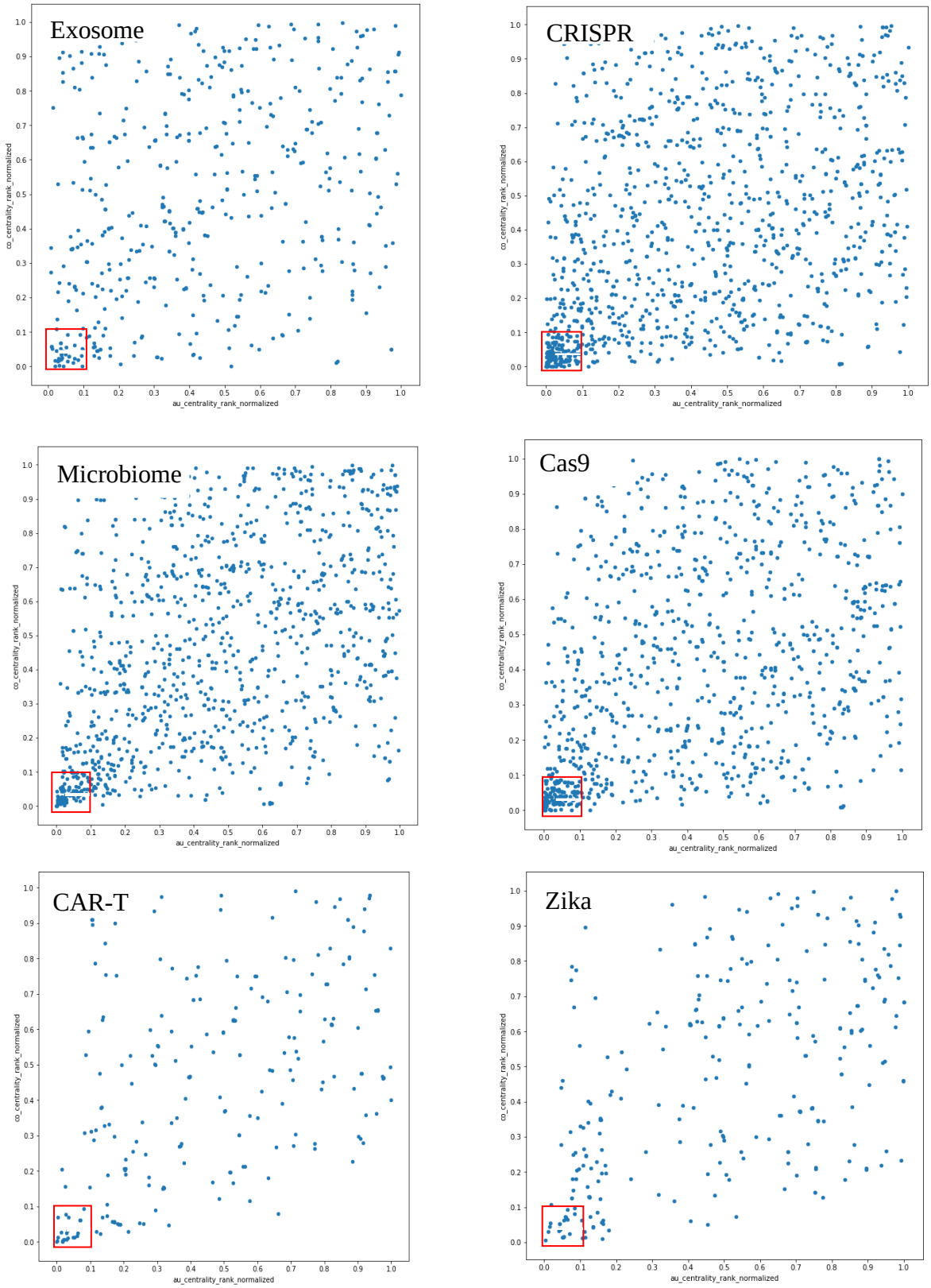
For each growing keyword above, a scatter diagram of the distribution of startup participant authors is mapped out, in terms of their rankings of degree centralities both in the author citation networks and the co-authorship networks (Figure 3-3).

For startup activities in which these startup participant authors are engaged, we can extract information from the VentureSource database related to the role of participants, company overview, financing to date, and so on. Although thorough case studies across all startup companies engaged by the all participants in this paper are not conducted, 18 top-degree centrality startup participants are listed herein with their 15 startups in each emerging research topic as previously shown in Table 1.3 on Page 5, to verify the collectiveness and relevance of the startup participant author pool and the significance of the selected names, by exemplifying several top startup participant authors for each

emerging research topic. These startups have been successful either at raising venture capital, achieving an IPO, or being acquired by big pharmaceutical companies according to the VentureSource database. We can access up-to-date information for each relevant startup by using VentureSource.

Results demonstrated that the startup participants concentrated near the top of both networks, especially around the top 10%, as shown in Figure 3-3. Correlation between both the ranks of the author citation networks and those of the co-authorship networks to the same startup participants was not strong: their correlation coefficients were 0.337 (Exosome), 0.464 (Microbiome), 0.371 (CRISPR), 0.337 (Cas9), 0.505 (CAR-T) and 0.528 (Zika).

Figure 3-3 Scatter Diagram, Distribution of Startup Participant Authors' Degree Centrality in Author Citation Networks & Co-authorship Networks based on Ratio from Top to Bottom, in Emerging Research Topics in Actively Financed Biopharmaceutical Industry Fields in 2014–2017



3.2.4. Hypothesis Testing of Top 10% Authors in Both Networks

From the observations presented in Figure 3-3 (B-1) regarding the six emerging research topics in the biopharmaceutical domain, it is hypothesized that the proportion of startup participants is higher among authors of the top 10% degree centrality in both networks (designated hereinafter as "Dual Top 10% Authors") than it is among authors who do not have such high centrality.

In order to conduct testing of the above hypothesis, ***Fisher's exact test*** is used to infer significance of differences in the observed proportions. Fisher's exact test, a test of statistical significance used for analysis of contingency tables, assesses significance of deviation from a null hypothesis, or *P*-value, calculated exactly as long as the contingency tables' row and column totals are fixed, rather than relying on an approximation, as does a chi-square approximation [69, 70, 71]. This section calculated the probability *P* that the number of startup participants is equal to or exceeds the observed number among "Dual Top 10% Authors," under the null hypothesis that startup participants are equally likely to be distributed among authors in both networks regardless of their degree centralities. Additionally, calculation was conducted on how many times higher the odds of being a startup participant is among "Dual Top 10% Authors" compared to other authors, i.e., odds ratio [72, 73] too.

Following are the findings relative to the six emerging research topics in Table 3.10. It is inferred that the results we observed in their odds ratios were statistically significant.

(i) The *P*-values were 1.321e-07 (Exosome), 2.714e-11 (Microbiome), 2.288e-48 (CRISPRR), 2.395e-36 (Cas9), 1.584e-3 (CAR-T), and 0.0401 (Zika), all of which were equal to or less than three places of decimals except for zika's *P*-value, which was still less than 0.05, the number used as the cutoff in most statistical hypothesis testing.

(ii) Odds ratios across all the emerging research topics were 2.899 (Exosome), 2.138 (Microbiome), 5.338 (CRISPR), 4.773 (Cas9), 2.222 (CAR-T) and 1.651 (Zika), all of which observed higher startup participant ratio in "Dual Top 10% Authors" than in other authors.

Table 3.10 Contingency Tables Related to the Number of Startup Participants and Non-Participants for Dual Top 10% Authors and Others with P-value and Odds Ratio for Each Research Topic

Exosome	Start-up Participant	Non-Participant
Dual Top 10% Authors	37	299
Other Authors	439	10,284
P-Value: 1.321e-07	Odds Ratio: 2.899	

Cas9	Start-up Participant	Non-Participant
Dual Top 10% Authors	118	520
Other Authors	870	18300
P-Value: 2.395*e*-36	Odds Ratio: 4.773	

Microbiome	Start-up Participant	Non-Participant
Dual Top 10% Authors	107	1712
Other Authors	9951	34040
P-Value: 2.714*e*-11	Odds Ratio: 2.138	

CAR-T	Start-up Participant	Non-Participant
Dual Top 10% Authors	23	147
Other Authors	206	2906
P-Value: 1.585e-3	Odds Ratio: 2.222	

CRISPRR	Start-up Participant	Non-Participant
Dual Top 10% Authors	142	611
Other Authors	1018	23380
P-Value: 2.288e-48	Odds Ratio: 5.338	

Zika	Start-up Participant	Non-Participant
Dual Top 10% Authors	20	568
Other Authors	257	12096
P-Value: 0.0401	Odds Ratio: 1.651	

3.3. Evaluating Network Centrality, Co-authorship Centrality, and Hot Topic Factors as Potential Distinctive Features

As shown in 3.1 and 3.2, in this chapter, in order to explore features specifically suited to academic researchers in the emerging biopharmaceutical research fields to assess their startup readiness, this thesis developed authors' network centrality, co-authorship centrality and hot topic factors to measure the degree of how much emerging their fields/topics are.

Both of centralities were specifically computed on the basis of authors among their networks of author citation and co-authorship respectively, derived from Web of Science Core Collection. Results show that both centralities could work as potential features for such academic researchers for the assessment of their startup readiness, on the premise that they are in the fields/topics with high degree of hot topic factors derived from VentureSource as well as Web of Science Core Collection. This suggests that hot topic factors could also function as desirable features for such researchers.

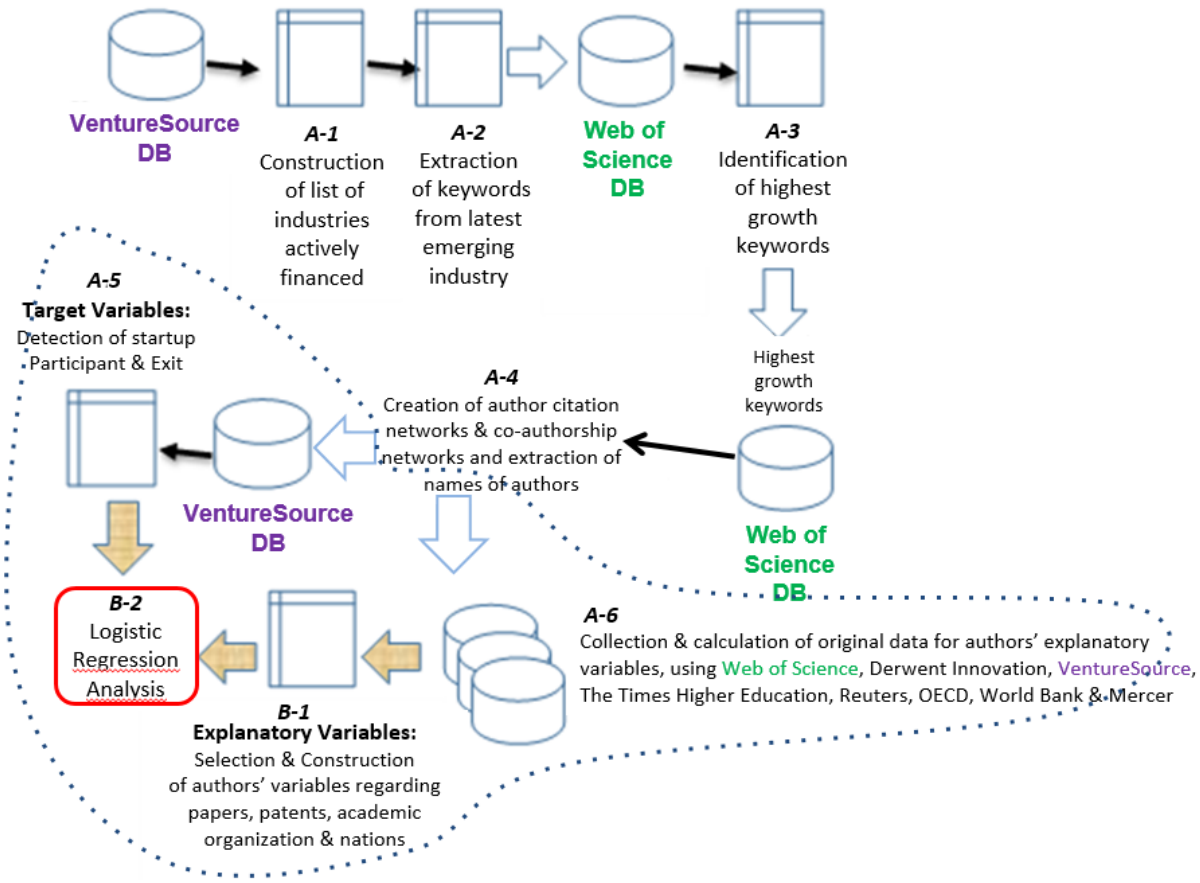
In the following chapters, these potential individual and non-individual features will be implemented and evaluated in the assessment model of this thesis, to test their possibility of being distinctively useful to academic researchers in emerging research fields with intense science linkage such as the biopharmaceutical domain.

Chapter 4. Designing the Assessment Model with Features

This chapter will propose an assessment model of startup readiness for academic researchers in emerging research fields with intense science linkage, primarily in the biopharmaceutical domain as this thesis's case studies, partly by referring to my prior work (2019) with co-authors regarding the startup participant researchers related to Cas9 [20]. This thesis attempts to expand and modify earlier literature on explanatory variables of academic startups, by introducing new data sources, and potential features derived from them that are explored in Chapter 3.

The analytical scheme of this thesis is shown in Figure 4-1 as follows, of which steps until A-3 are virtually the same as those in Figure 3-2.

Figure 4-1 Methodology Proposed in Chapter 4



Using VentureSource, all financing deals between January 1, 2017 and December 31, 2017, were analyzed, to compile a list of industry segments that were the most active in venture capital finance in 2017 **(A-1)**. From the VentureSource database, we can extract data of daily global startup investment deals, with respect to each industry field with its specific industry code/subcode. Information related to the amount of financing, the number of financing rounds, keywords, and participants of the startups are available. Using VentureSource, a ranked list of the 30 most active financed industry codes/subcodes were compiled among all 281 VentureSource industry codes/subcodes in 2017, based on the sum of the ranks of both the average financing size and the number of financing rounds (See Table 3.1). Then, industry codes/subcodes belonging to industry segment designated as "Biopharmaceutical" were extracted. For startups in target codes/subcodes belonging to the biopharmaceutical segment on VentureSource, keywords that appear multiple times were surveyed **(A-2)** (See Table 3.7). Then, differences in keyword frequency in research papers from 2014 through 2017 were analyzed, by searching the Web of Science Core Collection. Through this process, the highest-growth keywords for the period above for each target code/subcode were identified, as emerging research topics **(A-3)** (See Table Table 3.8).

The following is newly addressed in this chapter. For the aforementioned highest-growth keywords, author citation networks and co-authorship networks were created and the names of all the relevant authors who belong to these networks were extracted **(A-4)**. Among them, in order to prepare target variables, authors who became participants of startups and those who achieved exits were detected, as of December 2018, using the VentureSource database again **(A-5)**. Furthermore, original data for these authors' explanatory variables were collected and calculated, using following data sources: Web of Science Core Collection, Derwent Innovation, VentureSource, The Times Higher Education, Reuters, OECD, World Bank and Mercer **(A-6)**.

Finally, explanatory variables were selected and constructed from their original data composed of features regarding their papers, patents, academic organizations and nations, based on hypotheses of this thesis that might be useful to assess their startup readiness **(B-1)**. Logistic regression analysis will be conducted then **(B-2)**. Details of this paragraph will be addressed in Chapter 5, as the implementation of the assessment model.

In this way, Web of Science Core Collection and VentureSource are continuously used as data sources herein. Through the same procedure as that in

Figure 3-2 up to A-4, research topics to explore were identified, which are highest-growth keywords within the target period.

Regarding explanatory variables related to papers, patents, academic organizations, and nations to assess startup readiness, in addition to the above Web of Science Core Collection database, the following several data sources were incorporated: *Derwent Innovation - Derwent World Patents Index; The Times Higher Education - World University Rankings 2017; Reuters - The World's Most Innovative Universities 2017; OECD - Entrepreneurship at a Glance 2017; World Bank - Doing Business 2017;* and *Mercer - 2017 Workforce Turnover Around the World,* as discussed in 2.3.

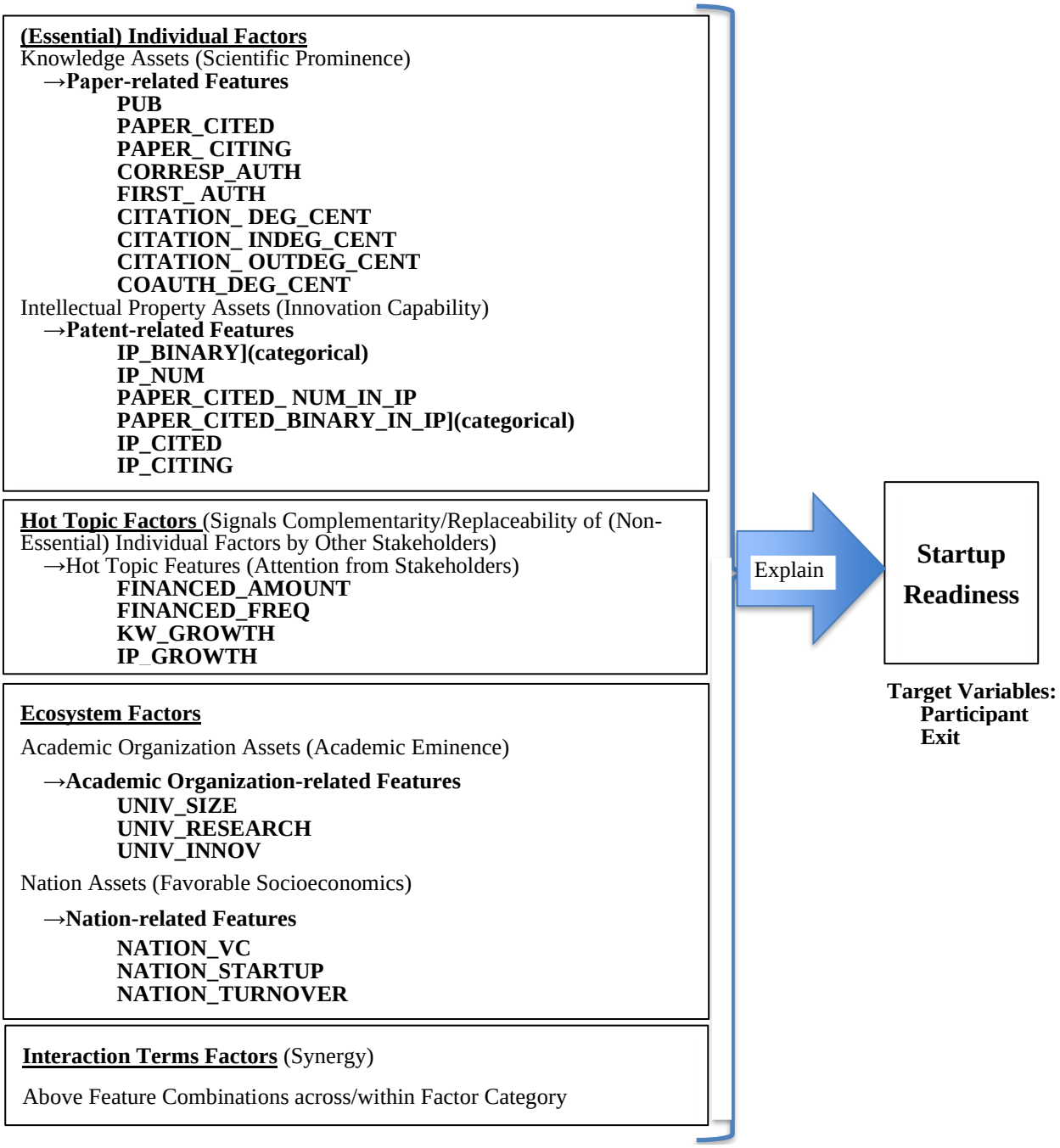


Figure 4-2 Conceptual Framework with Features to Assess Academic Researchers' Startup Readiness

As target binary variables to judge startup readiness, this thesis questioned whether academic researchers are recognized in VentureSource, not only as startup participants as surveyed in 3.1 and 3.2, but also as those who achieved an exit (i.e., IPO and M&A). Part of VentureSource data is related to exits associated with academic startups that those researchers are involved with. The reason exits are added as another target variable is that exits are a crucial factor to consider whether the relevant startup achieves success or not.

With these two target variables of startup participants and exits, it is hypothesized that explanatory variables composed of features of papers, patents, universities, and nations related to relevant research topics, could do a good assessment of researchers' startup readiness. To construct assessment models with effective sets of explanatory variables, this thesis selected and derived features from relevant data sources, by transforming some of them and creating interaction terms out of them, as shown in Figure 4-2.

As will be seen in Chapter 5 later, results show that the proposed assessment model yields good assessment and classifying performance, and carries specific implications about which variables and their combinations demand attention.

4.1. Preprocessing Data

Prior to the construction and implementation of the assessment model, this section followed basically the same but extended methodology relative to that in 3.2.1, using the data sources such as VentureSource and Web of Science Core Collection (See 2.3), as follows.

4.1.1. Construction of the List of Biopharmaceutical Industries Most Actively Financed (A-1)

Among Venture Source's 281 industry codes/subcodes, top 30 ranking of the most actively financed are arranged, according to the sum of the orders of both the average financing size and the number of financing rounds, based on 17,681 financing deals that took place during January 1, 2017 through December 31, 2017, as depicted in Table 3.1. From those top 30 industry codes/subcodes, those that belong to segment "Biopharmaceuticals" were five industry codes/subcodes: "Pharmaceuticals," "Biotechnology Therapeutics," "Immunotherapy / Vaccines," "Small Molecule Therapeutics," and "Gene Therapy," of which average financing sizes were 32.28, 23.87, 29.05, 25.72, and 30.58 million USD respectively, and of which number of rounds were 129, 154, 54, 67 and 47 respectively.

This step was exactly the same as A-1 in 3.2.

4.1.2. Extraction of Keywords from the Above Biopharmaceutical Industries (A-2)

Sourced from the VentureSource database, keywords for the startups that belonged to the above five biopharmaceutical industry codes/subcodes in Table 3.1, were surveyed. Then, keyword lists for each industry code/subcode were constructed. Keywords that appeared twice or more on each list are presented in Table 3.7.

This step was exactly the same as A-2 in 3.2.

4.1.3. Identification of Highest Growth Keywords (A-3)

Using Web of Science Core Collection, rankings of research topic keywords showing the highest growth in incidence, which correspond to each industry code/subcode in VentureSource, are constructed based on growth multiples during 2014–2017 (Table 3.8). From the rankings, for an additional survey, this thesis extracts keywords with incidence that more than doubled between 2014 (Start Year) and 2017 (End Year) and that appeared more than 100 times in 2017 (End Year). Keywords with growth multiples that surpassed 50 times, however, are excluded, as such words might just be an outbreak. Eventually, keywords that met this criterion from each industry code/subcode turned out to be the following five keywords: Exosome, Microbiome, CRISPR, CAR-T, and Cas9, of which growth multiples were 2.20, 2.54, 4.52, 5.00 and 5.32 respectively. Zika, which has the largest multiplicative factor of x88.85, was excluded, as Zika is the name of a fever that became an outbreak starting in Brazil in 2015. In any case, all of these five keywords are emerging research topics that have attracted academic attention to a rapidly increasing degree.

This step was almost the same as A-3 in 3.2, albeit Zika was removed herein.

4.1.4. Creation of Author Citation Networks and Co-Authorship Networks and Extraction of the Names of Authors from These Networks (A-4)

This step is similar to what was performed in 3.2.2, while this section completed this step for the authors related to both **Cas9** and **Microbiome** (the reason for selection of these two research topics will be touched upon at the beginning of Chapter 5), and also for the authors related to all five keyword research topics combined: **Exosome**, **Microbiome**, **CRISPR**, **CAR-T** and **Cas9**, while 3.2.2 covered six biopharmaceutical research topics individually including **Zika**. Through this step, both author citation networks and co-authorship networks, derived from paper citation networks based on papers published during 2013-2017 regarding **Cas9**, **Microbiome** and the rest of the

above five topics were created, and the names of authors relevant to these topics were extracted.

Table 3.9 presents a comparison among paper citation networks, author citation networks, and co-authorship networks relative to these topics. These networks can be used for the purposes of calculating several authors' centralities as relevant paper-related variables, to be discussed in 4.4.1.1.

4.1.5. Creation of Binary Variables Regarding Participants and Exits as Authors' Target Variables (A-5)

While this step is similar to what was performed in 3.2.3, not only startup participants but also exits of their startups are considered herein. The names of the authors related to the above five research topics who were included in the *VentureSource* database, as of December 31, 2018, were queried and detected if they are participants in startups and if their involved startups had achieved exits such as an M&A and an IPO in the database.

Therefore, two kinds of binary variables were created as the target variables for each author (academic researcher), indicating whether the author appears as "a participant in a startup [Participant]" (coded as 1) or not (coded as 0), and whether the researcher appears as "a participant whose startup achieved an exit [Exit]" (coded as 1) or not (coded as 0).

4.1.6. Collection and Calculation of Original Data for Authors' Explanatory Variables (A-6)

As described in the conceptual framework in Figure 4-2, four categories of explanatory variables were conceptualized for this thesis: (1) Individual Factors (composed of Paper-related Features and Patent-related Features), (2) Hot Topic Factors/Features, (3) Ecosystem Factors (composed of Academic Organization-related Features and Nation-related Features) and (4) Interaction Terms Factors. Thus, original data for these variable categories were collected and calculated from the data sources accordingly as follows. Each explanatory variable will be described later in detail in 4.4.

Across all academic researchers (authors), this thesis collected data of the common duration/timing corresponding to each type of explanatory variables, irrespective of the timing of an event ("Participant" or "Exit") taking place for each researcher, rather than changing data collection period depending on researchers. This issue will be described at the beginning of Chapter 5 in more detail.

4.1.6.1. Individual Factors

(i) Paper-related Features

Using *Web of Science Core Collection*, this thesis calculated the counts of each author's publications [PUB], frequency of citation in other papers [PAPER_CITED], frequency of citing other papers [PAPER_CITING], frequency of being a corresponding author [CORRESP_AUTH], and frequency of being a first author [FIRST_AUTH], derived from papers published during 2013-2017 that have either of the aforementioned five research topics: Exosome, Microbiome, CRISPR, CAR-T and Cas9 respectively, in their titles, abstracts or keywords.

Moreover, this thesis calculated each author's author citation degree centrality [CITAION_DEG_CENT], author citation in-degree centrality [CITATION_INDEG_CENT], author citation out-degree centrality [CITATION_OUTDEG_CENT] and co-author degree centrality [COAUTH_DEG_CENT], derived from author citation networks and co-authorship networks described in 4.1.5 based on papers published during 2013-2017 that include either of the aforementioned five research topics. The normalized degree centrality, the normalized in-degree centrality, the normalized out-degree centrality in author citation networks and the normalized degree centrality in co-authorship networks were computed, as demonstrated in 3.1.1.

(ii) Patent-related Features

Using *Derwent Innovation – Derwent World Patents Index, sourcing* patent publications issued during 2013-2017 that include either of the five research topics: Exosome, Microbiome, CRISPR, CAR-T and Cas9, in their abstracts, this thesis first queried each author's name relative to the above five research topics, to detect if the author is included as an inventor of a patent in the research topic [IP_BINARY] in the database. This thesis then created binary variables that indicate whether the author appears as an inventor (coded as 1) or not (coded as 0). Moreover, in a similar fashion, it is queried whether there is(are) paper(s) cited in patents that he/she is an inventor of [PAPER_CITED_BINARY_IN_IP] in the database, and created binary variables to indicate whether the author cited (a) paper(s) in the patent publications that included him/her as an inventor.

Furthermore, using the same data source, this thesis calculated the counts of each author's (inventor's) number of patents that the author is an inventor of [IP_NUM], frequency of citing academic papers in patens [PAPER_CITED_NUM_IN_IP], frequency of being cited by other patents [IP_CITED], and frequency of citing other patents [IP_CITING].

4.1.6.2. Hot Topic Factors/Features

Using *VentureSource*, based on 17681 financing deals in 2017, this thesis calculated the average financing deal size [FINANCED_AMOUNT] and the number of financing deals [FINANCED_FREQ] of the five biopharmaceutical industry codes/subcodes: Pharmaceuticals, Biotechnology Therapeutics, Immunotherapy/Vaccines, Small Molecule Therapeutics, and Gene Therapy, either of which each author's research topic belongs to, according to 3.2.1. (See Table 3.1)

Moreover, this thesis then calculated (i) the keyword frequency growth multiple [KW_GROWTH], i.e., growth multiple in the annual frequency of each author's research topic that appeared in the title, abstract, keywords, or Keyword Plus of the papers published between both the years 2014 and 2017 (See Table 3.8) using *Web of Science Core Collection*, and (ii) IP growth multiple [IP_GROWTH], i.e., growth multiple in the annual frequency of patent publication, of which the abstract each author's research topic appeared in, between both the years 2014 and 2017 using *Derwent Innovation's Derwent World Patents Index.*

4.1.6.3. Ecosystem Factors

(i) Academic-organization-related Features

To construct Academic-organization-related Features, relevant data was extracted from Times Higher Education's *World University Rankings* 2017 for the number of full-time students and the research score, and Reuters' *The World's Most Innovative Universities 2017* for the score of innovativeness, with respect to academic organizations that each author belongs to. Then, by referring to all the relevant data for each author, this thesis calculated the weighted number of full-time students of the academic organization to which an author's corresponding author belongs [UNIV_SIZE], the weighted research score of the academic organization to which an author's paper's corresponding author belongs [UNIV_RESEARCH], and the weighted score of innovativeness of the academic organization to which an author's paper's corresponding author belongs [UNIV_INNOV].

(ii) Nation-related Features

Furthermore, in order to build Nation-related Features, relevant national socioeconomics data was extracted from OECD's *Entrepreneurship at a Glance* 2017 for venture capital investment as a percentage of GDP, World Bank's *Doing Business* 2017 for the score for starting business, and Mercer*'s Workforce Turnover Around the World* 2017 for life science workforce voluntary turnover, with respect to countries that each author's corresponding author belongs to. Next, in the same fashion as (i), by

referring to all the relevant data for each author, this thesis calculated the weighted venture capital investment as a percentage of GDP of the country to which an author's corresponding author belongs [NATION_VC], weighted World Bank score for starting business in the country to which an author's corresponding author belongs [NATION_STARTUP], and weighted life science workforce voluntary turnover of the country to which an author's corresponding author belongs [NATION_TURNOVER].

4.2. Detection and Assessment of Startup Readiness Using Logistic Regression

Then, to assess the categorical probability of an event ("Participant" or "Exit") occurring given a select number of continuous and categorical variables, this thesis designs a logistic regression model, the specification of which is described in Chapter 5. In contrast to logistic regression, machine learning methods such as random forest, boosting, or neural networks have no underlying distributional assumptions, can handle complex relationships between explanatory variables and the outcome, and require no model specification. However, machine learning methods are often considered "black box" methods as they do not readily provide the user with any indication of the importance of individual explanatory variables that are used for the prediction output. Logistic regression models provide effect estimates (odds ratios) that are easily interpretable, and the advantages of logistic regression models include the comparatively easy implementation, the availability in all standard statistical software packages, and short computation times. Since in order to assess startup readiness, the interpretability of determinants for a variety of stakeholders easily is critical, logistic regression was used in this research. [74]

To examine startup readiness of academic researchers, a relevant logistic regression model was designed to determine the probability of academic researchers displaying startup readiness, be it participation in academic startups or exits of them. The following logistic regression model was constructed to calculate the odds ratio regarding the probability.

$$\log(\text{ODDS}) = \log\left(\frac{P_i}{1-P_i}\right) = \beta_0 + \beta_i(X_i) \tag{4-1}$$

where P_i is the probability of researcher i displaying startup readiness, while β_0 is y-intercept that is the log-odds of the event that log (ODDS) = 1 when all the explanatory variables belonging to reference group X_i associated with researcher i are 0, and $\beta_i(X_i)$ are the regression coefficients of the explanatory variables belonging to reference group X_i associated with researcher i.

As discussed in 2.1, it is hypothesized that the explanatory variables composed of the features of papers, patents, universities, and nations related to relevant research topics, could well assess researchers' startup readiness. As seen in 2.3, relevant data sources from which this thesis explored variables in this thesis were: *Web of Science Core Collection, Derwent Innovation - Derwent World Patents Index; The Times Higher Education - World University Rankings 2017; Reuters - The World's Most Innovative Universities 2017; OECD - Entrepreneurship at a Glance 2017; World Bank - Doing Business 2017;* and *Mercer - 2017 Workforce Turnover Around the World*.

The strategy of this thesis to explore and integrate variables is that they should be well accountable and understandable for a wide range of stakeholders of academic startups - not only for academic researchers, but also for practitioners who would complement or even replace academic researchers' incomplete financial, social, personal assets and traits, other than their proprietary knowledge assets and intellectual property assets. Constraints on the variables to explore herein are that they should be useful and effective for important practitioners outside the academia too, such as venture capitalists and managerial talents who have stakes in academic startups, thus limiting the data sources only to those that are purchasable or publicly available on a real-time or timely basis digitally. Even though these variables lack some attributes that could be attained by sources like personal interviews or customized surveys to scientists, this strategy allows us to utilize data in a scalable manner without limitation to our acquainted sources.

4.3. Target Variables

The target variable of a dataset is the feature of a dataset whose values are to be modeled and predicted by other variables (explanatory variables). It is or should be the output about which we want to gain a deeper understanding. It is important to have a well-defined target since what a predictor model does is to learn a function that maps relationships between input data (explanatory variables) and the target.

In this thesis, startup readiness refers to the concept describing the state when one is prepared for initiating startups and willing to do so with a hope to be successful (See 1.1). Thus, as target variables, it is necessary to identify features that signal participation by academic researchers and success of startups. In this regard, participation by academic researchers is measured by observing whether academic researchers are registered as startups' participants in the VentureSource database as of December 31, 2018. On the other hand, further argument is needed regarding the definition of startups' success. In this dissertation, however, success by academic researchers is measured by observing whether academic researchers are registered as those who have experienced startups' exits in the VentureSource database as of December 31, 2018. It is presumed that an

academic startup achieves success when it accomplishes an exit, be it an IPO or an M&A, from the points of views of a wide range of stakeholders who could be equity holders of startups. Information regarding the timing of startup's IPO and M&A can also be retrieved from VentureSource .

4.4. Explanatory Variables

The explanatory variable of a dataset is the feature of a dataset whose values are to be used to explain differences in the target variable. Also known as the independent variable, it explains variations in the target variable.

As described in the conceptual framework in Figure 4-2, this thesis conceptualized four categories of explanatory variables for this thesis: (1) Individual Factors, (2) Hot Topic Factors, (3) Ecosystem Factors and (4) interaction terms factors, then constructed each feature as explanatory variables, as follows.

4.4.1. Individual Factors

To shed new light on startup readiness, this thesis emphasizes the influence of individual factors of academic researchers in terms of their prominence as scientists and its profile, together with their intention to apply their scientific outcome in society and its feasibility.

Academic researchers' prominence as scientists and its profile, together with their intention to apply their scientific outcome in society and its feasibility, act as incubators of startup readiness because these features provide more focus than conventional startups on fundamental scientific discoveries aimed at solving scientific problems and at providing immediate social benefits.

It is presumed that academic startups in the biopharmaceutical field have intensive scientific, innovative linkage among research and social utilization, of which indicators include bibliometric data of paper-related features retrieved from the Web of Science Core Collection database and patent-related features retrieved from the Derwent Innovation database, as follows.

4.4.1.1. Paper-related Features

Most research knowledge produced by academic researchers contributes to the pool of open science, measurable by paper-related indexes. The traditional vision of university research is that faculty members who are highly active in academic research exhibit a strong commitment of time and orientation to advancing research knowledge at the expense of knowledge transfer. However, the entrepreneurial vision of the university

induces researchers to consider their publication of papers as knowledge assets that can be transferred and commercialized outside the scholarly community [75].

Based on this rationale, this thesis hypothesizes the following statement, and, by using Web of Science Core Collection, computes the following features later for all the authors among two groups: (1) papers related to **Cas9** and **Microbiome**, and (2) papers related to highest-growth five biopharmaceutical research topics: **Cas9**, **CAR-T**, **CRISPR**, **Microbiome** and **Exosome** combined. As discussed in 3.2.1, these five topics matched the criteria therein as highest-growth research topics in this order, and Cas9 was first-ranked.

H1. Higher paper-related indexes reflect higher startup readiness by researchers.

Paper-related Features (9 continuous variables)

- Publications [PUB]:
  Counts of the author's publications
- Frequency of citation in other papers [PAPER_CITED]:
  How frequent the author's papers were cited in other papers
- Frequency of citing other papers [PAPER_ CITING]:
  How frequent the author cited other papers in his/her papers
- Frequency of being a corresponding author [CORRESP_AUTH]:
  How frequent the author was a corresponding author in papers he/she co-authored
- Frequency of being a first author [FIRST_ AUTH]:
  How frequent the author was a first author in papers he/she co-authored
- Author citation degree centrality [CITATION_ DEG_CENT]:
  The normalized degree centrality of the author in the author citation network as a whole: the number of researchers citing or being cited by the author's papers in the network, divided by the maximum possible degree, i.e., the total number of researchers in the network minus one
- Author citation in-degree centrality [CITATION_ INDEG_CENT]:
  The normalized in-degree centrality of the author in the author citation network as a whole: the number of researchers citing the author's papers in the network, divided by the maximum possible degree, i.e., the total number of researchers in the network minus one
- Author citation out-degree centrality [CITATION_ OUTDEG_CENT]:
  The normalized out-degree centrality of the author in the author citation network as a whole: the number of researchers cited by the author's papers in

the network, divided by the maximum possible degree, i.e., the total number of researchers in the network minus one

- Co-author degree centrality [COAUTH_DEG_CENT]:
  The normalized degree centrality of the author in the co-authorship network as a whole: the number of researchers who co-author with the author in the network, divided by the maximum possible degree, i.e., the total number of researchers in the network minus one

It is assumed that the paper-related variables showing researchers' prominence as scientists would work well as explanatory variables for their startup readiness. Such researchers' prominence can be categorized into three aspects: publication of scientific activities, academic attention, and centrality in academic networks. (i) Publications [PUB], Frequency of citing other papers [PAPER_CITING], Frequency of being a corresponding author [CORRESP_AUTH], Frequency of being a first author [FIRST_AUTH], and Author citation out-degree centrality [CITATION_ OUTDEG_CENT] can be considered as variables that show researchers' activeness in publishing their scientific activities; (ii) Frequency of being cited by other papers [PAPER_CITED] and Author citation in-degree centrality [CITATION_ INDEG_CENT] show how much academic attention these researchers receive; (iii) Author citation degree centrality [CITATION_DEG_CENT] and Co-author degree centrality [COAUTH_DEG_CENT] are indexes that show researchers' overall centrality among author citation networks and co-authorship networks.

4.4.1.2. Patent-related Features

Patents indicate academic researchers' innovativeness: they are used most frequently to indicate the entrepreneurial activities of academic researchers.

Based on this rationale, this thesis hypothesizes the following statement, and, by using *Derwent Innovation - Derwent World Patents Index*, computes the following features for the authors in 4.4.1.1 who were also found in the Derwent Innovation database as inventors. The computation among the two groups is conducted as performed in 4.4.1.1: (1) papers related to **Cas9** and **Microbiome**, and (2) papers related to the five biopharmaceutical topics: **Cas9**, **CAR-T**, **CRISPR**, **Microbiome** and **Exosome** combined.

H2. Higher patent-related indexes are associated with higher startup readiness by researchers.

Patent-related Features (4 continuous, 2 categorical variables)

- Inventor of a patent in the research topic [IP_BINARY] (categorical):

Whether the author is an inventor of patents in the relevant research topic

- Number of patents that the author is an inventor of [IP_NUM]:

  Counts of the patents that the author is an inventor of in the relevant research topic

- Frequency of citing academic papers in patents [PAPER_CITED_ NUM_IN_IP]:

  How frequent the author cited academic papers in the patents that he/she is an inventor of

- Paper cited in patents that he/she is an inventor of [PAPER_CITED_BINARY_IN_IP] (categorical):

  Whether the author cited a paper in the patents that he/she is an inventor on

- Frequency of being cited by other patents [IP_CITED]:

  How frequent the patents that the author is an inventor of were cited by other patents

- Frequency of citing other patents [IP_CITING]:

  How frequent the author cited other patents in the patents he/she is an inventor of

It is assumed that patent-related variables that show researchers' innovation mindset would work well as explanatory variables for their startup readiness. This thesis hypothesizes that such researchers' innovativeness can fall into three aspects: invention of patents, transformation of scientific outcomes into patents, and patent attention. (i) Inventor on a patent in the research topic [IP_BINARY], Number of patents that the author is an inventor of [IP_NUM], and Frequency of citing other patents [IP_CITING] can be considered as variables showing researchers' activity in the invention of patents; (ii) Frequency of citing academic papers in patents [PAPER_CITED_ NUM_IN_IP] and Author citing a paper in patents that he/she invented [PAPER_CITED_BINARY_IN_IP] represent how active the researchers are in transforming scientific outcomes into patents; (iii) Frequency of being cited by other patents [IP_CITED] shows how much patent attention the researchers receive.

4.4.2. Hot Topic Factors/Features

As discussed in 2.3, Hot Topic Factors/Features are incorporated into this thesis, to measure the degree of social attention from financial, scientific and innovative perspectives to specified discipline or research topic in question.

Since they signal the extent to which academic researchers can attract competent stakeholders, such as venture capitalists and managerial talents, to their academic

startups, this thesis presumes that those researchers' lack of financial, social, personal assets and other relevant personal traits can be complemented or even replaced by stakeholders other than academic researchers. These features should be extremely valuable when we compare academic researchers across different industry segments and research topics.

Based on this rationale, this thesis hypothesizes the following statement, and, by using *VentureSource*, *Web of Science Core Collection*, and *Derwent Innovation*, computes the following features for the authors in 4.4.1.1 who were also found in the Derwent Innovation database as inventors. The computation among two groups is conducted as performed in 4.4.1.1: (1) papers related to **Cas9** and **Microbiome**, and (2) papers related to the five biopharmaceutical topics: **Cas9**, **CAR-T**, **CRISPR**, **Microbiome** and **Exosome** combined.

H3. Higher hot topic indexes are associated with unimportance of researchers' lack of individual factors other than paper-related and patent-related features.

Hot Topic Features (4 continuous variables)

- Average financing deal size [FINANCED_AMOUNT]:
  Average financing size per deal among startups in the relevant biopharmaceutical industry code/subcode in 2017, according to VentureSource
- Number of financing deals [FINANCED_FREQ]:
  The number of financing deals among startups in the relevant biopharmaceutical industry code/subcode in 2017, according to VentureSource
- Keyword frequency growth multiple [KW_GROWTH]:
  Growth multiple in the annual frequency of the relevant research topic that appeared in the title, abstract, keywords, or Keyword Plus of the papers during 2014-2017, according to Web of Science Core Collection
- IP growth multiple [IP_GROWTH]:
  Growth multiple in the annual frequency of patent publication, of which the abstract the relevant research topic appeared in, during 2014-2017, according to Derwent Innovation

Average financing deal size [FINANCED_AMOUNT], and Number of financing deals [FINANCED_FREQ] are variables showing how much and how often the relevant biopharmaceutical segment attracted venture capital. Keyword frequency growth multiple [KW_GROWTH] shows how much growth in attention the relevant research topic attained, which could signal the trend of attention from a wide range of

stakeholders. Lastly, IP growth multiple [KW_GROWTH] is a variable indicating how much growth in attention the relevant research topic attracted, which potentially signals the growth in the interest regarding application in society of the relevant research findings.

4.4.3. Ecosystem Factors

Scientific startups are founded and supported not only by individual factors of scientific founders, but also by the ecosystem surrounding the scientists, most typically composed of their research institutes and nations. Indeed, the larger the size of universities and the higher the level of the academic research, the larger the reservoir of resources and expertise linked to laboratories, technology transfer offices and star scientists' expertise that can be mobilized to foster the entrepreneurial vision of university research [76, 77, 78, 79]. Moreover, some attributes of national innovation systems are at the core of entrepreneurship and entrepreneurial innovation because they enable scientists to access crucially important resources such as capital, labor, and environments favorable to their research-based startups [80]. Strong positive externalities for research-based startups can be generated within leading research institutes and nations with a proactive innovation environment. Therefore, one can hypothesize the following.

H4. Better academic-organization-related features are associated with higher startup readiness by researchers.

Academic Organization-related Features (3 continuous variables)

- Weighted number of full-time students of the academic organization to which an author's corresponding author belongs [UNIV_SIZE]
- Weighted research score of the academic organization to which the paper's corresponding author belongs [UNIV_RESEARCH]
- Weighted score of innovativeness of the academic organization to which the paper's corresponding author belongs [UNIV_INNOV]

H5. Stronger nation-related features are associated with higher startup readiness by researchers.

Nation-related Features (three continuous variables)

- Weighted venture capital investment as a percentage of GDP of the country to which an author's corresponding author belongs [NATION_VC]

- Weighted World Bank score for starting business in the country to which an author's corresponding author belongs [NATION_STARTUP]
- Weighted life science workforce voluntary turnover in the country to which an author's corresponding author belongs [NATION_TURNOVER]

4.4.4. Multi-Variable Fractional Polynomials (MFPs) for Above Factors

One assumption for logistic regression analysis, which is conducted in this research, is linearity in its link function. Variables are assumed to be associated linearly with the response variable in logit scale. However, such is not always the case; the assumption might therefore be wrong. In exploratory studies, investigators must rely on data to ascertain the functional form. Multivariable fractional polynomial (MFP) method is such a method that it allows software to determine the functional form of an explanatory variable whether it is important to the model, or not. MFP is convenient when investigators want to preserve the continuous nature of variables when the relation is nonlinear [81]. This will be discussed later in 5.1.1.2.

All continuous sole variables, i.e., all continuous variables other than interaction terms described in 4.4.5, are checked regarding whether their observed probability plots are clustered linearly.

4.4.5. Interaction Terms Factors

Other than the above variables, this thesis considers coordination of these variables too, because, rather than relying solely on each variable individually, the coordination of variables can predict startup readiness more effectively. Therefore this thesis takes into account all feasibly possible "interaction terms factors," or combinations of practically possible two features across individual, hot topic and Ecosystem Factors. It is hypothesized that startup readiness can vary depending on the combination of these variables too, rather than depending solely on individual variables. This thesis surveys following combinations of groups of features.

- (Paper-related Features) * (Paper-related Features)
- (Paper-related Features) * (Patent-related Features)
- (Paper-related Features) * (Hot Topic Features)
- (Paper-related Features) * (Academic Organization-related Features)
- (Paper-related Features) * (Nation-related Features)
- (Patent-related Features) * (Patent-related Features)
- (Patent-related Features) * (Hot Topic Features)
- (Patent-related Features) * (Academic Organization-related Features)

- (Patent-related Features) * (Nation-related Features)
- (Hot Topic Features) * (Hot Topic Features)
- (Hot Topic Features) * (Academic Organization-related Features)
- (Hot Topic Features) * (Nation-related Features)
- (Academic Organization-related Features) * (Academic Organization-related Features)
- (Academic Organization-related Features) * (Nation-related Features)
- (Nation-related Features) * (Nation-related Features)

Since each group of features include various types of variables internally, it is necessary to verify what effect each combination might have in each setting, rather than making general hypothesis for each pair of groups of features. This will be discussed later in 5.1.1.3.

Chapter 5. Implementing the Assessment Model

As introduced in Chapter 4, this chapter addresses selection and construction of final explanatory variables (B-1) and logistic regression analysis (B-2) (See the introduction of Chapter 4 and Figure 4-1). Modeling is conducted herein both on two individual biopharmaceutical research topics: **Cas9** and **Microbiome**, and on the top five biopharmaceutical research topics combined (**Exosome**, **Microbiome**, **CRISPR**, **CAR-T** and **Cas9**) for comparison as well. One reason this thesis selected **Cas9** and **Microbiome** for comparative individual topics is that **Cas9** exhibited the largest frequency growth multiple (5.32) whereas **Microbiome** had the second least multiple (2.54) among the top five topics in Table 3.8, albeit similar network size of both topics (the node counts and edge counts of author citation networks: **Cas9**: 19,893 and 1,415,583 vs. **Microbiome**: 37,116 and 1,694,176, as seen in Table 3.9), which could provide distinguishing implications. **Cas9** and **Microbiome** are not considered to be scientifically related to each other, either (See 3.2.1 (ii)).

This chapter implements a logistic regression assessment model to assess (i) the importance of each explanatory variable on an event (Participant or Exit) for academic researchers with the relevant research topic, and (ii) the categorical probability of an event (Participant or Exit) occurring for each academic researcher, given a select number of continuous and categorical variables, i.e., startup readiness (See 1.1). Although the logistic regression model analyzes each academic researcher's categorical probability to be part of a startup participant class and a startup exit class, the goal of this thesis is not to predict whether an event (related to the researcher) will occur or not in the coming future by itself, but to assess and express each researcher's startup readiness state with a value between 0 and 1, as well as each variable's effect in terms of how many times the odds of each author's event increase associated with a one-unit increase in its distribution. In fact, the modeling that this thesis mainly tries to achieve is not so much *predictive modeling* as *explanatory modeling*, such that predictive modeling is defined as the process of applying a statistical model to data for the purpose of predicting new or future observations, whereas explanatory modeling is defined as the method of using a statistical model for testing relational explanations. Predictive modeling and explanatory modeling, however, can be considered to be two dimensions rather than extremes on continuum in that explanatory power and predictive accuracy are different qualities, a model will possess some level of each of them. [82]

Thus, this assessment model should be conducted in a consistent manner across all relevant researchers, irrespective of each researcher's difference in his/her actual state regarding the relevant event. With respect to the duration for which data for explanatory

variables is collected and arranged for each academic researcher, this thesis collected data of the common duration/timing corresponding to each type of explanatory variables, irrespective of the timing of an event (Participant or Exit) taking place for each researcher. The reason is that, since this research conducts explanatory modeling rather than predictive modeling, each observation regarding academic researchers should be assessed on a uniformed duration basis to analyze each researcher's individual prominence and capability. Incidentally, data sources used in this thesis, such as VentureSource and Web of Science Core Collection, will not allow us to collect data on explanatory variables such as Paper-related Features precisely prior to an event (Participant or Exit), hindering sufficient predictive modeling. We can neither tell the exact participation dates of academic researchers in startups, nor collect paper-related features on a daily or monthly basis, due to the limitation of VentureSource and Web of Science Core Collection.

Furthermore, throughout the specification of the logistic regression modeling in this chapter, validations regarding several assumptions associated with logistic regression analysis are conducted. Logistic regression analysis uses maximum likelihood estimation to estimate group membership. However, to interpret the results of probabilities regarding group membership, a preliminary analysis of the cleaned dataset was conducted to observe if the assumptions of logistic regression were met. The following are those assumptions.

Linearity of the Logit. One assumption of logistic regression is that the continuous predictors of the model are linear with the logit of the target variables. If the assumption is not valid, transformation of the variables should be considered [83]. Validations are presented in Table 5.1, Table 5.2 and Table 5.7.

Absence of Multicollinearity. A limitation of logistic regression is that it is sensitive to variables that have very high correlations with each other. Variables that are highly collinear often produce very large standard errors and inflated regression estimates [83] Therefore, the collinearity between the explanatory variables in the model had to be observed. A standard procedure that allows for this is the calculation of tolerance for each variable. The tolerance statistic is the calculation of the variance of each of the explanatory variables in the model not explained by all of the other explanatory variables in the model. A higher tolerance value suggests low levels of collinearity. Calculations are presented in Table 5.3

Absence of Small-Sample Bias Toward Variables. When there are too few cases in relation to the number of discrete variables, parameter estimates may inflate, which could produce large standard errors, and ultimately cause the model not to converge [83]. The problem is that maximum likelihood estimation of the logistic model suffers from

small-sample bias. Since the degree of bias is strongly dependent on the number of cases in the less frequent of the two categories, the problem is not the rarity of events specifically, but rather the possibility of a small number of cases on the rarer of the two outcomes [84]. Therefore, the cell counts were observed for each variable, and for each category of the categorical variables (See the numbers of YES (=1) of Participant and Exit in the Target Variable section on Appendices C-1, C-2 and C-3).

5.1. Preparing Variables

5.1.1. Selection and Construction of Explanatory Variables Related to Cas9 and Microbiome

The original data for the authors with research topics **Cas9** and **Microbiome**, on a name-based aggregation basis, were arranged as described in 4.1, with their two types of target variables: Participant and Exit that are explained in 4.3. Derived from the original data, explanatory variables described in 4.4 ((i) Individual Factors – Paper-related features and Patent-related features; (ii) Ecosystem Factors – Academic-organization-related features and Nation-related features and some of their (iii) Multi-variable Fractional Polynomials (MFPs) Factors, as well as the (iv) Interaction Terms Factors, that are possible pairs of the above explanatory variables), are prepared except for Hot Topic Factors. The reason for Hot Topic Factors not being considered herein was that there was no difference regarding the factors across authors since all authors belonged to the same research topic **Cas9** or **Microbiome**.

5.1.1.1. Stepwise Selection (1)

Firstly, this section conducted so-called ***stepwise regression*** (or ***stepwise selection***) in order to select the potential explanatory variables, specifically suited to the regression analysis of this thesis.

Stepwise regression consists of iteratively adding and removing predictors (explanatory variables) in the predictive model to find the subset of variables in the data set resulting in the best performing model, which is a model that lowers prediction error. Three strategies can be used for stepwise selection [85, 86].

- Forward selection, which starts with no predictors (explanatory variables) in the model, iteratively adds the most contributive predictors. It stops when the improvement is no longer significant.
- Backward selection (or backward elimination), which starts with all predictors in the model (full model), iteratively removes the least contributive predictors. It stops when one has a model where all predictors are significant.

- Stepwise selection (or sequential replacement), which is a combination of forward and backward selections, starts with no predictors, then sequentially adds the most contributive predictors (like forward selection). After adding each new variable, it removes any variable that no longer provides improvement in the model fit (similarly to backward selection).

Among these methods, this thesis applied stepwise selection based on Akaike information criterion (AIC) (Akaike, 1974) using the "step" function in R's "stats" package because the stepwise AIC method is a model selection method that can be extended widely to more generalized models and can be applied to non-normally distributed data. Essentially, AIC is a technique based on in-sample fit to estimate the likelihood of a model to predict/estimate the future values. A good model is one that has minimum AIC among all other models. The following equation is used to estimate the AIC of a model:

$$\text{AIC} = -2 \times \log(L) + 2 \times k \tag{5-1}$$

where L represents the likelihood value, and k denotes the number of estimated parameters [87, 88].

In results, the values of the AIC test static for the base models with no explanatory variables, for **Cas9** (1) Participant and (2) Exit were -11270.24 and -23924.82 respectively. For the final model, containing all variables chosen by the above stepwise selection procedure, the values of the AIC test statistics reduced to -11780 and -24280 respectively. The chosen explanatory variables were (1) eleven variables: PUB, IP_NUM, IP_CITED, FIRST_AUTH, CORRESP_AUTH, NATION_VC, PAPER_CITED_ NUM_IN_IP, NATION_STARTUP, CITATION_OUTDEG_CENT, IP_BINARY, and PAPER_CITED_BINARY_IN_IP, and (2) fourteen variables: PUB, IP_NUM, CORRESP_AUTH, NATION_VC, NATION_STARTUP, IP_CITING, COAUTH_ DEG_CENT, IP_CITED, CITATION_OUTDEG_CENT, FIRST_AUTH, UNIV_ RESEARCH, NATION_TURNOVER, IP_BINARY, and PAPER_CITED_ BINARY_ IN_IP, respectively. In a similar fashion, the values of the AIC test static for the base models with no explanatory variables, for **Microbiome** (1) Participant and (2) Exit were -22447.02 and -48245.44 respectively. For the final model, containing all variables chosen by the above stepwise selection procedure, the values of the AIC test statistics reduced to -25780 and -51730 respectively. The chosen explanatory variables were (1) thirteen variables: CORRESP_AUTH, IP_NUM, CITATION_OUTDEG_CENT, PAPER_CITED_BINARY_IN_IP, NATION_VC, FIRST_AUTH, NATION_ STARTUP, PAPER_CITED_NUM_IN_IP, UNIV_RESEARCH, UNIV_INNOV,

COAUTH_DEG_CENT, UNIV_SIZE and IP_CITED., and (2) eleven variables: CORRESP_AUTH, CITATION_OUTDEG_CENT, PAPER_CITED_BINARY_IN_IP, PAPER_CITED_NUM_IN_IP, PUB, NATION_VC, FIRST_AUTH, COAUTH_DEG_CENT, CITATION_INDEG_CENT, UNIV_INNOV and IP_BINARY respectively.

5.1.1.2. Multivariable fractional polynomials (MFPs)

Secondly, regarding *Linearity of the Logit* as discussed earlier in this chapter, for potential explanatory variables that are not clustered around a straight line, they were transformed into multivariable fractional polynomials (MFPs) using the MFP method with the “mfp” function in the “mfp” package of R. It selects the MFP model which best predicts the outcome [89]. This algorithm uses a form of backward elimination. It starts from a most complex permitted fractional polynomial (FP) model and attempts to simplify it by reducing the degrees of freedom (df). The selection algorithm is inspired by the so-called "closed test procedure": a sequence of tests with the "familywise error rate" or *P*-value maintained at a prespecified nominal value. The "closed test" algorithm for choosing an FP model with maximum permitted degree m=2 (4 df) for a single continuous predictor, x, is explained below.

- Inclusion: test the FP in x for possible omission of x (4 df test, significance level determined by select). If x is significant, then continue; otherwise drop x from the model.
- Nonlinearity: test the FP in x against a straight line in x (3 df test, significance level determined by alpha). If significant, then continue; otherwise the chosen model is a straight line.
- Simplification: test the FP with m=2 (4 df) against the best FP with m=1 (2 df) (2 df test at alpha level). If significant, then choose m=2; otherwise choose m=1. All significance tests are carried out using an approximate *P*-value calculation based on a difference in deviances (-2 x log likelihood) having a chi-squared or F distribution, depending on the regression in use. Therefore, each test in the procedure maintains a significance level only approximately equal to select. The algorithm is therefore not truly a closed procedure. However, for a given significance level, it does provide some protection against overfitting, which is against choosing over-complex MFP models.

Using the results obtained for the continuous explanatory variables for each target variable, this thesis constructed MFPs corresponding to them, instead of using their original potential variables. Matrixes including the best fractional polynomial powers for those variables are presented in Table 5.1 for academic regarding **Cas9**, and in Table 5.2 for those regarding **Microbiome**. If a variable’s *P*-value indicates significance at the 5%

level and if its corresponding power(s) is (are) not one, then we can transform the variable by raising it to its corresponding power(s) to create the corresponding MFP(s).

Table 5.1 MFP Transformation of Continuous Potential Explanatory Variables Using Closed Test Procedure for ***Cas9*** Academic Researchers

			Constructed MFP's)		
for Participant	p.lin [a,b]	p.FP [a,b]	power2	power4.1	power4.2
PUB	0.000 ***	0.699	0	0.5	0.5
NATION_VC	0.864	0.708	0.5	3	3
IP_BINARY	1.000	1.000	-2	-2	-2
CORRESP_AUTH	0.480	0.693	2	0.5	1
NATION_STARTUP	0.528	0.499	0.5	3	3
FIRST_AUTH	0.002 **	0.088 +	-2	0.5	2
IP_NUM	0.936	0.811	1	0.5	3
CITATION_OUTDEG_CENT	0.876	0.987	-1	1	2
IP_CITED	0.656	0.446	1	3	3
PAPER_CITED_NUM_IN_IP	0.520	0.622	0.5	0.5	3
PAPER_CITED_BINARY_IN_IP	1.000	1.000	-2	-2	-2
			Constructed MFP's)		
for Exit	p.lin [a,b]	p.FP [a,b]	power2	power4.1	power4.2
NATION_VC	0.427	0.259	2	-2	-2
NATION_STARTUP	0.329	0.463	0	3	3
PUB	0.000 ***	0.794	0	-2	0
IP_BINARY	1.000	1.000	-1	-2	-2
NATION_TURNOVER	0.134	0.073 +	2	3	3
COAUTH_DEG_CENT	0.521	0.495	2	3	3
CORRESP_AUTH	0.154	0.165	2	-2	1
CITATION_OUTDEG_CENT	0.444	0.694	-2	3	3
IP_CITING	0.542	0.505	0.5	0	3
UNIV_RESEARCH	0.915	0.893	2	3	3
PAPER_CITED_BINARY_IN_IP	1.000	1.000	-2	-2	-2
IP_NUM	0.709	0.830	-2	-2	3
FIRST_AUTH	0.011 *	0.053	-2	1	1
IP_CITED	0.701	0.564	0	-2	-1

a) +, *****, ******, and ******* respectively denote that the *P*-value is significant at 10%, 5%, 1%., and 0.1% .
b) **p.lin** corresponds to the test of nonlinearity and **p.FP** the test of simplification.
The maximum permitted degree (*m*) equals 1 when degrees of freedom (df) equal 2 on the fractional polynomial transformation, whereas *m* = 2 when df = 4.

Table 5.2 MFP Transformation of Continuous Potential Explanatory Variables Using Closed Test Procedure for ***Microbiome*** Academic Researchers

			Constructed MFP's)		
for Exit	p.lin [a,b]	p.FP [a,b]	power2	power4.1	power4.2
CORRESP_AUTH	0.000 ***	0.016 *	0	0	0.5
NATION_VC	0.017 *	0.457	-2	-2	-2
PAPER_CITED_BINARY_IN_IP	1.000	1.000	-2	-2	1
PAPER_CITED_NUM_IN_IP	0.687	0.907	0	-2	0.5
COAUTH_DEG_CENT	0.527	0.566	2	0.5	3
CITATION_OUTDEG_CENT	0.999	0.989	0.5	-2	3
FIRST_AUTH	0.592	0.728	-1	-2	3
PUB	0.074 +	0.152	0	-2	-2
IP_BINARY	1.000	1.000	-2	-2	-2
CITATION_INDEG_CENT	0.244	0.308	-2	3	3
UNIV_INNOV	0.905	0.888	2	-2	3
			Constructed MFP's)		
for Participant	p.lin [a,b]	p.FP [a,b]	power2	power4.1	power4.2
CORRESP_AUTH	0.000 ***	0.003 **	0	-0.5	0.5
PAPER_CITED_BINARY_IN_IP	1.000	1.000	-2	-2	-2
NATION_VC	0.546	1	-2	-2	3
UNIV_RESEARCH	0.299	0.297	0.5	2	2
NATION_STARTUP	0.000 ***	0.000 ***	-2.0	3	3
FIRST_AUTH	0.333	0.791	0	0.5	3
IP_NUM	0.567	0.441	0.5	-1	-1
CITATION_OUTDEG_CENT	0.976	0.902	1	-2	3
COAUTH_DEG_CENT	0.981	0.968	0.5	0.5	3
UNIV_SIZE	0.850	0.671	1	2	2
PAPER_CITED_NUM_IN_IP	0.780	0.790	0.5	0	3
UNIV_INNOV	0.653	0.443	1	3	3
IP_CITED	0.007 **	0.014 *	-0.500	2	3

a) b): as are the cases with Table 5.2

5.1.1.3. Interaction Terms

Thirdly, as discussed in 4.4.5, this thesis also created all feasibly possible interaction terms variables composed of the combinations among the above; for **Cas9** (1) eleven variables for Participant and (2) fourteen variables for Exit, two variables (PUB and FIRST_AUTH) of both of which were transformed into their MFPs, as well as for **Microbiome:** (1) thirteen variables for Participant, out of which three variables (CORRESP_AUTH, NATION_STARTUP and IP_CITED) were transformed into their MFPs, and (2) eleven variables for Exit, out of which two variables (CORRESP_AUTH and NATION_VC) were transformed into their MFPs. These interaction terms variables with each group of existing explanatory variables (which were partially transformed into MFPs as described above) were aggregated, to attain new sets of explanatory variables for further stepwise selection of variables as follows.

5.1.1.4. Stepwise Selection (2)

Fourthly, stepwise selection was conducted again with the above MFPs and Interaction Terms included in the explanatory variables. The values of the AIC test statistics regarding target variables (1) Participant and (2) Exit reduced to, for **Cas9**, (1) -11940 and (2) -24550, and, for **Microbiome**, (1) -26370 and (2) -50840 respectively, from their values for the base models with no explanatory variables, for **Cas9** (1) -11270.24 and (2) -23924.82 and for **Microbiome** (1) -22447.02 and (2) -48245.44 respectively. Except for **Mircobiome**'s Exit, these values showed considerable improvement from their prior values in 5.1.1.1 (for ***Cas9*** (1) -11780 and (2) -24280 and for **Microbiome** -25780 and -51730 respectively) that appeared without MFPs and interaction terms variables.

5.1.1.5. Addressing Multicollinearity and Referencing Correlations

Finally, multicollinearity using variance inflation factors (VIFs) is detected for each variable, as shown in Table 5.3.

Multicollinearity refers to a situation in which two or more explanatory variables in a regression model are closely linearly related. More commonly, the issue of multicollinearity arises when an approximate linear relation is found among two or more explanatory variables. Multivariate regression model with multicollinearity can indicate how well the entire bundle of predictors predicts the outcome variable, but it might not give a valid result for any individual predictor, or about which predictors are redundant with respect to others. In practice, we can detect multicollinearity using the variance inflation factor (VIF) as

$$\text{tolerance} = 1 - R_j^2, \quad \text{VIF} = \frac{1}{tolerance} \tag{5-2}$$

where R_j^2 is the coefficient of multiple correlation for the regression of explanatory variable j on all the other explanatory variables. The regression does not involve the target variable. In other words, R_j^2 is a measure of how well a given variable can be estimated using a linear function of a set of other variables and it is the correlation between the variable's values and the best estimate that can be computed linearly from the explanatory variables. In general, tolerances of less than 0.10 or VIFs of 10 or greater are often used to indicate a multicollinearity problem [90].

By applying the two techniques above, this thesis selected potential explanatory variables for the analyses both for Target Variable (i) Participant and (ii) Exit, both of which have main effect variables and interaction terms variables, as depicted with their VIFs in Table 5.3 for **Cas9** and Table 5.4 for **Microbiome**. These VIFs were computed using the "vif" function of R's "car" package. This thesis removed potential variables with VIFs larger than 10, or with correlation greater than 90% with other variables.

Table 5.3 VIFs of Cas9: Selected and Constructed Explanatory Variables

for Participant: Selected Main Explanatory Variables [a,b]	VIF
NATION_VC.c	1.200
NATION_STARTUP.c	1.330
FIRST_AUTH_MFP.c	1.676
CORRESP_AUTH.c	1.734
CITATION_OUTDEG_CENT.c	2.142
PUB_MFP.c	2.213
IP_NUM.c	2.530
PAPER_CITED_BINARY_IN_IP	3.054
IP_BINARY	3.267
PAPER_CITED_NUM_IN_IP.c	3.839
IP_CITED.c	5.111

for Participant: Constructed Explanatory Interaction Term Variables [a,b,c]	VIF
FIRST_AUTH_MFP.c:NATION_STARTUP.c	1.393
FIRST_AUTH_MFP.c:CORRESP_AUTH.c	1.416
FIRST_AUTH_MFP.c:NATION_VC.c	1.465
FIRST_AUTH_MFP.c:IP_BINARY	1.505
CORRESP_AUTH.c:IP_BINARY	2.150
IP_CITED.c:CORRESP_AUTH.c	4.286
IP_NUM.c:IP_CITED.c	4.747

for Exit: Selected Main Explanatory Variables [a,b]	VIF
NATION_TURNOVER.c	1.429
UNIV_RESEARCH.c	1.452
NATION_VC.c	1.501
NATION_STARTUP.c	1.637
CORRESP_AUTH.c	2.091
FIRST_AUTH_MFP.c	2.252
IP_NUM.c	2.690
PUB_MFP.c	2.929
COAUTH_DEG_CENT.c	3.382
CITATION_OUTDEG_CENT.c	3.822
PAPER_CITED_BINARY_IN_IP	5.552
IP_CITED.c	5.643
IP_BINARY	6.117

for Exit: Constructed Explanatory Interaction Term Variables [a,b,c]	VIF
NATION_STARTUP.c:FIRST_AUTH_MFP.c	1.669
IP_BINARY:UNIV_RESEARCH.c	1.840
NATION_VC.c:FIRST_AUTH_MFP.c	1.848
FIRST_AUTH_MFP.c:IP_BINARY	1.892
PUB_MFP.c:UNIV_RESEARCH.c	1.965
COAUTH_DEG_CENT.c:NATION_TURNOVER.c	2.299
FIRST_AUTH_MFP.c:CITATION_OUTDEG_CENT.c	3.028
COAUTH_DEG_CENT.c:IP_BINARY	3.439
CORRESP_AUTH.c:CITATION_OUTDEG_CENT.c	4.109
NATION_VC.c:IP_CITED.c	5.305
CORRESP_AUTH.c:PAPER_CITED_BINARY_IN_IP	6.266
CORRESP_AUTH.c:IP_BINARY	7.162
CITATION_OUTDEG_CENT.c:PAPER_CITED_BINARY_IN_IP	7.539
PUB_MFP.c:CITATION_OUTDEG_CENT.c	9.241
CITATION_OUTDEG_CENT.c:IP_BINARY	9.951

a) **".c"** indicates that variables were centered to mitigate multicollinearity, so that their means became zero. This indication is omitted from paper.
b) "_MFP" indicates that variables were turned into their multivariable fractional polynomial forms. Regarding Cas9, the same MFPs were applied both to Participant and Exit.
c) ":" indicates the multiplication between the first and the second variables to create Interaction Term Variables.

Table 5.4 VIFs of Microbiome: Selected and Constructed Explanatory Variables

for Participant: Selected Main Explanatory Variables [a,b]	VIF
CORRESP_AUTH_MFPp.c	1.097
CITATION_OUTDEG_CENT.c	1.691
PAPER_CITED_BINARY_IN_IP	1.547
FIRST_AUTH.c	1.450
NATION_VC.c	1.085
PAPER_CITED_NUM_IN_IP.c	2.300
UNIV_INNOV.c	2.310
UNIV_RESEARCH.c	2.103
UNIV_SIZE.c	1.970

for Participant: Constructed Explanatory Interaction Term Variables [a,b,c]	VIF
CORRESP_AUTH_MFPp.c:CITATION_OUTDEG_CENT.c	1.899
FIRST_AUTH.c:IP_NUM.c	1.453
CORRESP_AUTH_MFPp.c:PAPER_CITED_BINARY_IN_IP	1.197
IP_NUM.c:IP_CITED_MFPp.c	1.799
CORRESP_AUTH_MFPp.c:UNIV_INNOV.c	1.426
FIRST_AUTH.c:UNIV_RESEARCH.c	1.074
NATION_VC.c:UNIV_RESEARCH.c	1.236
CITATION_OUTDEG_CENT.c:UNIV_SIZE.c	1.229
CORRESP_AUTH_MFPp.c:UNIV_SIZE.c	1.481
UNIV_INNOV.c:UNIV_RESEARCH.c	1.922
PAPER_CITED_NUM_IN_IP.c:UNIV_SIZE.c	1.914
CITATION_OUTDEG_CENT.c:FIRST_AUTH.c	1.794

for Exit: Selected Main Explanatory Variables [a,b]	VIF
CORRESP_AUTH_MFPe.c	1.042
CITATION_OUTDEG_CENT.c	2.313
FIRST_AUTH.c	1.762
PAPER_CITED_BINARY_IN_IP	1.808
NATION_VC_MFPe.c	1.066
PAPER_CITED_NUM_IN_IP.c	1.779
UNIV_INNOV.c	1.120
CITATION_INDEG_CENT.c	1.794
COAUTH_DEG_CENT.c	2.095

for Exit: Constructed Explanatory Interaction Term Variables [a,b,c]	VIF
CORRESP_AUTH_MFPe.c:CITATION_OUTDEG_CENT.c	1.665
CORRESP_AUTH_MFPe.c:FIRST_AUTH.c	1.646
CITATION_OUTDEG_CENT.c:FIRST_AUTH.c	3.068
FIRST_AUTH.c:NATION_VC_MFPe.c	1.081
CITATION_OUTDEG_CENT.c:CITATION_INDEG_CENT.c	4.991
FIRST_AUTH.c:COAUTH_DEG_CENT.c	3.175
CITATION_INDEG_CENT.c:COAUTH_DEG_CENT.c	5.815
UNIV_INNOV.c:CITATION_INDEG_CENT.c	1.671
CORRESP_AUTH_MFPe.c:CITATION_INDEG_CENT.c	1.435

a) **".c"** indicates that variables were centered to mitigate multicollinearity, so that their means became zero. This indication is omitted from paper.
b) "_MFP" indicates that variables were turned into their multivariable fractional polynomial forms. _MFPp's and _MFPe's are specifically for Participant and Exit respectively.
c) ":" indicates the multiplication between the first and the second variables to create Interaction Term Variables.

In addition, by using the "cor" function of R's "stats" package, the correlations between original solo explanatory variables (Individual Factors and Ecosystem Factors, depicted in 4.4.1 and 4.4.3) regarding **Cas9** and **Microbiome** are computed in Table 5.5 and Table 5.6 respectively. Highly correlated pairs with correlations over 0.5 or less than -0.5 (highlighted in yellow in the tables) are as follows, none of which are observed in Table 5.3 and Table 5.4. For **Cas9**, they are PUB & PAPER_CITED (0.599), PUB & PAPER_CITING (0.557), PUB & CITATION_DEG_CENT (0.674), PUB & CITATION_INDEG_CENT (0.597), PUB & CITATION_OUTDEG_CENT (0.553), PAPER_CITED & CITATION_DEG_CENT (0.943), PAPER_CITED & CITATION_INDEG_CENT (1.000), PAPER_CITING & CITATION_DEG_CENT (0.943), PAPER_CITING & CITATION_OUTDEG_CENT (1.000), CITATION_ DEG_CENT & CITATION_INDEG_CENT (0.943), CITATION_DEG_CENT & CITATION_OUTDEG_CENT (0.614), IP_BINARY & IP_NUM (0.531), IP_BINARY & PAPER_CITED_BINARY_IN_IP (0.833), IP_NUM & PAPER_CITED_NUM_IN_ IP (0.746), IP_NUM & PAPER_CITED_BINARY_IN_IP (0.511), IP_NUM & IP_ CITED (0.758), IP_NUM & IP_CITING (0.640), PAPER_CITED_NUM_IN_IP & IP_ CITED (0.839), PAPER_CITED_NUM_ IN_IP & IP_CITING (0.779) and IP_CITED & IP_CITING (0.557). Similarly for **Microbiome**, any relevant pairs are PUB &

PAPER_CITED (0.600), PUB & PAPER_CITING (0.521), PUB & CITATION_DEG_CENT (0.641), PUB & CITATION_INDEG_CENT (0.596), PUB & CITATION_OUTDEG_CENT (0.515), PUB & COAUTH_DEG_CENT (0.559), PAPER_CITED & CITATION_DEG_CENT (0.944), PAPER_CITED & CITATION_INDEG_CENT (0.999), PAPER_CITED & COAUTH_DEG_CENT (0.583), PAPER_CITING & CITATION_DEG_CENT (0.720), PAPER_CITING & CITATION_OUTDEG_CENT (0.999), CITATION_DEG_CENT & CITATION_INDEG_CENT (0.944), CITATION_DEG_CENT & CITATION_OUTDEG_CENT (0.719), CITATION_DEG_CENT & COAUTH_DEG_CENT (0.628), CITATION_INDEG_CENT & COAUTH_DEG_CENT (0.585), IP_BINARY & PAPER_CITED_ BINARY_IN_IP (0.921), IP_NUM & IP_CITING (0.947), PAPER_CITED_NUM_IN_IP & PAPER_CITED_BINARY_IN_IP (0.536) and UNIV_SIZE & UNIV RESEARCH (0.600).

All of these combinations are composed of variables belonging to the same group of features (Individual Factors, Hot Topic Factors or Ecosystem Factors), which are expected results.

Table 5.5 Correlations Between Applied Original Explanatory Variables in Cas9

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
1	PUB	1																				
2	PAPER_CITED	0.599	1																			
3	PAPER_CITING	0.557	0.326	1																		
4	CORRESP_AUTH	0.247	0.117	0.164	1																	
5	FIRST_AUTH	0.305	0.166	0.248	0.064	1																
6	CITATION_DEG_CENT	0.674	0.943	0.617	0.149	0.221	1															
7	CITATION_INDEG_CENT	0.597	1.000	0.325	0.116	0.166	0.943	1														
8	CITATION_OUTDEG_CENT	0.553	0.322	1.000	0.163	0.247	0.614	0.321	1													
9	COAUTH_DEG_CENT	0.434	0.314	0.199	0.093	0.004	0.322	0.314	0.197	1												
10	IP_BINARY	0.304	0.271	0.213	0.113	0.149	0.295	0.271	0.211	0.130	1											
11	IP_NUM	0.451	0.400	0.226	0.134	0.106	0.399	0.398	0.223	0.204	0.531	1										
12	PAPER_CITED_NUM_IN_IP	0.373	0.407	0.174	0.082	0.073	0.389	0.407	0.171	0.175	0.231	0.746	1									
13	PAPER_CITED_BINARY_IN_IP	0.306	0.302	0.213	0.090	0.147	0.320	0.301	0.211	0.130	0.833	0.511	0.278	1								
14	IP_CITED	0.356	0.370	0.128	0.073	0.056	0.343	0.369	0.127	0.171	0.170	0.758	0.839	0.199	1							
15	IP_CITING	0.288	0.330	0.155	0.090	0.061	0.319	0.330	0.153	0.132	0.199	0.640	0.779	0.237	0.557	1						
16	UNIV_SIZE	0.081	0.032	0.044	0.040	0.023	0.040	0.032	0.044	0.092	0.016	0.006	-0.004	0.010	0.000	0.000	1					
17	UNIV_RESEARCH	0.098	0.145	0.033	-0.028	0.018	0.130	0.145	0.032	0.192	0.060	0.049	0.045	0.073	0.031	0.049	0.470	1				
18	UNIV_INNOV	0.058	0.012	0.040	0.046	0.021	0.023	0.012	0.039	0.041	0.021	0.013	-0.001	0.010	0.007	-0.004	0.317	0.330	1			
19	NATION_VC	0.071	0.114	0.046	-0.059	0.026	0.110	0.114	0.045	0.074	0.072	0.061	0.046	0.090	0.031	0.045	0.184	0.350	0.063	1		
20	NATION_STARTUP	-0.003	0.029	-0.001	-0.105	0.006	0.024	0.029	-0.001	0.023	-0.016	-0.001	0.018	0.011	0.012	0.016	0.091	0.210	0.033	0.330	1	
21	NATION_TURNOVER	0.051	0.074	0.033	0.166	-0.003	0.073	0.074	0.032	0.070	0.081	0.047	0.018	0.056	0.013	0.020	0.109	0.138	0.036	0.369	0.096	1

Table 5.6 Correlations Between Applied Original Explanatory Variables in Microbiome

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
1	**PUB**	**1**																				
2	**PAPER_CITED**	0.600	**1**																			
3	**PAPER_CITING**	0.521	0.457	**1**																		
4	**CORRESP_AUTH**	0.194	0.107	0.137	**1**																	
5	**FIRST_AUTH**	0.246	0.143	0.246	0.090	**1**																
6	**CITATION_DEG_CENT**	0.641	0.944	0.720	0.127	0.199	**1**															
7	**CITATION_INDEG_CENT**	0.596	0.999	0.456	0.103	0.142	0.944	**1**														
8	**CITATION_OUTDEG_CENT**	0.515	0.454	0.999	0.136	0.245	0.719	0.454	**1**													
9	**COAUTH_DEG_CENT**	0.559	0.583	0.498	0.101	0.066	0.628	0.585	0.494	**1**												
10	**IP_BINARY**	0.182	0.126	0.115	0.041	0.034	0.137	0.124	0.115	0.108	**1**											
11	**IP_NUM**	0.025	0.017	0.018	0.006	0.006	0.019	0.016	0.018	0.014	0.247	**1**										
12	**PAPER_CITED_NUM_IN_IP**	0.090	0.123	0.093	0.027	0.012	0.129	0.124	0.095	0.109	0.494	0.283	**1**									
13	**PAPER_CITED_BINARY_IN_IP**	0.170	0.128	0.114	0.035	0.026	0.138	0.125	0.114	0.108	0.921	0.250	0.536	**1**								
14	**IP_CITED**	0.196	0.121	0.075	0.035	0.015	0.121	0.121	0.075	0.066	0.396	0.222	0.220	0.362	**1**							
15	**IP_CITING**	0.053	0.039	0.042	0.013	0.007	0.045	0.039	0.042	0.031	0.313	0.947	0.436	0.326	0.251	**1**						
16	**UNIV_SIZE**	0.091	0.068	0.108	0.025	0.050	0.094	0.068	0.108	0.097	0.009	-0.005	-0.002	0.006	0.010	-0.007	**1**					
17	**UNIV_RESEARCH**	0.137	0.153	0.157	0.026	0.066	0.179	0.153	0.156	0.162	0.033	-0.003	0.015	0.029	0.029	-0.002	0.600	**1**				
18	**UNIV_INNOV**	0.101	0.100	0.101	0.032	0.039	0.116	0.100	0.101	0.104	0.019	0.005	0.029	0.017	0.008	0.005	0.334	0.401	**1**			
19	**NATION_VC**	0.100	0.072	0.041	-0.009	0.047	0.070	0.071	0.038	0.089	0.033	0.015	0.009	0.028	0.023	0.021	0.129	0.272	0.133	**1**		
20	**NATION_STARTUP**	0.029	0.037	0.045	-0.025	0.026	0.046	0.036	0.045	0.026	0.006	0.003	0.001	0.006	0.005	0.004	0.127	0.181	0.039	0.237	**1**	
21	**NATION_TURNOVER**	0.047	0.042	0.033	0.076	0.020	0.044	0.041	0.032	0.057	0.008	0.008	-0.004	0.008	0.015	0.007	0.083	0.190	0.120	0.432	0.199	**1**

5.1.1.6. Descriptive Statistics

Using the "describe" function of R's "psych" package, this thesis computed the descriptive statistics of the original potential variables and the final explanatory variables herein, as reported in Appendix C-1 and C-2. In these tables, 19611 and 35932 academic researchers with research topics **Cas9** and **Microbiome** are surveyed respectively as follows, according to their target variables and explanatory variables.

Regarding the dual target variables: Participant and Exit, this thesis measured them with binary variables defined as follows: Participant is a binary variable coded 1 if the researcher was found participating in (a) startup(s), and 0 otherwise; Exit is a binary variable coded 1 if the researcher was found to have exited (a) startup(s), and 0 otherwise. Among the whole researcher base in question, regarding **Cas9**, 669 researchers (3.41%) were found as startup participants, and 345 of them (1.76%) as those who have achieved (an) exit(s), while regarding **Microbiome**, 1164 (3.24%) as startup participants and 558 of them (1.55%) as those who have achieved (an) exit(s). Even though Participant and Exit are minorities among the researchers, these numbers are not as small as problematic rarity to cause the model to produce too large standard errors to converge.

With respect to explanatory variables, this thesis analyzed the original variables of **Cas9** and **Microbiome**, as well as their selected, applied variables in the proposed logistic regression model for this dissertation.

(i) Original variables composed of Individual Factors and Ecosystem Factors are grouped into two types of groups: categorical variables: Yes=1, No=0 (i.e., binary variables as seen in target variables), and continuous variables: integral or numerical. Herein, continuous variables are analyzed with descriptive statistics such as Minimum, Mean, Median, Standard Deviation (S.D.), Skewness and Kurtosis. Skewness is the degree of distortion from the symmetrical bell curve or the normal distribution. In other words, it measures the lack of symmetry in data distribution, and positive/negative skewness means the tail on the right/left side of the distribution is longer or fatter, respectively. Conventionally, if the skewness is greater than 1(positively skewed) or less than -1 (negatively skewed), the data is considered to be highly skewed. Kurtosis is a measure of whether the data is heavy-tailed or light-tailed relative to a normal distribution. In this descriptive statistics, the standard normal distribution has a kurtosis of 0, which has been standardized, so that data sets with kurtosis over/less than 0 are considered to have heavy/light tails, or outliers/lack of outliers, respectively. Regarding categorical variables: IP_BINARY and PAPER_CITED_BINARY_IN_IP, for **Cas9,** 1090 researchers (5.56%) were found as inventors of patents related to Cas9 and 769 of them (3.92%) as those who have cited

(a) paper(s) in such patents, while for **Microbiome**, 198 (0.55%) as inventors and 168 (0.47%) of them as those who have cited (a) paper(s) likewise. With respect to continuous variables, on the other hand, we attain 10 integral variables (PUB, PAPER_CITED, PAPER_CITING, CORRESP_AUTH, FIRST_AUTH, PAPER_CITED_NUM_IN_IP, IP_NUM, IP_CITED, IP_CITING, and UNIV_SIZE) and 9 numerical variables (CITATION_DEG_CENT, CITATION_INDEG_CENT, CITATION_ OUTDEG_CENT, COAUTH_DEG_CENT, UNIV_RESEARCH, UNIV_ INNOV, NATION_VC, NATION_STARTUP, NATION_ TURNOVER), descriptive statistics of all of which are to be computed.

(ii) To construct the assessment model, two types of variables are implemented as the selected, applied variables herein: solo variables derived from Individual Factors and Ecosystem Factors, some of which are turned into multivariable fractional polynomial (MFP) forms, and interaction terms factors, i.e., the products of such solo variables. To alleviate multicollinearity issues among these final explanatory variables and their interaction terms factors, this thesis centered solo continuous variables, i.e., variables other than categorical ones, so that their means become close to zero. Regarding categorical variables, IP_BINARY and PAPER_CITED_BINARY_IN_IP are employed for ***Cas9*** and ***Microbiome***. Continuous variables for both topics herein are all numerical, not integral, due to their centering, transformation to MFP's, and multiplication. Descriptive statistics was computed for all of the above variables.

5.1.2. Selection and Construction of Explanatory Variables Related to Five Biopharmaceutical Topics Combined (5-Biopharma-Topics)

This section analyzed data of the authors related to the top five biopharmaceutical topics combined: **Exosome**, **Microbiome**, **CRISPR**, **Cas9**, and **CAR-T**, as described in 4.1 (Referred to ***5-Biopharma-Topics*** hereafter), for their two types of target variables: Participant and Exit, as explained in 4.2 and 4.3 as in 5.1.1. In a manner different from 5.1.1, this thesis herein examined all the potential explanatory variables that are described in 4.4 including Hot Topic Factors (See B-1 of Figure 4-1). As opposed to 5.1.1 regarding individual research topics of **Cas9** and **Microbiome** respectively, for the analysis of the authors related to the five topics combined, Hot Topic Factors were added regarding each research topic, because these factors could help explain different results among the five topics.

5.1.2.1. Stepwise Selection (1)

Similar to the case in 5.1.1.1, for the authors regarding **5-Biopharma-Topics**, stepwise selection to select the potential explanatory variables was conducted.

In the results, the values of the AIC test static for the base models with no predictors (explanatory variables), for (1) Participant and (2) Exit were -56524 and -121832 respectively. For the final model, containing all variables chosen by the stepwise selection procedure, the values of the AIC test statistics reduced to -61031 and -126198 respectively. The chosen explanatory variables were (1) nineteen variables: CORRESP_AUTH, IP_BINARY, COAUTH_DEG_CENT, IP_NUM, NATION_VC, FINANCED_FREQ, FIRST_AUTH, IP_GROWTH, PUB, IP_CITING, IP_CITED, NATION_STARTUP, FINANCED_AMOUNT, KW_GROWTH, NATION_ TURNOVER, PAPER_CITED_BINARY_IN_IP, UNIV_RESEARCH, UNIV_INNOV, and PAPER_CITED_NUM_IN_IP, and (2) sixteen variables: CORRESP_AUTH, IP_CITING, NATION_VC, IP_BINARY, COAUTH_DEG_CENT, FINANCED_FREQ, FIRST_AUTH, KW_GROWTH, FINANCED_AMOUNT, IP_GROWTH, PUB, NATION_STARTUP, NATION_TURNOVER, UNIV_SIZE, PAPER_CITED_ BINARY_IN_IP, and IP_NUM, respectively.

5.1.2.2. Multivariable fractional polynomials (MFPs)

Secondly, as described in 5.1.1.2, *Linearity of the Logit* for potential explanatory variables regarding the authors with **5-Biopharma-Topics**, that are not clustered around a straight line were examined. This thesis transformed them into multivariable fractional polynomials (MFPs) using the MFP method exactly as described in 5.1.1.2.

Using the results obtained for the continuous explanatory variables for each target variable ((1) for Participant: CORRESP_AUTH, NATION_TURNOVER, NATION_VC, NATION_STARTUP, FIRST_AUTH, PUB, and COAUTH_DEG_CENT; (2) for Exit: CORRESP_AUTH, NATION_TURNOVER, NATION_STARTUP, PUB, and FIRST_AUTH, this thesis constructed MFPs corresponding to them, instead of using their original potential variables. A matrix including the best fractional polynomial powers for those variables is presented as Table 5.7. As demonstrated in 5.1.1.2, if a variable's *P*-value indicates significance at the 5% level and if its corresponding power(s) is (are) not one, then the variable is transformed by raising it to its corresponding power(s) to create the corresponding MFP(s).

Table 5.7 MFP Transformation of Continuous Potential Explanatory Variables Using Closed Test Procedure for Academic Researchers in 5- Biopharma-Topics

				Constructed MFP's)	
for Participant	p.lin	p.FP	power2	power4.1	power4.2
CORRESP_AUTH	0.000 ***	0.000 ***	0.5	-2	0.5
IP_BINARY	1.000	1.000	-2	-1	-1
NATION_TURNOVER	0.000 ***	0.000 ***	1	-2	-1
NATION_VC	0.000 ***	0.000 ***	-2	-2	-2
IP_GROWTH	1.000	1.000	-2	-2	1
FINANCED_FREQ	1.000	1.000	-2	-2	-2
FINANCED_AMOUNT	1.000	1.000	-2	-2	-2
NATION_STARTUP	0.001 **	0.042 *	-2	3	3
FIRST_AUTH	0.000 ***	0.367	-1	0.5	0.5
KW_GROWTH	1.000	1.000	-2	-2	-2
PUB	0.000 ***	0.000 ***	-0.5	-2	-2
IP_NUM	0.262	0.320	0.5	-2	0
UNIV_RESEARCH	0.112	0.119	3	-2	-2
PAPER_CITED_BINARY_IN_IP	1.000	1.000	-2	-0.5	3
IP_CITED	0.141	0.073 +	2	-1	-0.5
PAPER_CITED_NUM_IN_IP	0.259	0.551	0.5	-2	0
UNIV_INNOV	0.259	0.144	0.5	3	3
COAUTH_DEG_CENT	0.021 *	0.719	-0.5	-2	-1
IP_CITING	0.066 +	0.657	0	-0.5	0
				Constructed MFP's)	
for Exit	p.lin [a,b]	p.FP [a,b]	power2	power4.1	power4.2
CORRESP_AUTH	0.000 ***	0.000 ***	0.5	-0.5	0.5
NATION_TURNOVER	0.000 ***	0.000 ***	1	-2	-2
FINANCED_FREQ	1.000	1.000	-1	-2	0
FINANCED_AMOUNT	1.000	1.000	-2	-2	-0.5
NATION_VC	0.943	0.824	1	-2	3
IP_BINARY	1.000	1.000	1	-0.5	0.5
NATION_STARTUP	0.000 ***	0.000 ***	-2	3	3
KW_GROWTH	1.000	1.000	1	0	0.5
IP_GROWTH	1.000	1.000	-0.5	-1	1
PUB	0.000 ***	0.000 ***	-0.5	-2	-2
IP_CITING	0.323	0.586	0.5	0	0
FIRST_AUTH	0.018 *	0.888	-2	-2	3
PAPER_CITED_BINARY_IN_IP	1.000	1.000	-2	-0.5	0.5
UNIV_SIZE	0.138	0.171	0.5	1	2
IP_NUM	0.804	0.897	-0.5	0.5	3
COAUTH_DEG_CENT	0.252	0.350	-0.5	-0.5	3

a) **+, *, **,** and ******* respectively denote that the *P*-value is significant at 10%, 5%, 1%., and 0.1% .
b) **p.lin** corresponds to the test of nonlinearity and **p.FP** the test of simplification.
The maximum permitted degree (*m*) equals 1 when degrees of freedom (df) equal 2 on the fractional polynomial transformation, whereas *m* = 2 when df = 4.

5.1.2.3. Interaction Terms

Thirdly, this thesis created all feasibly possible interaction terms variables composed of the combinations among (1) for Participant: the above nineteen variables, some of which were transformed into their MFPs (i.e., MFPs corresponding to CORRESP_AUTH, NATION_TURNOVER, NATION_VC, NATION_STARTUP, FIRST_AUTH, PUB, and COAUTH_DEG_CENT), and (2) for Exit: the above sixteen variables, some of which were transformed into their MFPs (i.e., MFPs corresponding to CORRESP_AUTH, NATION_TURNOVER, NATION_STARTUP, PUB, and FIRST_AUTH), as demonstrated in 5.1.1.3. Then, this thesis created their interaction

terms variables by making pairs of each group of explanatory variables (which were partially transformed into MFPs as described above). Such newly combined interaction terms variables join explanatory variables for further stepwise selection as follows.

5.1.2.4. Stepwise Selection (2)

Fourthly, stepwise selection was conducted again, in the same manner as 5.1.1.4. The values of the AIC test static for the base models with no predictors (explanatory variables), for Target Variables (1) Participant and (2) Exit, were, again, -56524 and -121832 respectively. For the logistic regression model containing all the explanatory variables including MFPs and interaction terms variables that were chosen by the stepwise selection procedure herein, the values of the AIC test statistics reduced to (1) –62290 and (2) –127200 respectively, both of which show considerable improvement from their prior values ((1) -61031 and -126198 respectively) that appeared without MFPs and interaction terms variables.

5.1.2.5. Addressing Multicollinearity and Referencing Correlations

Finally, multicollinearity was detected using variance inflation factors (VIFs) for each variable. In the same way as described in 5.1.1.5, potential explanatory variables were selected both for Target Variable (i) Participant and (ii) Exit, both of which have solo variables and interaction terms variables, as depicted with their VIFs in Table 5.8. This thesis removed potential variables with VIFs larger than 10, or with correlation greater than 90% with other variables.

Table 5.8 VIFs of 5-Biopharma- Topics Selected and Constructed Explanatory Variables

(… Continued on Next Page)

for Participant: Selected Main Explanatory Variables [a,b]	VIF
FIRST_AUTH_MFPp.c	1.374
CORRESP_AUTH_MFPp.c	1.418
FINANCED_AMOUNT.c	1.498
NATION_VC_MFPp.c	1.585
UNIV_INNOV.c	1.635
COAUTH_DEG_CENT_MFPp.c	1.692
IP_GROWTH.c	2.117
PUB_MFPp.c	2.146
NATION_TURNOVER_MFPp.c	2.306
UNIV_RESEARCH.c	2.377
IP_BINARY	3.018
IP_NUM.c	3.400
NATION_STARTUP_MFPp.c	4.193
IP_CITED.c	4.617
IP_CITING.c	8.062

a) **".c"** indicates that variables were centered to mitigate multicollinearity, so that their means became zero. This indication is omitted from paper.

b) "_MFP" indicates that variables were turned into their multivariable fractional polynomial forms. _MFPp's and _MFPe's are specifically for Participant and Exit respectively.

c) ":" indicates the multiplication between the first and the second variables to create Interaction Term Variables.

for Participant: Constructed Explanatory Interaction Term Variables [a,b,c]	VIF
CORRESP_AUTH_MFPp.c:IP_GROWTH.c	1.092
CORRESP_AUTH_MFPp.c:FINANCED_AMOUNT.c	1.109
FIRST_AUTH_MFPp.c:KW_GROWTH.c	1.122
IP_GROWTH.c:PAPER_CITED_NUM_IN_IP.c	1.209
IP_GROWTH.c:UNIV_INNOV.c	1.226
NATION_VC_MFPp.c:PUB_MFPp.c	1.252
FIRST_AUTH_MFPp.c:UNIV_RESEARCH.c	1.284
IP_GROWTH.c:UNIV_RESEARCH.c	1.364
COAUTH_DEG_CENT_MFPp.c:UNIV_RESEARCH.c	1.378
CORRESP_AUTH_MFPp.c:COAUTH_DEG_CENT_MFPp.c	1.388
COAUTH_DEG_CENT_MFPp.c:NATION_STARTUP_MFPp.	1.399
PUB_MFPp.c:UNIV_INNOV.c	1.426
UNIV_RESEARCH.c:FINANCED_FREQ.c	1.464
IP_BINARY:FINANCED_AMOUNT.c	1.472
IP_GROWTH.c:NATION_TURNOVER_MFPp.c	1.541
FIRST_AUTH_MFPp.c:NATION_STARTUP_MFPp.c	1.566
NATION_VC_MFPp.c:FIRST_AUTH_MFPp.c	1.738
PUB_MFPp.c:UNIV_RESEARCH.c	1.757
NATION_TURNOVER_MFPp.c:NATION_STARTUP_MFPp.	1.776
CORRESP_AUTH_MFPp.c:PUB_MFPp.c	1.794
NATION_VC_MFPp.c:UNIV_RESEARCH.c	1.970
NATION_VC_MFPp.c:UNIV_INNOV.c	2.055
FIRST_AUTH_MFPp.c:PAPER_CITED_BINARY_IN_IP	2.179
IP_CITED.c:NATION_STARTUP_MFPp.c	2.270
PUB_MFPp.c:KW_GROWTH.c	2.289
IP_GROWTH.c:PUB_MFPp.c	2.382
NATION_STARTUP_MFPp.c:UNIV_INNOV.c	2.449
NATION_STARTUP_MFPp.c:UNIV_RESEARCH.c	2.472
IP_BINARY:NATION_STARTUP_MFPp.c	2.539
IP_BINARY:UNIV_RESEARCH.c	2.730
CORRESP_AUTH_MFPp.c:PAPER_CITED_BINARY_IN_IP	2.799
IP_NUM.c:NATION_STARTUP_MFPp.c	3.277
CORRESP_AUTH_MFPp.c:IP_CITED.c	3.423
NATION_TURNOVER_MFPp.c:KW_GROWTH.c	3.456
CORRESP_AUTH_MFPp.c:IP_NUM.c	3.568
FIRST_AUTH_MFPp.c:IP_CITED.c	3.813
IP_NUM.c:FIRST_AUTH_MFPp.c	3.882
FINANCED_AMOUNT.c:NATION_TURNOVER_MFPp.c	3.952
NATION_VC_MFPp.c:NATION_STARTUP_MFPp.c	4.015
IP_CITING.c:UNIV_INNOV.c	4.031
UNIV_INNOV.c:PAPER_CITED_NUM_IN_IP.c	4.414
IP_NUM.c:UNIV_RESEARCH.c	4.699
COAUTH_DEG_CENT_MFPp.c:PAPER_CITED_BINARY_IN_	5.031
IP_CITED.c:UNIV_RESEARCH.c	5.767
CORRESP_AUTH_MFPp.c:IP_CITING.c	5.916
IP_BINARY:COAUTH_DEG_CENT_MFPp.c	6.957
IP_NUM.c:COAUTH_DEG_CENT_MFPp.c	7.161
UNIV_RESEARCH.c:PAPER_CITED_NUM_IN_IP.c	7.831
IP_CITING.c:PAPER_CITED_NUM_IN_IP.c	7.922
COAUTH_DEG_CENT_MFPp.c:IP_CITING.c	8.153
CORRESP_AUTH_MFPp.c:PAPER_CITED_NUM_IN_IP.c	8.161

(Continued from Previous Page)

for Exit: Selected Main Explanatory Variables [a,b]	VIF
NATION_VC.c	1.100
FINANCED_AMOUNT.c	1.170
CORRESP_AUTH_MFPe.c	1.178
PUB_MFPe.c	1.189
UNIV_SIZE.c	1.277
FIRST_AUTH_MFPe.c	1.403
IP_BINARY	1.517
IP_CITING.c	1.737
IP_NUM.c	2.257

a) **".c"** indicates that variables were centered to mitigate multicollinearity, so that their means became zero. This indication is omitted from paper.
b) "_MFP" indicates that variables were turned into their multivariable fractional polynomial forms. _MFPp's and _MFPe's are specifically for Participant and Exit respectively.
c) ":" indicates the multiplication between the first and the second variables to create Interaction Term Variables.

for Exit: Constructed Explanatory Interaction Term Variables [a,b,c]	VIF
COAUTH_DEG_CENT.c:NATION_STARTUP_MFPe.c	1.080
FINANCED_AMOUNT.c:NATION_STARTUP_MFPe.c	1.108
FINANCED_AMOUNT.c:PUB_MFPe.c	1.136
CORRESP_AUTH_MFPe.c:UNIV_SIZE.c	1.140
FINANCED_AMOUNT.c:COAUTH_DEG_CENT.c	1.146
PUB_MFPe.c:UNIV_SIZE.c	1.218
IP_BINARY:FIRST_AUTH_MFPe.c	1.276
FIRST_AUTH_MFPe.c:NATION_STARTUP_MFPe.c	1.320
NATION_VC.c:FIRST_AUTH_MFPe.c	1.484
CORRESP_AUTH_MFPe.c:KW_GROWTH.c	1.656
CORRESP_AUTH_MFPe.c:FINANCED_AMOUNT.c	1.665
CORRESP_AUTH_MFPe.c:PAPER_CITED_BINARY_IN_IP	2.437
CORRESP_AUTH_MFPe.c:IP_CITING.c	3.243
CORRESP_AUTH_MFPe.c:IP_NUM.c	3.441
IP_CITING.c:IP_NUM.c	4.162

As in the case of 5.1.1.5, the correlations between the original solo explanatory variables (Individual Factors, Hot Topic Factors and Ecosystem Factors, depicted in 4.4.1 to 4.4.3) are also presented in Table 5.9. Highly correlated pairs with correlations over 0.5 or less than -0.5 (highlighted in yellow in the table) are as follows, none of which we observe in Table 5.8: PUB & PAPER_CITED (0.557), PUB & PAPER_CITING (0.514), PAPER_CITED & CITATION_DEG_CENT (0.619), PAPER_CITED & CITATION_INDEG_CENT (0.679), PAPER_CITING & CITATION_OUTDEG_CENT (0.618), CITATION_DEG_CENT & CITATION_ INDEG_CENT (0.930), CITATION_DEG_CENT & CITATION_ OUTDEG_CENT (0.652), CITATION_DEG_CENT & COAUTH_DEG_CENT (0.658), CITATION_ INDEG_CENT & COAUTH_DEG_CENT (0.564), CITATION_ OUTDEG_CENT & COAUTH_DEG_CENT (0.551), CITATION_OUTDEG_CENT & IP_GROWTH (0.528), COAUTH_DEG_CENT & IP_GROWTH (0.512), IP_BINARY & PAPER_ CITED_BINARY_IN_IP (0.857), IP_NUM & PAPER_CITED_NUM_IN_IP (0.626), IP_NUM & IP_CITED (0.626), IP_NUM & IP_CITING (0.596), PAPER_ CITED_NUM_IN_IP & IP_CITED (0.861), PAPER_CITED_NUM_IN_IP & IP_ CITING (0.870), IP_CITED & IP_CITING (0.680), UNIV_SIZE & UNIV_RESEARCH (0.517), FINANCED_FREQ & KW_GROWTH (-0.946). Unexpected discoveries among the above include positively highly correlated relationships between CITATION_OUTDEG_CENT & IP_GROWTH (0.528) and COAUTH_DEG_CENT & IP_GROWTH (0.512) and negatively highly correlated relationship between FINANCED_FREQ & KW_GROWTH (-0.946), the former of which occurred across Independent Factors and Hot Topic Factors whereas the latter within Ecosystem Factors.

Table 5.9 Correlations Between Applied Original Explanatory Variables in 5-Biopharma-Topics

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
1	PUB	1																								
2	PAPER_CITED	0.557	1																							
3	PAPER_CITING	0.514	0.377	1																						
4	CORRESP_AUTH	0.176	0.097	0.126	1																					
5	FIRST_AUTH	0.271	0.158	0.244	0.055	1																				
6	CITATION_DEG_CENT	0.415	0.619	0.415	0.057	0.121	1																			
7	CITATION_INDEG_CENT	0.381	0.679	0.226	0.050	0.091	0.930	1																		
8	CITATION_OUTDEG_CENT	0.323	0.204	0.618	0.054	0.134	0.652	0.332	1																	
9	COAUTH_DEG_CENT	0.348	0.266	0.203	0.027	0.013	0.658	0.564	0.551	1																
10	IP_BINARY	0.221	0.227	0.187	0.067	0.100	0.168	0.156	0.122	0.101	1															
11	IP_NUM	0.227	0.277	0.157	0.068	0.061	0.166	0.170	0.091	0.082	0.449	1														
12	PAPER_CITED_NUM_IN_IP	0.233	0.362	0.147	0.054	0.054	0.200	0.218	0.083	0.086	0.221	0.626	1													
13	PAPER_CITED_BINARY_IN_IP	0.222	0.247	0.183	0.057	0.096	0.178	0.170	0.120	0.105	0.857	0.434	0.258	1												
14	IP_CITED	0.212	0.312	0.111	0.052	0.039	0.166	0.185	0.060	0.066	0.158	0.626	0.861	0.179	1											
15	IP_CITING	0.203	0.325	0.135	0.052	0.045	0.183	0.197	0.080	0.078	0.204	0.596	0.870	0.237	0.680	1										
16	UNIV_SIZE	0.082	0.042	0.078	0.035	0.032	0.036	0.019	0.055	0.077	0.023	0.008	0.004	0.019	0.003	0.003	1									
17	UNIV_RESEARCH	0.118	0.141	0.099	0.011	0.039	0.067	0.071	0.031	0.072	0.067	0.036	0.037	0.068	0.026	0.041	0.517	1								
18	UNIV_INNOV	0.075	0.042	0.073	0.038	0.030	0.018	0.011	0.027	0.017	0.031	0.016	0.003	0.022	0.007	0.000	0.312	0.367	1							
19	NATION_VC	0.093	0.104	0.051	-0.020	0.038	0.097	0.102	0.041	0.082	0.061	0.039	0.031	0.068	0.021	0.032	0.158	0.301	0.091	1						
20	NATION_STARTUP	0.021	0.032	0.020	-0.047	0.022	0.024	0.027	0.008	-0.019	-0.004	0.003	0.010	0.010	0.007	0.009	0.105	0.186	0.032	0.277	1					
21	NATION_TURNOVER	0.054	0.065	0.057	0.097	0.007	0.072	0.062	0.057	0.057	0.062	0.031	0.013	0.046	0.010	0.016	0.113	0.178	0.091	0.394	0.161	1				
22	FINANCED_AMOUNT	0.003	0.004	0.009	0.001	-0.038	0.103	0.062	0.129	0.177	0.073	0.028	0.013	0.062	0.010	0.012	0.022	0.012	0.025	-0.053	-0.029	0.047	1			
23	FINANCED_FREQ	-0.092	-0.063	-0.143	-0.051	0.021	-0.163	-0.100	-0.207	-0.172	-0.118	-0.053	-0.033	-0.097	-0.024	-0.031	-0.054	-0.117	-0.048	-0.023	0.041	-0.110	-0.468	1		
24	KW_GROWTH	0.098	0.068	0.153	0.039	-0.014	0.161	0.099	0.205	0.155	0.110	0.051	0.032	0.089	0.024	0.031	0.055	0.123	0.048	0.041	-0.039	0.109	0.232	-0.946	1	
25	IP_GROWTH	0.084	0.047	0.105	-0.006	0.004	0.417	0.255	0.528	0.512	0.025	0.010	0.009	0.021	0.003	0.010	0.022	0.013	-0.006	0.067	-0.055	0.044	-0.153	-0.381	0.479	1

5.1.2.6. Descriptive Statistics

Just as in 5.1.1.6, the descriptive statistics of the original potential variables and the final explanatory variables for academic researchers with **5-Biopharma-Topics** were computed and presented in Appendix C-3. In this combined table, 94669 academic researchers related to **5-Biopharma-Topics** are surveyed.

With respect to the two target variables Participant and Exit, the table found 3156 startup participants (3.33%) and 1556 of them who have achieved exits (1.64%), out of the combined researcher base.

Regarding the explanatory variables, this thesis analyzed the **5-Biopharma-Topics** dataset's original variables as well as their selected, applied variables in the assessment model of this thesis, as performed in 5.1.1.6.

(i) Original variables are not only Individual Factors and Ecosystem Factors, but also Hot Topic Factors herein, which this thesis added since different research topics across emerging biopharmaceutical fields need to be dealt with, as opposed to 5.1.1.6 in which **Cas9** and **Microbiome** were addressed individually. These original variables are grouped into two types of categories in the same way as in 5.1.1.6: categorical variables: Yes=1, No=0 (i.e., binary variables as seen in target variables), and continuous variables: integral or numerical. Then, continuous variables are analyzed with the same descriptive statistics in the same manner: Minimum, Mean, Median, Standard Deviation (S.D.), Skewness and Kurtosis. Regarding categorical variables: IP_BINARY and PAPER_CITED_BINARY_IN_IP, 2884 researchers (3.05%) were found as inventors of patents relative to **5-Biopharma-Topics** and 2133 of them (2.25%) as those who have cited (a) paper(s) in such patents. Regarding continuous variables, on the other hand, we attain 10 integral variables (PUB, PAPER_CITED, PAPER_CITING, CORRESP_AUTH, FIRST_AUTH, PAPER_CITED_NUM_ IN_IP, IP_NUM, IP_CITED, IP_CITING, and UNIV_SIZE) and 13 numerical variables including four Hot Topic Factor additions additionally herein (CITATION_DEG_CENT, CITATION_INDEG_CENT, CITATION_OUTDEG_ CENT, COAUTH_DEG_CENT, FINANCED_AMOUNT, FINANCED_FREQ, KW_GROWTH, IP_GROWTH, UNIV_RESEARCH, UNIV_INNOV, NATION_ VC, NATION_ STARTUP, NATION_TURNOVER), descriptive statistics of all of which are to be computed.

(ii) In the same way as in 5.1.1.6, solo variables and interaction terms factors are arranged to complete the selected, applied variables for this thesis's assessment model. Some of those solo variables are turned into MFP forms and all variables are centered so that their means turn close to zero in order to mitigate multicollinearity issues

among those final explanatory variables. IP_BINARY is employed finally as the only categorical variable, while as many as 85 continuous either for Participant or Exit are employed, all of which are numerical in the same fashion as in 5.1.1.6. Descriptive statistics for all of the above variables are computed as well.

5.2. Preparing Models

Startup readiness is measured by the assessment models of this thesis, using two types of binary expressing target variables: (i) Participant and (ii) Exit, as depicted in 4.3, associated with the four categories of explanatory variables comprised of (1) Individual Factors, (2) Hot Topic Factors, (3) Ecosystem Factors and (4) Interaction Terms Factors, regarding **Cas9**, **Microbiome** and **5-Biopharma-Topics**. To assess authors' startup readiness regarding (i) Participant and (ii) Exit, models are designed and tested by configuring determinants in this section.

For the model regarding startup readiness in terms of Participant, target variables take a value of 1 when the author of related paper(s) appears as Participant in the VentureSource database, and 0 otherwise. Regarding startup readiness in terms of Exit, target variables take a value of 1 when the author appears as Participant who achieved an IPO or M&A in the VentureSource database and 0 otherwise.

As described in this section, although the authors belonging to Exit are part of those belonging to Participant, the startup readiness assessment models in terms of Participant and Exit are built independently, not in a stepwise fashion, constructing explanatory variables best suited for each target variables separately. The reason is that the purpose of building this Exit-oriented startup readiness assessment model is to assess the authors who have Exit potential in particular, not the authors who combine voluntary Participant potential and subsequent Exit potential. As argued in 2.1 and depicted in Figure 2-4, academic researchers' "Individual Factors" can be interpreted as the composition of (i) Essential Individual Factors: Knowledge Assets and Intellectual Property Assets represented by Paper-related Features and Patent-related Features, and (ii) other Individual Factors (such as Financial Assets, Social Capital Assets and Personal Assets for startup creation). Researchers with sufficient Essential Individual Factors but without other Individual Factors could be encouraged to participate in startups together with venture capital firms and management people, whose financial, social capital and personal assets could effectively complement and replace researchers' weak other Individual Factors, when relevant research topics are "hot". Since it is presumed that a significant portion of Participants herein initiated participation in startups due to such other Individual Factors, not thanks to their Essential Individual Factors which should be the key for Exit, it is reasonable to build the assessment models in terms of Participant and Exit independently for practical purposes.

5.2.1. Design and Test of Assessment Models Relating to Cas9 and Microbiome

5.2.1.1. Designing Assessment Models

After preparing the variables regarding academic researchers related to **Cas9** and **Microbiome** as depicted in 5.1.1, the estimated logistic regression assessment models for Participant and Exit were designed as following:

(a) ***Cas9***

for Participant

$$
\begin{aligned}
log\left(\frac{P_i}{1-P_i}\right) = & \beta_0 + \beta_1 IP\_NUM + \beta_2 PUB\_MFP + \beta_3 IP\_CITED + \beta_4 FIRST\_AUTH\_MFP \\
& + \beta_5 CORRESP\_AUTH + \beta_6 CITATION\_OUTDEG\_CENT + \beta_7 PAPER\_CITED\_NUM\_IN\_IP \\
& + \beta_8 NATION\_VC + \beta_9 NATION\_STARTUP + \beta_{10} IP\_BINARY \\
& + \beta_{11} PAPER\_CITED\_BINARY\_IN\_IP + \beta_{12} FIRST\_AUTH\_MFP * NATION\_STARTUP \\
& + \beta_{13} IP\_CITED * CORRESP\_AUTH + \beta_{14} IP\_NUM * IP\_CITED \\
& + \beta_{15} CORRESP\_AUTH * IP\_BINARY + \beta_{16} FIRST\_AUTH\_MFP * IP\_BINARY \\
& + \beta_{17} FIRST\_AUTH\_MFP * NATION\_VC + \beta_{18} FIRST\_AUTH\_MFP * CORRESP\_AUTH
\end{aligned}
\tag{5-3}
$$

where $\beta_i (i = 0, \ldots, 18)$ are the coefficients; the explanatory variables used in the model have been defined in 4.4. $log\left(\frac{P_i}{1-P_i}\right)$ is the logarithm of the ratio of the probability that an author *i* has become Participant relative to the probability that the same author has not become Participant.

for Exit

$$
\begin{aligned}
log\left(\frac{P_i}{1-P_i}\right) = & \beta_0 + \beta_1 IP\_NUM + \beta_2 PUB\_MFP + \beta_3 CORRESP\_AUTH + \beta_4 COAUTH\_DEG\_CENT \\
& + \beta_5 NATION\_VC + \beta_6 NATION\_STARTUP + \beta_7 IP\_CITED + \beta_8 FIRST\_AUTH\_MFP \\
& + \beta_9 CITATION\_OUTDEG\_CENT + \beta_{10} IP\_BINARY + \beta_{11} PAPER\_CITED\_BINARY\_IN\_IP \\
& + \beta_{12} NATION\_TURNOVER + \beta_{13} UNIV\_RESEARCH \\
& + \beta_{14} NATION\_STARTUP * FIRST\_AUTH\_MFP + \beta_{15} NATION\_VC * FIRST\_AUTH\_MFP \\
& + \beta_{16} CORRESP\_AUTH * CITATION\_OUTDEG\_CENT \\
& + \beta_{17} PUB\_MFP * CITATION\_OUTDEG\_CENT \\
& + \beta_{18} FIRST\_AUTH\_MFP * CITATION\_OUTDEG\_CENT \\
& + \beta_{19} COAUTH\_DEG\_CENT * IP\_BINARY + \beta_{20} FIRST\_AUTH\_MFP * IP\_BINARY \\
& + \beta_{21} CORRESP\_AUTH * IP\_BINARY \\
& + \beta_{22} CITATION\_OUTDEG\_CENT * PAPER\_CITED\_BINARY\_IN\_IP \\
& + \beta_{23} CORRESP\_AUTH * PAPER\_CITED\_BINARY\_IN\_IP
\end{aligned}
\tag{5-4}
$$

$+ \beta_{24} CITATION\_OUTDEG\_CENT * IP\_BINARY$

$+ \beta_{25}\ COAUTH\_DEG\_CENT\ \ *\ \ NATION\_TURNOVER\ \ + \beta_{26}\ IP\_BINARY\ \ *\ \ UNIV\_RESEARCH$

$+ \beta_{27} PUB\_MFP * UNIV\_RESEARCH\ + \beta_{28}.NATION\_VC * IP\_CITED$

where $\beta_i (i = 0, ..., 28)$ are the coefficients; the explanatory variables used in the model have been defined in 4.4.

Table 5.10 Estimated Logit Model of Variables Affecting Startup Participant in Cas9

	Explanatory Variables	Coefficients(β)	p-value [a]	
1	IP_NUM	0.127	0.007	**
2	PUB_MFP	1.176	0.000	***
3	IP_CITED	-0.002	0.551	
4	FIRST_AUTH_MFP	-3.579	0.000	***
5	CORRESP_AUTH	0.020	0.170	
6	CITATION_OUTDEG_CENT	-36.922	0.000	***
7	PAPER_CITED_NUM_IN_IP	0.001	0.175	
8	NATION_VC	1.224	0.000	***
9	NATION_STARTUP	-0.012	0.006	**
10	IP_BINARY	0.869	0.000	***
11	PAPER_CITED_BINARY_IN_IP	-0.419	0.078	+
12	FIRST_AUTH_MFP * NATION_STARTUP	0.049	0.000	***
13	IP_CITED * CORRESP_AUTH	0.000	0.218	
14	IP_NUM * IP_CITED	0.000	0.266	
15	CORRESP_AUTH * IP_BINARY	0.035	0.186	
16	FIRST_AUTH_MFP * IP_BINARY	1.433	0.000	***
17	FIRST_AUTH_MFP * NATION_VC	-3.452	0.001	***
18	FIRST_AUTH_MFP * CORRESP_AUTH	-0.073	0.027	*
	Likelihood Ratio Test			
	Number of Cases	19611		
	Likelihood Ratio (chi-square, deviance)	350.382		
	Degree of Freedom (d.f.)	18		
	p-value [a]	0.000	***	
	a) +,*,**,*** respectively denote that the variable is significant at 10%, 5%, 1% and 0.1%			

Results of the estimated logit models' explanatory variables are summarized in Table 5.10 and Table 5.11 for Participant and Exit respectively. Using the "glm" function of R's "stats" package, the logistic regression was computed in terms of whether or not the authors have become Participant/Exit to attain coefficients of all explanatory variables and their *P*-values.

Table 5.11 Estimated Logit Model of Variables Affecting Startup Exit in Cas9

	Explanatory Variables	Coefficients(β)	p-value [a]	
1	IP_NUM	0.077	0.164	
2	PUB_MFP	1.252	0.000	***
3	CORRESP_AUTH	0.010	0.611	
4	COAUTH_DEG_CENT	-8.072	0.929	
5	NATION_VC	1.388	0.001	***
6	NATION_STARTUP	-0.019	0.002	**
7	IP_CITED	0.004	0.200	
8	FIRST_AUTH_MFP	-3.818	0.002	**
9	CITATION_OUTDEG_CENT	-39.584	0.011	*
10	IP_BINARY	0.215	0.576	
11	PAPER_CITED_BINARY_IN_IP	0.303	0.478	
12	NATION_TURNOVER	0.070	0.003	**
13	UNIV_RESEARCH	-0.003	0.141	
14	NATION_STARTUP * FIRST_AUTH_MFP	0.054	0.000	***
15	NATION_VC * FIRST_AUTH_MFP	-4.446	0.004	**
16	CORRESP_AUTH * CITATION_OUTDEG_CENT	3.559	0.015	*
17	PUB_MFP * CITATION_OUTDEG_CENT	-36.029	0.023	*
18	FIRST_AUTH_MFP * CITATION_OUTDEG_CENT	-79.220	0.002	**
19	COAUTH_DEG_CENT * IP_BINARY	219.021	0.089	+
20	FIRST_AUTH_MFP * IP_BINARY	1.605	0.004	**
21	CORRESP_AUTH * IP_BINARY	0.061	0.278	
22	CITATION_OUTDEG_CENT * PAPER_CITED_BINARY_IN_IP	-65.519	0.061	+
23	CORRESP_AUTH * PAPER_CITED_BINARY_IN_IP	-0.075	0.225	
24	CITATION_OUTDEG_CENT * IP_BINARY	53.511	0.131	
25	COAUTH_DEG_CENT * NATION_TURNOVER	50.650	0.161	
26	IP_BINARY * UNIV_RESEARCH	-0.008	0.136	
27	PUB_MFP * UNIV_RESEARCH	0.009	0.020	*
28	NATION_VC * IP_CITED	-0.029	0.101	
	Likelihood Ratio Test			
	Number of Cases	19611		
	Likelihood Ratio (chi-square, deviance)	256.528		
	Degree of Freedom (d.f.)	28		
	p-value [a]	0.000	***	
	a) +,*,**,*** respectively denote that the variable is significant at 10%, 5%, 1% and 0.1%			

(b) ***Microbiome***

for Participant

$$
\begin{aligned}
log\left(\frac{P_i}{1-P_i}\right) = & \beta_0 + \beta_1 CORRESP\_AUTH\_MFPp + \beta_2 CITATION\_OUTDEG\_CENT \\
& + \beta_3 PAPER\_CITED\_BINARY\_IN\_IP + \beta_4 FIRST\_AUTH + \beta_5 NATION\_VC \\
& + \beta_6 PAPER\_CITED\_NUM\_IN\_IP + \beta_7 UNIV\_INNOV + \beta_8 UNIV\_RESEARCH \\
& + \beta_9 UNIV\_SIZE + \beta_{10} CORRESP_{AUTH_{MFPp}} * CITATION\_OUTDEG\_CENT \\
& + \beta_{11} FIRST\_AUTH * IP\_NUM + \beta_{12} CORRESP\_AUTH\_MFPp * PAPER\_CITED\_BINARY\_IN\_IP \\
& + \beta_{13} IP\_NUM * IP\_CITED\_MFPp + \beta_{14} CORRESP\_AUTH\_MFPp * UNIV\_INNOV \\
& + \beta_{15} FIRST\_AUTH * UNIV\_RESEARCH + \beta_{16} NATION\_VC * UNIV\_RESEARCH \\
& + \beta_{17} CITATION\_OUTDEG\_CENT * UNIV\_SIZE + \beta_{18} CORRESP\_AUTH\_MFPp * UNIV\_SIZE \\
& + \beta_{18} UNIV\_INNOV * UNIV\_RESEARCH + \beta_{20} PAPER\_CITED\_NUM\_IN\_IP * UNIV\_SIZE
\end{aligned}
\tag{5-5}
$$

$+\ \beta_{21} CITATION\_OUTDEG\_CENT * FIRST\_AUTH$

where β_i(i = 0, …, 21) are the coefficients; whose explanatory variables used in the model have been defined in 4.4.

for Exit

$$log\left(\frac{P_i}{1-P_i}\right) = \beta_0 + \beta_1 CORRESP\_AUTH\_MFPe + \beta_2 CITATION\_OUTDEG\_CENT$$

$$+\ \beta_3 FIRST\_AUTH + \beta_4 PAPER\_CITED\_BINARY\_IN\_IP$$
$$+\ \beta_5 NATION\_VC\_MFPe + \beta_6 PAPER\_CITED\_NUM\_IN\_IP + \beta_7 UNIV\_INNOV$$
$$+\ \beta_8 CITATION\_INDEG\_CENT + \beta_9 COAUTH\_DEG\_CENT$$
$$+\ \beta_{10} CORRESP\_AUTH\_MFPe * CITATION\_OUTDEG\_CENT$$
$$+\ \beta_{11} CORRESP\_AUTH\_MFPe * FIRST\_AUTH + \beta_{12} CITATION\_OUTDEG\_CENT * FIRST\_AUTH \quad (5\text{-}6)$$
$$+\ \beta_{13} FIRST\_AUTH * NATION\_VC\_MFPe$$
$$+\ \beta_{14}\ CITATION\_OUTDEG\_CENT * CITATION\_INDEG\_CENT$$
$$+\ \beta_{15} FIRST\_AUTH * COAUTH\_DEG\_CENT$$
$$+\ \beta_{16} CITATION\_INDEG\_CENT * COAUTH\_DEG\_CENT$$
$$+\ \beta_{17} UNIV\_INNOV * CITATION\_INDEG\_CENT$$
$$+\ \beta_{18} CORRESP\_AUTH\_MFPe * CITATION\_INDEG\_CENT$$

where $\beta_i(i = 0, ..., 18)$ are the coefficients; whose explanatory variables used in the model have been defined in 4.4.

Table 5.12 Estimated Logit Model of Variables Affecting Startup Participant in Microbiome

	Explanatory Variables	Coefficients(β)	p-value [a]	
1	CORRESP_AUTH_MFPp	1.275	0.000	***
2	CITATION_OUTDEG_CENT	-14.828	0.396	
3	PAPER_CITED_BINARY_IN_IP	1.978	0.000	***
4	FIRST_AUTH	-0.178	0.006	**
5	NATION_VC	0.777	0.000	***
6	PAPER_CITED_NUM_IN_IP	-0.009	0.251	
7	UNIV_INNOV	0.001	0.674	
8	UNIV_RESEARCH	0.007	0.000	***
9	UNIV_SIZE	0.000	0.001	***
10	CORRESP_AUTH_MFPp * CITATION_OUTDEG_CENT	-27.432	0.006	**
11	FIRST_AUTH * IP_NUM	-0.069	0.277	
12	CORRESP_AUTH_MFPp * PAPER_CITED_BINARY_IN_IP	-0.994	0.007	**
13	IP_NUM * IP_CITED_MFPp	0.000	0.073	+
14	CORRESP_AUTH_MFPp * UNIV_INNOV	-0.006	0.004	**
15	FIRST_AUTH * UNIV_RESEARCH	-0.005	0.011	*
16	NATION_VC * UNIV_RESEARCH	-0.026	0.000	***
17	CITATION_OUTDEG_CENT * UNIV_SIZE	-0.002	0.028	
18	CORRESP_AUTH_MFPp * UNIV_SIZE	0.000	0.003	**
19	UNIV_INNOV * UNIV_RESEARCH	0.000	0.029	*
20	PAPER_CITED_NUM_IN_IP * UNIV_SIZE	0.000	0.057	+
21	CITATION_OUTDEG_CENT * FIRST_AUTH	10.325	0.214	
	Likelihood Ratio Test			
	Number of Cases	35932		
	Likelihood Ratio (chi-square, deviance)	1310.905		
	Degree of Freedom (d.f.)	21		
	p-value [a]	0.000	***	
	a) +,*,**,*** respectively denote that the variable is significant at 10%, 5%, 1% and 0.1%			

Table 5.13 Estimated Logit Model of Variables Affecting Startup Exit in Microbiome

	Explanatory Variables	Coefficients(β)	p-value [a]	
1	CORRESP_AUTH_MFPe	0.538	0.000	***
2	CITATION_OUTDEG_CENT	-26.992	0.423	
3	FIRST_AUTH	-0.086	0.443	
4	PAPER_CITED_BINARY_IN_IP	2.798	0.000	***
5	NATION_VC_MFPe	-0.667	0.000	***
6	PAPER_CITED_NUM_IN_IP	-0.078	0.041	+
7	UNIV_INNOV	-0.004	0.060	+
8	CITATION_INDEG_CENT	24.647	0.015	*
9	COAUTH_DEG_CENT	-300.995	0.077	+
10	CORRESP_AUTH_MFPe * CITATION_OUTDEG_CENT	-29.651	0.010	**
11	CORRESP_AUTH_MFPe * FIRST_AUTH	-0.052	0.175	
12	CITATION_OUTDEG_CENT * FIRST_AUTH	-106.218	0.009	**
13	FIRST_AUTH * NATION_VC_MFPe	0.544	0.025	*
14	CITATION_OUTDEG_CENT * CITATION_INDEG_CENT	5099.086	0.005	**
15	FIRST_AUTH * COAUTH_DEG_CENT	686.972	0.000	***
16	CITATION_INDEG_CENT * COAUTH_DEG_CENT	-26835.897	0.011	*
17	UNIV_INNOV * CITATION_INDEG_CENT	1.128	0.000	***
18	CORRESP_AUTH_MFPe * CITATION_INDEG_CENT	-10.133	0.064	+
	Likelihood Ratio Test			
	Number of Cases	35932		
	Likelihood Ratio (chi-square, deviance)	1000.665		
	Degree of Freedom (d.f.)	18		
	p-value [a]	0.000	***	
	a) +,*,**,*** respectively denote that the variable is significant at 10%, 5%, 1% and 0.1%			

Results of the estimated logit models' explanatory variables are summarized in Table 5.12 and Table 5.13 for Participant and Exit respectively. The logistic regression in terms of whether or not the authors have become Participant/Exit is computed, to attain coefficients of all explanatory variables and their *P*-values.

5.2.1.2. Goodness of Fit Test

Moreover, the goodness of fit of this model was examined, to check how well it fits a set of observations. In general, measures of goodness of fit summarize the discrepancy between observed values and the values expected under the model in question. Such measures can be used in statistical hypothesis testing, i.e., to test for normality of residuals, to test whether outcome frequencies follow a specified distribution, in other words, to test the goodness of fit. Herein, This thesis conducted the following three statistical tests to assess whether a given distribution is suited to a dataset – (i) Chi-squared test, (ii) Hosmer-Lemeshow test and (iii) Osius-Rojek test [91, 92], and also alternatively assessed (iv) the area under the ROC curve (AUC).

(i) A chi-squared test, also written as χ^2 test, which is also called a likelihood ratio test, is any statistical hypothesis test where the sampling distribution of the test statistic is a chi-squared distribution when the null hypothesis is true, which is calculated with the following formula:

$$X^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \tag{5-7}$$

where O_i = *observed frequency*, E_i = *expected frequency*, k = *number of categories*. Using the "Anova" function of R's "car" package, the likelihood ratio, i.e., chi-square or Deviance is computed.

(ii) The Hosmer-Lemeshow test is a statistical test for goodness of fit for logistic regression models, especially for risk prediction models. The test assesses whether or not the observed event rates match expected event rates in subgroups of the model population. Data is first regrouped by ordering the predicted probabilities and forming the number of groups, g (which has been conventionally 10, but open for change). Models for which expected and observed event rates in subgroups are similar are called well calibrated, which is the null hypothesis of this test. Using the "logitgof" function of R's "generalhoslem" package, the Hosmer and Lemeshow test is computed. To test the null hypothesis that the data fit the specified model, the Hosmer-Lemeshow test statistic is calculated with the following formula:

$$\overline{P_k} = \sum_{i=1}^{i=n_k} \frac{n_i P_i}{n_k}, \qquad k = 1,2,\dots,g \tag{5-8}$$

$$C = \sum_{k=1}^{g} \frac{(y_k - n_k \overline{P_k})^2}{n_k \overline{P_k}(1 - \overline{P_k})} \tag{5-9}$$

The Hosmer and Lemeshow C statistic is based on: y_k, the number of observations where $y = 1$, n_k, the number of observations and P_k, the average probability in group k. This should follow a $chiSq$ (X^2) distribution with $g - 2$ degrees of freedom.

(iii) Osius and Rojek derived a large-sample normal approximation for the Pearson chi-square test statistics, which is usually referred to as the scaled Pearson chi-square. These are based on a *power-divergence* statistic $PD[l]$ ($l = 1$ for Pearsons test) and the standard deviation (herein, of a binomial distribution) SD. The statistic is:

$$Z_{OR} = \frac{PD_\lambda - \mu_\lambda}{\sigma_\lambda} \tag{5-10}$$

For logistic regression, it is calculated as:

$$Z_{OR} = \frac{P\chi^2 - (n - p)}{\sqrt{2(n - \sum_{i=1}^{n} \frac{1}{n_i}) + RSS}} \tag{5-11}$$

Where RSS is the residual sum-of-squares from a weighted linear regression:

$$\frac{1 - 2P_i}{\sigma_i} \sim X, \qquad \text{weights} = \sigma_i \tag{5-12}$$

Here X is the matrix of the model's explanatory variables. A two-tailed test against a standard normal distribution $N(0,1)$ should not be significant. Likewise in the Hosmer - Lemeshow test, the null hypothesis is that the data fit the specified model.

(iv) An ROC curve is the curve that we plot the *sensitivity* (true positive rate) against the *fall-out* (false positive rate), or one minus the *specificity* (true negative rate), at various sequential threshold points associated with the model. The area under the ROC curve (AUC) is also calculated, which means that the AUC value (e.g. 0.700, in other words 70.0%) of the time, the assessment model ranks a random positive example higher than a random negative example, respectively, which shows the assessment model's good assessment performance, i.e. goodness of fit in this regard. Conventionally, the AUC is used for potential selection of possibly optimal and suboptimal models and cost/benefit analysis of decision making. Using the "roc" function of R's "pROC" package, receiver operating characteristic (ROC) curves are presented for Participant and for Exit herein, respectively, to plot the *sensitivity* against the *fall-out* at various sequential probability cutoffs.

(a) ***Cas9***

for Participant

(i) Chi-squared test: Its calculated value of 350.382 is much larger than the critical value of the chi-squared statistic with 18 degrees of freedom at the 0.1% significance level. From this result, it is inferred that the null hypothesis, that all the parameter coefficients (except the intercept) are all zeros, is strongly rejected. Consequently, the model is significant at the 0.1% level.

(ii) The Hosmer-Lemeshow test: Here, the output returned $X^2 = 9.734$, Pr $(p - value)$ = 0.2842, which meant that the null hypothesis was retained. The null hypothesis is that the observed event rates match expected event rates in subgroups of the model population.

(iii) The Osius and Rojek test: The output returned that Osius-Rojek test Z = -0.2869789 with p-value = 0.7741284, which meant that the null hypothesis that the data fit the specified model, was retained.

(iv) Derived from probability cutoff thresholds and sensitivities/specificities, the ROC curve is constructed and its AUC was calculated as 0.6629 (Figure 5-1). This meant that 66.29% of the time for Participant, the assessment model ranked a random "Participant" author higher than a random "non-Participant" author, which shows acceptable classification performance of this model.

for Exit

(i) Chi-squared test: The calculated value was 256.528, which is way greater than the critical value of the chi-squared statistic with 28 degrees of freedom at the 0.1% significance level. Thus, the null hypothesis, that all the parameter coefficients (except the intercept) are all zeros, is strongly rejected. Consequently, the model is significant at the 0.1% level.

(ii) The Hosmer-Lemeshow test: The output returned $X^2 = 3.9482$, Pr $(p - value)$ = 0.8617, which meant that the null hypothesis was retained. The null hypothesis is that the observed event rates match expected event rates in subgroups of the model population.

(iii) The Osius and Rojek test: The output returned was, Z = -0.3318388 with p-value = 0.74001, which meant that the null hypothesis was retained. The null hypothesis is that the data fit the specified model.

(iv) Derived from probability cutoff thresholds and sensitivities/specifities, the ROC curve is constructed and its AUC was calculated as 0.7029 (Figure 5-2), which meant that 70.29% of the time for Participant, the model ranked a random "Exit" author higher than a random "non-Exit" author, which displays good classification performance.

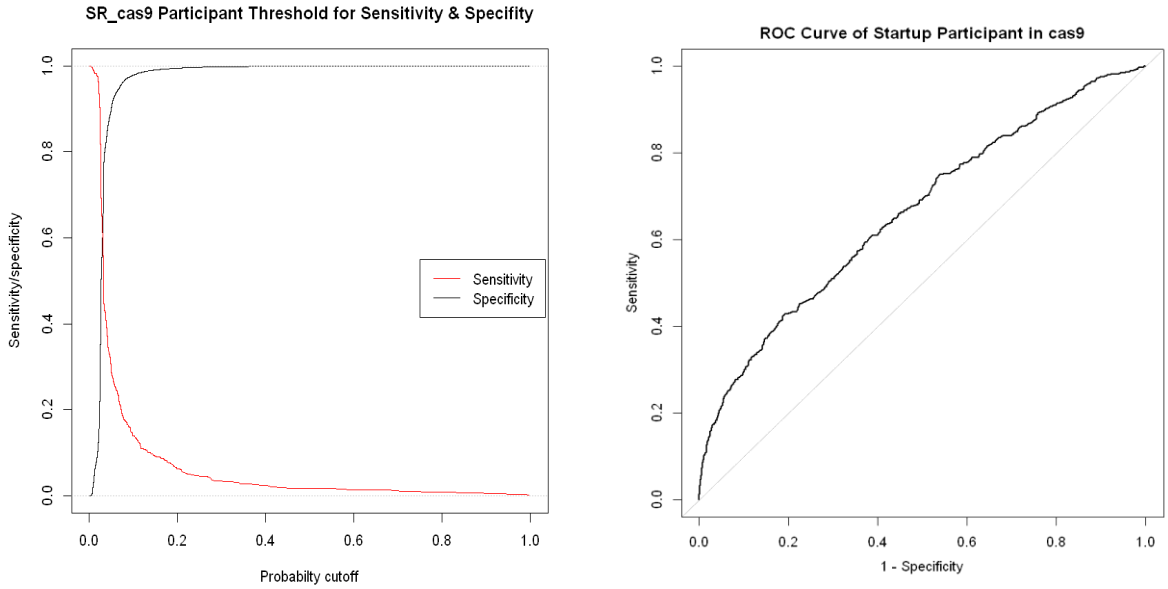


Figure 5-1 Sensitivity & Specifity vs. Probability Cutoff & ROC Curve for Cas9 Participants
AUC: 0.6629

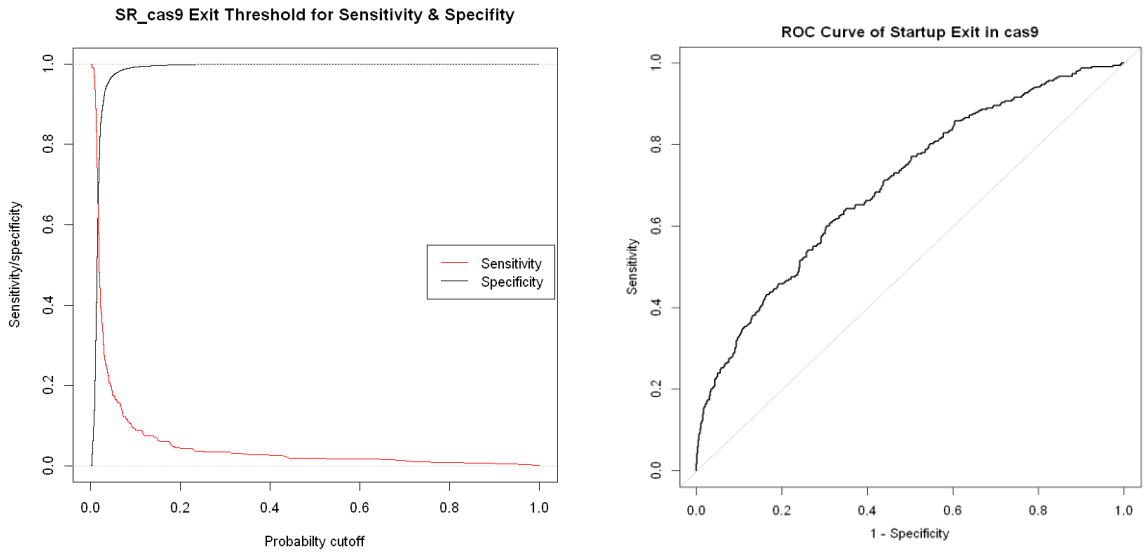


Figure 5-2 Sensitivity & Specifity vs. Probability Cutoff & ROC Curve for Cas9 Exits
AUC: 0.7029

(b) ***Microbiome***

for Participant

(i) The chi-squared test was calculated as 1310.905, much larger than the critical value of the chi-squared statistic with 21 degrees of freedom at the 0.1% significance level. Therefore, it is inferred that the null hypothesis, that all the parameter coefficients (except the intercept) are all zeros, is strongly rejected. Consequently, the model is significant at the 0.1% level.

(ii) The Hosmer-Lemeshow test: the output returned $X^2 = 43.667$ and $\Pr(p - value)$ = 0.000 with 8 degrees of freedom, which suggest that the null hypothesis that the observed event rates match expected event rates in subgroups of the model population, is strongly rejected. However, the Hosmer Lemeshow test has been criticized for several problems, such as lack of consideration of overfitting, and lack of guidance to selecting the number of subgroups. When the number of variables is large, small values for g give the test less opportunity to find mis-specifications. Larger values mean that the number of items in each subgroup may be too small to find differences between observed and expected values. As such, it seems inappropriate to reject the null hypothesis just because the Hosmer Lemeshow test suggests so.

(iii) The Osius and Rojek test: The output returned that Osius-Rojek test Z = -0.619 with p-value = 0.536, which meant that the null hypothesis that the data fit the specified model, was retained.

(iv) Probability cutoff thresholds and sensitivities/specificities lead to the ROC curve and its AUC that was calculated as 0.7407 (Figure 5-3). This meant that 74.07% of the time for Participant, the assessment model ranked a random "Participant" author higher than a random "non-Participant" author, which displays good classification performance.

for Exit

(i) Chi-squared test: The calculated value was 256.528, which is way greater than the critical value of the chi-squared statistic with 28 degrees of freedom at the 0.1% significance level. Thus, the null hypothesis, that all the parameter coefficients (except the intercept) are all zeros, is strongly rejected. Consequently, the model is significant at the 0.1% level.

(ii) The Hosmer-Lemeshow test: The output returned X^2 =9.139, $\Pr(p - value)$ = 0.3307 with 8 degrees of freedom, which meant that the null hypothesis, that the

observed event rates match expected event rates in subgroups of the model population, was retained for Exit, as opposed to the result for Participant.

(iii) The Osius and Rojek test: The output returned was, Z = 0.000 with p-value = 1.000, meaning that the null hypothesis, that the data fit the specified model, was retained.

(iv) The ROC curve was derived from probability cutoff thresholds and sensitivities/specificities, and its AUC was calculated as 0.7728 (Figure 5-4). In other words, 77.28% of the time for Exit, this model ranked a random "Exit" author higher than a random "non-Exit" author, whose classification performance was excellent.

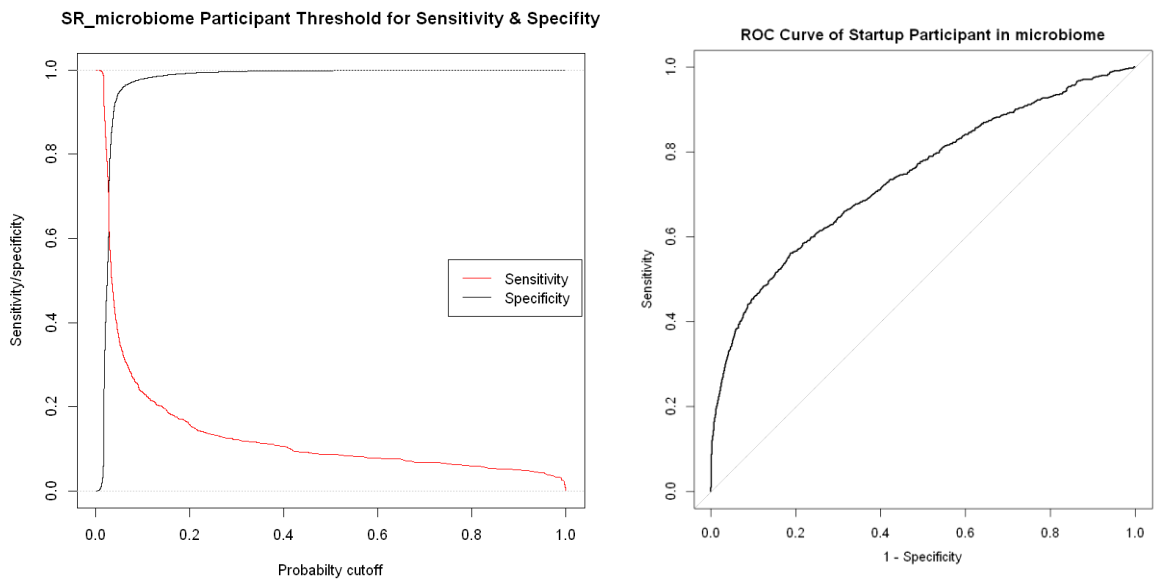


Figure 5-3 Sensitivity & Specifity vs. Probability Cutoff & ROC Curve for Microbiome Participants
AUC: 0.7407

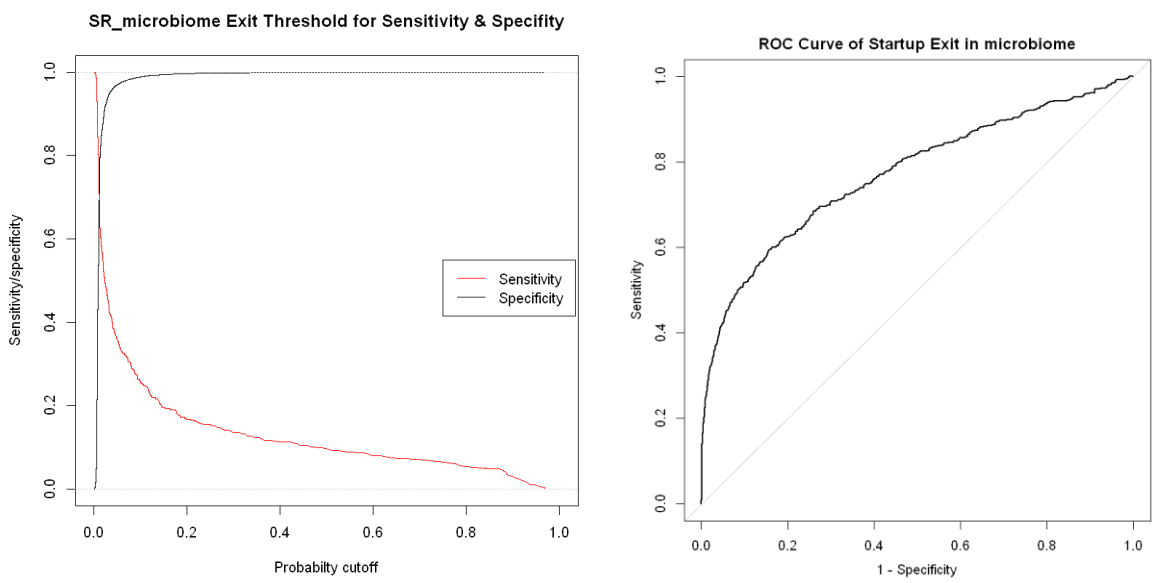


Figure 5-4 Sensitivity & Specifity vs. Probability Cutoff & ROC Curve for Microbiome Exits
AUC: 0.7728

5.2.2. Design and Test of Assessment Models Relating to 5-Biopharma-Topics

5.2.2.1. Designing Assessment Models

Given prepared variables regarding **5-Biopharma-Topics** as depicted in 5.1.1.6, the estimated assessment model for Participant was designed as following:

$$
\begin{aligned}
log\left(\frac{P_i}{1-P_i}\right) = & \beta_0 + \beta_1 CORRESP\_AUTH\_MFPp + \beta_2 IP\_BINARY + \beta_3 IP\_NUM \\
& + \beta_4 NATION\_VC\_MFPp + \beta_5 IP_{GROWTH} + \beta_6 PUB\_MFPp + \beta_7 FINANCED\_AMOUNT \\
& + \beta_8 FIRST\_AUTH\_MFPp + \beta_9 COAUTH\_DEG\_CENT\_MFPp + \beta_{10} IP\_CITED \\
& + \beta_{11} NATION\_TURNOVER\_MFPp + \beta_{12} NATION_{STARTUP_{MFPp}} + \beta_{13} UNIV_{RESEARCH} \\
& + \beta_{14} IP\_CITING + \beta_{15} UNIV_{INNOV} + \beta_{16} CORRESP_{AUTH_{MFPp}} * IP_{GROWTH} \\
& + \beta_{17} CORRESP\_AUTH\_MFPp * FINANCED\_AMOUNT \\
& + \beta_{18} IP\_BINARY * COAUTH\_DEG\_CENT\_MFPp + \beta_{19} IP\_NUM * FIRST\_AUTH\_MFPp \\
& + \beta_{20} COAUTH\_DEG\_CENT\_MFPp * PAPER\_CITED\_BINARY\_IN\_IP \\
& + \beta_{21} IP\_GROWTH * PAPER\_CITED\_NUM\_IN\_IP \\
& + \beta_{22} CORRESP\_AUTH\_MFPp * COAUTH\_DEG\_CENT\_MFPp \\
& + \beta_{23} NATION\_VC\_MFPp * PUB\_MFPp \\
& + \beta_{24} PUB\_MFPp * KW\_GROWTH + \beta_{25} FIRST\_AUTH\_MFPp * IP\_CITED \\
& + \beta_{26} CORRESP\_AUTH\_MFPp * PAPER\_CITED\_BINARY\_IN\_IP \\
& + \beta_{27} FIRST\_AUTH\_MFPp * PAPER\_CITED\_BINARY\_IN\_IP \\
& + \beta_{28} CORRESP\_AUTH\_MFPp * PAPER\_CITED\_NUM\_IN\_IP \qquad (5\text{-}13) \\
& + \beta_{29} CORRESP\_AUTH\_MFPp * PUB\_MFPp + \beta_{30} CORRESP\_AUTH\_MFPp * IP\_CITED \\
& + \beta_{31} IP\_GROWTH * PUB\_MFPp + \beta_{32} FINANCED\_AMOUNT * NATION\_TURNOVER\_MFPp \\
& + \beta_{33} FIRST\_AUTH\_MFPp * NATION\_STARTUP\_MFPp \\
& + \beta_{34} IP\_CITED * NATION\_STARTUP\_MFPp \\
& + \beta_{35} COAUTH\_DEG\_CENT\_MFPp * NATION\_STARTUP\_MFPp \\
& + \beta_{36} NATION\_TURNOVER\_MFPp * NATION\_STARTUP\_MFPp \\
& + \beta_{37} NATION\_VC\_MFPp * NATION\_STARTUP\_MFPp \\
& + \beta_{38} IP\_GROWTH * NATION\_TURNOVER\_MFPp \\
& + \beta_{39} KW\_GROWTH * NATION\_TURNOVER\_MFPp \\
& + \beta_{40} IP\_NUM * NATION\_STARTUP\_MFPp + \beta_{41} IP\_BINARY * NATION\_STARTUP\_MFPp \\
& + \beta_{42} FIRST\_AUTH\_MFPp * KW\_GROWTH + \beta_{43} IP\_NUM * UNIV\_RESEARCH \\
& + \beta_{44} IP\_BINARY * UNIV\_RESEARCH + \beta_{45} IP\_CITED * UNIV\_RESEARCH \\
& + \beta_{46} NATION\_VC\_MFPp * UNIV\_RESEARCH \\
& + \beta_{47} COAUTH\_DEG\_CENT\_MFPp * UNIV\_RESEARCH + \beta_{48} PUB\_MFPp * UNIV\_RESEARCH \\
& + \beta_{49} FINANCED\_FREQ * UNIV\_RESEARCH. + \beta_{50} FIRST\_AUTH\_MFPp * UNIV\_RESEARCH
\end{aligned}
$$

$+ \beta_{51} NATION\_VC\_MFPp * FIRST\_AUTH\_MFPp$
$+ \beta_{52} PAPER\_CITED\_NUM\_IN\_IP * UNIV\_RESEARCH$
$+ \beta_{53} PAPER\_CITED\_NUM\_IN\_IP * IP\_CITING + \beta_{54} CORRESP\_AUTH\_MFPp * IP\_CITING$
$+ \beta_{55} IP\_BINARY * FINANCED\_AMOUNT + \beta_{56} IP\_GROWTH * UNIV\_INNOV$
$+ \beta_{57} PUB\_MFPp * UNIV\_INNOV + \beta_{58} IP\_GROWTH * UNIV\_RESEARCH$
$+ \beta_{59} IP\_CITING * UNIV\_INNOV + \beta_{60} PAPER\_CITED\_NUM\_IN\_IP * UNIV\_INNOV$
$+ \beta_{61} CORRESP\_AUTH\_MFPp * IP\_NUM + \beta_{62} NATION\_STARTUP\_MFPp.c * UNIV\_INNOV$
$+ \beta_{63} NATION\_VC\_MFPp * UNIV\_INNOV + \beta_{64} NATION\_STARTUP\_MFPp * UNIV\_RESEARCH$
$+ \beta_{65} COAUTH\_DEG\_CENT\_MFPp * IP\_CITING + \beta_{66} IP\_NUM * COAUTH\_DEG\_CENT\_MFPp$

where $\beta_i (i = 0, ..., 66)$ are the coefficients for the explanatory variables.

Likewise, for Exit, the estimated assessment model was designed as follows:

$$\log\left(\frac{P_i}{1-P_i}\right) = \beta_0 + \beta_1 CORRESP\_AUTH\_MFPe + \beta_2 IP\_CITING + \beta_3 NATION\_VC$$
$$+ \beta_4 IP\_BINARY + \beta_5 FIRST\_AUTH\_MFPe + \beta_6 FINANCED\_AMOUNT + \beta_7 PUB\_MFPe$$
$$+ \beta_8 UNIV\_SIZE + \beta_9 IP\_NUM + \beta_{10} CORRESP\_AUTH\_MFPe * FINANCED\_AMOUNT$$
$$+ \beta_{11} CORRESP\_AUTH\_MFPe * KW\_GROWTH$$
$$+ \beta_{12} COAUTH\_DEG\_CENT * FINANCED\_AMOUNT$$
$$+ \beta_{13} IP\_BINARY * FIRST\_AUTH\_MFPe + \beta_{14} CORRESP\_AUTH\_MFPe * IP\_CITING$$
$$+ \beta_{15} CORRESP\_AUTH\_MFPe * UNIV\_SIZE + \beta_{16} PUB\_MFPe * UNIV\_SIZE \quad (5\text{-}14)$$
$$+ \beta_{17} FINANCED\_AMOUNT * PUB\_MFPe + \beta_{18} IP\_CITING * IP\_NUM$$
$$+ \beta_{19} CORRESP\_AUTH\_MFPe * IP\_NUM$$
$$+ \beta_{20} CORRESP\_AUTH\_MFPe * PAPER\_CITED\_BINARY\_IN\_IP$$
$$+ \beta_{21} FIRST\_AUTH\_MFPe * NATION\_STARTUP\_MFPe$$
$$+ \beta_{22} FINANCED\_AMOUNT * NATION\_STARTUP\_MFPe$$
$$+ \beta_{23} COAUTH\_DEG\_CENT * NATION\_STARTUP\_MFPe$$
$$+ \beta_{24} NATION\_VC * FIRST\_AUTH\_MFPe$$

where $\beta_i (i = 0, ..., 24)$ are the coefficients for the explanatory variables.

Results of the estimated logit models' explanatory variables herein are summarized in Table 5.14 and Table 5.15 for Participant and Exit respectively. Just like in the previous sections regarding **Cas9** and **Microbiome**, the logistic regression was computed in terms of whether or not authors have become Participant/Exit, to attain coefficients of all explanatory variables and their *P*-values.

Table 5.14 Estimated Logit Model of Variables Affecting Participant in 5-Biopharma-Topics

	Explanatory Variables	Coefficients(β)	p-value [a]	
1	CORRESP_AUTH_MFPp	0.681	0.000	***
2	IP_BINARY	0.800	0.000	***
3	IP_NUM	0.079	0.009	**
4	NATION_VC_MFPp	-0.433	0.000	***
5	IP_GROWTH	0.017	0.059	+
6	PUB_MFPp	-0.934	0.000	***
7	FINANCED_AMOUNT	-0.002	0.869	
8	FIRST_AUTH_MFPp	0.655	0.000	***
9	COAUTH_DEG_CENT_MFPp	0.018	0.471	
10	IP_CITED	0.005	0.051	+
11	NATION_TURNOVER_MFPp	0.000	0.039	*
12	NATION_STARTUP_MFPp	0.951	0.000	***
13	UNIV_RESEARCH	0.002	0.004	**
14	IP_CITING	0.003	0.081	+
15	UNIV_INNOV	-0.003	0.003	**
16	CORRESP_AUTH_MFPp * IP_GROWTH	-0.022	0.196	
17	CORRESP_AUTH_MFPp * FINANCED_AMOUNT	-0.088	0.000	***
18	IP_BINARY * COAUTH_DEG_CENT_MFPp	-0.589	0.002	**
19	IP_NUM * FIRST_AUTH_MFPp	0.172	0.182	
20	COAUTH_DEG_CENT_MFPp * PAPER_CITED_BINAR	0.543	0.003	**
21	IP_GROWTH * PAPER_CITED_NUM_IN_IP	0.000	0.064	+
22	CORRESP_AUTH_MFPp * COAUTH_DEG_CENT_MFP	0.006	0.811	
23	NATION_VC_MFPp * PUB_MFPp	0.539	0.016	*
24	PUB_MFPp * KW_GROWTH	-0.411	0.000	***
25	FIRST_AUTH_MFPp * IP_CITED	-0.003	0.633	
26	CORRESP_AUTH_MFPp * PAPER_CITED_BINARY_IN	-0.210	0.052	+
27	FIRST_AUTH_MFPp * PAPER_CITED_BINARY_IN_IP	0.870	0.043	*
28	CORRESP_AUTH_MFPp * PAPER_CITED_NUM_IN_II	-0.001	0.250	
29	CORRESP_AUTH_MFPp * PUB_MFPp	0.482	0.000	***
30	CORRESP_AUTH_MFPp * IP_CITED	0.005	0.004	**
31	IP_GROWTH * PUB_MFPp	-0.061	0.136	
32	FINANCED_AMOUNT * NATION_TURNOVER_MFPp	0.000	0.000	***
33	FIRST_AUTH_MFPp * NATION_STARTUP_MFPp	5.195	0.000	***
34	IP_CITED * NATION_STARTUP_MFPp	-0.108	0.000	***
35	COAUTH_DEG_CENT_MFPp * NATION_STARTUP_M	0.927	0.000	***
36	NATION_TURNOVER_MFPp * NATION_STARTUP_M	0.000	0.000	***
37	NATION_VC_MFPp * NATION_STARTUP_MFPp	-2.974	0.000	***
38	IP_GROWTH * NATION_TURNOVER_MFPp	0.000	0.000	***
39	NATION_TURNOVER_MFPp * KW_GROWTH	0.000	0.002	**
40	IP_NUM * NATION_STARTUP_MFPp	0.859	0.005	**
41	IP_BINARY * NATION_STARTUP_MFPp	-1.034	0.152	
42	FIRST_AUTH_MFPp * KW_GROWTH	0.131	0.106	
43	IP_NUM * UNIV_RESEARCH	0.003	0.008	**
44	IP_BINARY * UNIV_RESEARCH	-0.006	0.024	*
45	IP_CITED * UNIV_RESEARCH	0.000	0.005	**
46	NATION_VC_MFPp * UNIV_RESEARCH	0.011	0.000	***
47	COAUTH_DEG_CENT_MFPp * UNIV_RESEARCH	-0.002	0.014	*
48	PUB_MFPp * UNIV_RESEARCH	-0.008	0.009	**
49	UNIV_RESEARCH * FINANCED_FREQ	0.000	0.001	***
50	FIRST_AUTH_MFPp * UNIV_RESEARCH	0.005	0.076	*
51	NATION_VC_MFPp * FIRST_AUTH_MFPp	1.134	0.001	***
52	UNIV_RESEARCH * PAPER_CITED_NUM_IN_IP	0.000	0.003	**
53	IP_CITING * PAPER_CITED_NUM_IN_IP	0.000	0.089	+
54	CORRESP_AUTH_MFPp * IP_CITING	0.002	0.108	
55	IP_BINARY * FINANCED_AMOUNT	-0.050	0.241	
56	IP_GROWTH * UNIV_INNOV	-0.002	0.000	***
57	PUB_MFPp * UNIV_INNOV	-0.012	0.002	**
58	IP_GROWTH * UNIV_RESEARCH	0.000	0.057	+
59	IP_CITING * UNIV_INNOV	0.000	0.031	*
60	UNIV_INNOV * PAPER_CITED_NUM_IN_IP	0.000	0.069	+
61	CORRESP_AUTH_MFPp * IP_NUM	-0.050	0.086	+
62	NATION_STARTUP_MFPp * UNIV_INNOV	-0.036	0.000	***
63	NATION_VC_MFPp * UNIV_INNOV	-0.009	0.004	**
64	NATION_STARTUP_MFPp * UNIV_RESEARCH	0.013	0.032	*
65	COAUTH_DEG_CENT_MFPp * IP_CITING	0.003	0.066	+
66	IP_NUM * COAUTH_DEG_CENT_MFPp	-0.096	0.077	+
	Likelihood Ratio Test			
	Number of Cases	94669		
	Likelihood Ratio (chi-square, deviance)	2254.460		
	Degree of Freedom (d.f.)	66		
	p-value [a]	0.000	***	
	a) +,*,**,*** respectively denote that the variable is significant at 10%, 5%, 1% and 0.1%			

Table 5.15 Estimated Logit Model of Variables Affecting Exit in 5-Biopharma-Topics

	Explanatory Variables	Coefficients(β)	p-value [a]	
1	CORRESP_AUTH_MFPe	0.773	0.000	***
2	IP_CITING	0.002	0.015	*
3	NATION_VC	1.348	0.000	***
4	IP_BINARY	0.923	0.000	***
5	FIRST_AUTH_MFPe	0.442	0.000	***
6	FINANCED_AMOUNT	0.014	0.227	
7	PUB_MFPe	-0.653	0.000	***
8	UNIV_SIZE	0.000	0.298	
9	IP_NUM	0.040	0.133	
10	CORRESP_AUTH_MFPe * FINANCED_AMOUNT	0.039	0.004	**
11	CORRESP_AUTH_MFPe * KW_GROWTH	-0.316	0.000	***
12	FINANCED_AMOUNT * COAUTH_DEG_CENT	50.840	0.000	***
13	IP_BINARY * FIRST_AUTH_MFPe	0.652	0.017	*
14	CORRESP_AUTH_MFPe * IP_CITING	0.001	0.134	
15	CORRESP_AUTH_MFPe * UNIV_SIZE	0.000	0.004	**
16	PUB_MFPe * UNIV_SIZE	0.000	0.010	**
17	FINANCED_AMOUNT * PUB_MFPe	-0.043	0.248	
18	IP_CITING * IP_NUM	0.000	0.012	*
19	CORRESP_AUTH_MFPe * IP_NUM	0.054	0.013	*
20	CORRESP_AUTH_MFPe * PAPER_CITED_BINARY_IN_IP	-0.583	0.000	***
21	FIRST_AUTH_MFPe * NATION_STARTUP_MFPe	4.096	0.000	***
22	FINANCED_AMOUNT * NATION_STARTUP_MFPe	-0.244	0.001	***
23	COAUTH_DEG_CENT * NATION_STARTUP_MFPe	-254.432	0.025	*
24	NATION_VC * FIRST_AUTH_MFPe	-1.964	0.002	**
	Likelihood Ratio Test			
	Number of Cases	94669		
	Likelihood Ratio (chi-square, deviance)	1640.170		
	Degree of Freedom (d.f.)	24		
	p-value [a]	0.000	***	
	a) +,*,**,*** respectively denote that the variable is significant at 10%, 5%, 1% and 0.1%			

5.2.2.2. Goodness of Fit Test

As in 5.2.1.2, the goodness of fit of this statistical model was examined for the dataset regarding academic researchers related to **5-Biopharma-Topics** too, to check how well it fits a set of observations. The following are the results of Participants and Exits, with respect to (i) Chi-squared test, (ii) Hosmer-Lemeshow test and (iii) Osius-Rojek test, and (iv) the area under the ROC curve (AUC).

for Participant

(i) Chi-squared test: Its calculated value of 2254.46 is way larger than the critical value of the chi-squared statistic with 66 degrees of freedom at the 0.1% significance level. This result lets us infer that the null hypothesis, that all the parameter coefficients (except the intercept) are all zeros, is strongly rejected. Consequently, the model is significant at the 0.1% level.

(ii) The Hosmer-Lemeshow test: Here, the output returned X^2 =30.023, Pr $(p - value)$ = 0.000 with 8 degrees of freedom, when g, the number of subgroups, was set to conventional 10, which allegedly seemed that the null hypothesis that the observed event rates match expected event rates in subgroups of the model population, is strongly rejected. However, as discussed in 5.2.1.2 (ii), Hosmer Lemeshow test has been criticized for several deficiencies, in that it seems inappropriate to reject the null hypothesis just because the Hosmer Lemeshow test suggests so.

(iii) The Osius and Rojek test: The output returned that Osius-Rojek test Z = 0.000, with p-value = 1.000, which meant that the null hypothesis that the data fit the specified model, was retained.

(iv) Probability cutoff thresholds and sensitivities/specificities lead to the ROC curve and its AUC that was calculated as 0.6898 (Figure 5-5). In other words, 68.98% of the time, the model ranked a random "Participant" author higher than a random "non-Participant" author, whose classification performance is within an acceptable range.

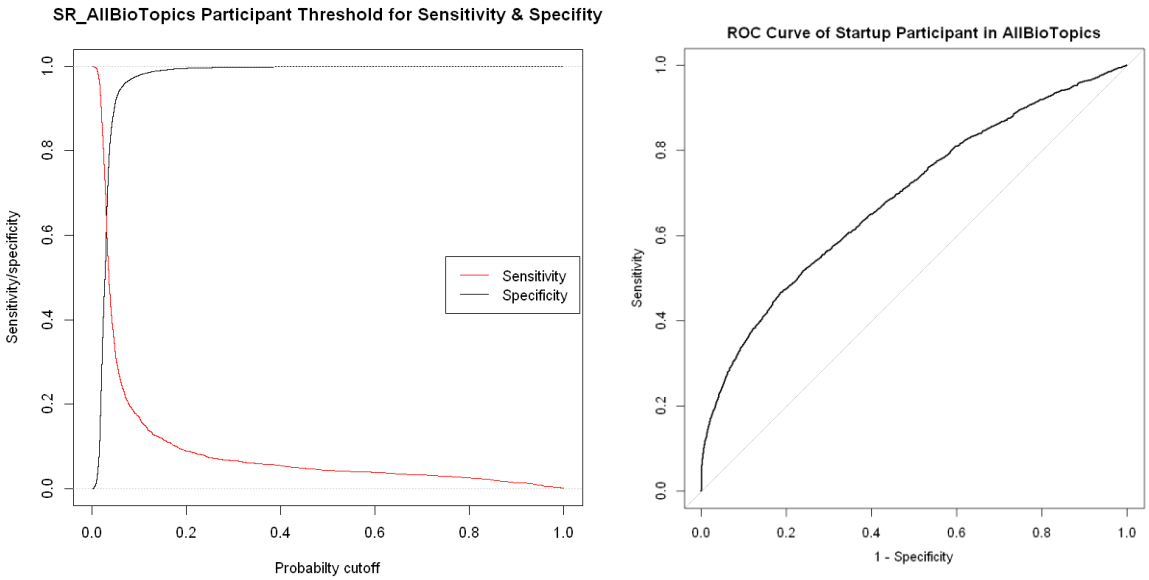


Figure 5-5 Sensitivity & Specifity vs. Probability Cutoff & ROC Curve for 5-Biopharma-Topics Participants AUC: 0.6898

for Exit

(i) Chi-squared test: The calculated value was 1640.17, which is much larger than the critical value of the chi-squared statistic with 24 degrees of freedom at the 0.1% significance level. Thus, the null hypothesis, that all the parameter coefficients (except the intercept) are all zeros, is strongly rejected. Consequently, the model is significant at the 0.1% level.

(ii) The Hosmer-Lemeshow test: The output returned $X^2 = 33.213$, Pr $(p - value)$ = 0.000 with 8 degrees of freedom seemingly meant that the null hypothesis that the observed event rates match expected event rates in subgroups of the model population, was rejected. As described above, this does not necessarily mean that the null hypothesis is not acceptable.

(iii) The Osius and Rojek test: The output returned was, Z = -0.018 with p-value = 0.986, which meant that the null hypothesis was retained. The null hypothesis is that the data fit the specified model.

(iv) The AUC was calculated as 0.7228 from probability cutoff thresholds and sensitivities/ specifities as shown in Figure 5-6, which means that the classification performance is that 72.28% of the time, the model ranked a random "Exit" author higher than a random "non-Exit" one, which is good.

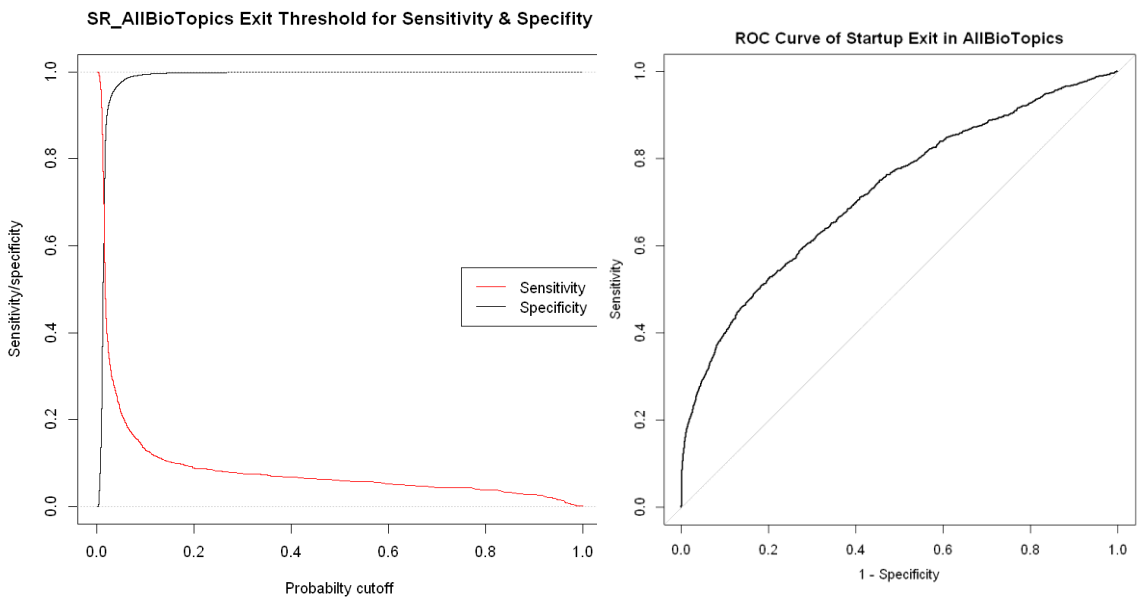


Figure 5-6 Sensitivity & Specifity vs. Probability Cutoff & ROC Curve for 5-Biopharma-Topics Exits
AUC: 0.7228

5.3. Assessing Startup Readiness

As introduced in 1.1, this dissertation defines startup readiness such that they can be calculated, expressed and assessed with a value between 0 and 1, determined as the probability of an academic researcher to be part of a startup participant class and a startup exit class, with 0 being the least startup preparedness and the least startup promise, while with 1 being the most such preparedness and such promise.

When assessing startup readiness of academic researchers, stakeholders such as venture capitalists and managerial talents might want to identify researchers with high startup readiness regarding the promise of success exemplified by exit potentials, who have not yet been involved in startups. They might want as many potentially startup-ready promising scientists to be identified as reasonably possible. On the other hand, academic researchers themselves might be interested in explanatory variables that could improve their startup readiness for success.

This chapter deals with how to assess startup readiness for such purposes.

5.3.1. Assessing Academic Researchers Related to Cas9 and Microbiome

5.3.1.1. Computation of Startup Readiness and Its Implication for Practical Use

By computing the assessment model shown in (5-3) and (5-4) for **Cas9** and (5-5) and (5-6) for **Microbiome** in 5.2.1.1, we can attain values of each author's startup readiness for Participant and Exit respectively, expressed between 0 and 1. In order to exemplify the practical use of such values, this section arranged top 30 startup readiness authors regarding **Cas9** and **Microbiome** in descending order for Participant and Exit with their computed startup readiness values, actual event status (i.e., Participant/Exit or not) and relevant event timing if any, as shown in Table 5.16.

In this way, we can identify and compare academic researchers with high values of startup readiness regarding Participant and Exit, albeit as mainly explanatory modeling. It is notable that, the lineups of highly ranked authors for Participant and Exit, together with those of positive ones (i.e., Participant or Exit), are considerably different from each other. We can also detect whether a relevant author is actually positive or not.

One implication is that, by computing startup readiness values of academic researchers for Participant and Exit, this assessment model potentially enables stakeholders such as venture capitalists and managerial talents not only to detect (i) the ones likely to be Participant and (ii) the ones likely to achieve Exit, but also to identify (iii) the ones who are promising in terms of Exit among researchers who are already Participant and (iv) the ones who are promising in terms of Exit among researchers who have not yet become Participant, as demonstrated in Table 5.16.

Table 5.16 Top 30 Startup Readiness Researchers (for Participant & Exit) in Cas & Microbiome
(Researchers Whose Family and Given Names in This Order Are Identified Only)
(Positive Researchers Highlighted in Yellow for Participant and in Orange for Exit)

Cas9

for Participant

Rank	Startup Readiness	Author	Participant	Est. Date
1	0.997	Liu, David	1	1/1/1997
2	0.997	Zhang, Feng	1	2/11/2004
3	0.983	Doudna, Jennifer	1	11/1/2013
4	0.967	Church, George	1	1/1/2009
5	0.870	Li, Jun	1	1/1/1996
6	0.869	Cowan, Chad	1	10/31/2013
7	0.720	Joung, Keith	1	11/1/2013
8	0.698	Porteus, Matthew	1	10/31/2013
9	0.696	May, Andrew	1	1/1/2011
10	0.651	Bumcrot, David	0	
11	0.612	Li, Li	1	1/1/2014
12	0.580	Charpentier, Emmanuelle	1	10/31/2013
13	0.559	Zhang, Yi	1	11/1/2007
14	0.540	Kim, Jin-Soo	0	
15	0.534	Mahfouz, Magdy	0	
16	0.523	Wang, Hui	0	
17	0.521	Gao, Caixia	0	
18	0.502	Jinek, Martin	0	
19	0.465	Donohoue, Paul	0	
20	0.456	Kim, Jungeun	0	
21	0.455	Zhang, Yong	0	
22	0.452	Khalili, Kamel	0	
23	0.447	Quake, Stephen	1	5/9/2003
24	0.444	Gersbach, Charles	0	
25	0.428	Li, Jing	0	
26	0.428	Liu, Jun	1	12/1/2002
27	0.422	Bengtsson, Niclas	0	
28	0.416	Wang, Haoyi	0	
29	0.407	Zhang, Lei	1	5/12/2005
30	0.406	Nureki, Osamu	1	11/1/2015

for Exit

Rank	Startup Readiness	Author	Exit	IPO Date	M&A Date	Participant	Est. Date
1	0.999	Zhang, Feng	1	2/3/2016		1	2/11/2004
2	0.999	Doudna, Jennifer	1	2/3/2016		1	11/1/2013
3	0.870	Liu, David	1	9/24/2008		1	1/1/1997
4	0.723	Church, George	1	2/3/2016	1/20/2017	1	1/1/2009
5	0.713	Bradner, James	1	6/30/2016	4/14/2014	1	1/1/2008
6	0.676	Li, Wei	1		4/26/2012	1	1/1/2007
7	0.580	Zhang, Yong	0			0	
8	0.534	Yang, Hui	1	10/28/2014		1	6/9/2008
9	0.446	Hu, Bian	0			0	
10	0.438	Jacks, Tyler	1	8/7/2014		1	1/1/2006
11	0.436	Wang, Yong	0			0	
12	0.431	Liu, Wei	1	5/19/2017		1	10/16/2001
13	0.421	Kim, Jin-Soo	0			0	
14	0.416	Wang, Ying	0			0	
15	0.397	Liu, Jun	1		10/27/2010	1	12/1/2002
16	0.379	Wang, Hui	0			0	
17	0.371	Musunuru, Kiran	0			0	
18	0.343	Wang, Xin	1	7/9/2014		1	6/8/1994
19	0.318	Li, Li	0			1	1/1/2014
20	0.315	Zhu, Jian-Kang	0			0	
21	0.311	Porteus, Matthew	1	10/19/2016		1	10/31/2013
22	0.310	Joung, Keith	0			1	1/25/2017
23	0.301	Li, Xiao-Jiang	0			0	
24	0.295	Cowan, Chad	0			1	10/31/2013
25	0.279	Yang, Huan	0			0	
26	0.273	Bengtsson, Niclas	0			0	
27	0.271	Li, Jian	0			0	
28	0.270	Zhao, Yu	0			0	
29	0.269	Zhao, Yunde	0			0	
30	0.253	Hahn, William	0			0	

Microbiome

for Participant

Rank	Startup Readiness	Author	Participant	Est. Date
1	0.975	Apte, Zachary	1	10/15/2012
2	0.974	Richman, Jessica	1	10/15/2012
3	0.838	Wang, Jianping	0	
4	0.727	Li, Laixing	0	
5	0.715	Bajaj, Jasnriohan	0	
6	0.673	Li, Lingzhi	0	
7	0.672	Lu, Tse-Yuan	0	
8	0.654	Quigley, Eamonn	1	1/1/1999
9	0.599	Li, Leyuan	0	
10	0.593	Li, Li	1	4/21/1998
11	0.585	Li, Lingling	0	
12	0.552	Dominguez-Bello, Maria	0	
13	0.542	Li, Liming	0	
14	0.529	Wang, Jia-Sheng	0	
15	0.526	Mills, David	1	1/1/1998
16	0.505	Lukiw, Walter	0	
17	0.503	Li, Lianshuo	0	
18	0.502	Kim, Ho, Cheol	0	
19	0.491	Wang, Joseph	0	
20	0.491	Wang, Jianxing	0	
21	0.484	Li, Huiying	0	
22	0.481	Li, Lu-Quan	0	
23	0.480	Li, Linlin	0	
24	0.479	Li, Lingyu	0	
25	0.477	Li, Longqing	0	
26	0.475	Kim, Helen	1	1/1/2002
27	0.471	Li, Lai-Xing	0	
28	0.470	Wang, Junshuai	0	
29	0.469	Li, Lei	0	
30	0.466	Bais, Harsh	0	

for Exit

Rank	Startup Readiness	Author	Exit	IPO Date	M&A Date	Participant	Est. Date
1	0.540	Clark, Andrew	1	1/7/2008		1	7/1/1998
2	0.522	Huang, Chun-Ming	0			0	
3	0.464	Wang, Yan	1		4/17/2014	1	1/1/1976
4	0.448	Bajaj, Jasnriohan	0			0	
5	0.446	Wolchok, Jedd	0			0	
6	0.435	Mizrahi, Itzhak	0			0	
7	0.401	Kim, Byung	0			0	
8	0.332	Lu, Tse-Yuan	0			0	
9	0.324	Wang, Jin	1	12/30/2014		1	1/1/2001
10	0.313	Li, Liming	0			0	
11	0.312	Wang, Jia-Sheng	0			0	
12	0.293	Li, Lu-Quan	0			0	
13	0.284	Li, Linlin	0			0	
14	0.282	Li, Lingyu	0			0	
15	0.282	Li, Leyuan	0			0	
16	0.281	Gootenberg, David	0			0	
17	0.279	Li, Li	1	5/6/2010		1	4/21/1998
18	0.277	Wang, Jessica	0			0	
19	0.276	Li, Longqing	0			0	
20	0.274	Kim, Hyungsuk	0			0	
21	0.269	Wang, Joseph	0			0	
22	0.269	Wang, Jianxing	0			0	
23	0.262	Berry, David	1	2/6/2014	7/8/2010	1	1/1/1998
24	0.259	Li, Lai-Xing	0			0	
25	0.252	Mack, David	1	10/24/2013	2/15/2005	1	1/1/2000
26	0.252	Moitinho-Silva, Lucas	0			0	
27	0.251	Kim, Hae, Jin	0			0	
28	0.249	Kim, Hyunho	0			0	
29	0.249	Li, Yuan	0			0	
30	0.247	Dominguez-Bello, Maria	0			0	

This model can be also simply used to measure startup readiness for Exit. Such usage is feasible even before researchers become Participant, in order to judge the proper timing of founding startups. Thus, another implication is that, this model can possibly urge academic researchers and their stakeholders, such as university administrators and policymakers, to understand their limitations of startup readiness regarding Exit in

relative terms early on, and to improve their features of explanatory variables upon or prior to engaging in or creating startups. Importance of each explanatory variable to startup readiness per researchers group will be examined in 5.3.1.2

As discussed in 2.1 and Figure 2-4, attentive business stakeholders can complement or even replace researchers' insufficient financial assets, social capital assets, personal assets that might have hindered researchers from engaging in startups without such professionals' help. Therefore, the proper use of this model could enable the formation and the development of promising academic startups in a mutually complementary fashion in that researchers are evaluated based on, and are enabled to focus on their Essential Individual Factors as scientists (i.e., Knowledge Assets and Intellectual Property Assets) in question, whereas business stakeholders play pivotal roles with respect to financing, management and business that are presumably not critical expertise for academic researchers as scientists.

5.3.1.2. Assessing Importance of Each Explanatory Variable

As presented in Table 5.10 and Table 5.11 for **Cas9** and Table 5.12 and Table 5.13 for **Microbiome**, it is apparent which variables increase and decrease when the likelihood that authors become Participants/Exits increases, respectively, but the results shown in the tables are based on the signs and significance of the coefficients of the explanatory variables. They do not incorporate the scope of these coefficients because, in the logistic functional form upon which logistic recession is based, the estimated values of coefficients, as those presented in these tables, use different scales. We cannot compare them directly. They are not coefficients reflecting the marginal effects of the explanatory variables. To compare the scope of the impact of these explanatory variables, we have standardized their coefficients so that they are based on the same scale and can explain each author's likelihood of being a Participant/Exit. Moreover, we attained exponential transformation of these standardized coefficients, or, raised e (the base of the natural logarithm) to the power of standardized β_j(for Participant, j = 1, …, 18, and for Exit, j = 1, …, 28), so that we can usefully detect and compare the effects of each variable in terms of how many times the odds of each author's being a Participant/Exit (i.e., $\frac{P_i}{1-P_i}$) increase associated with a one-unit (i.e., one S.D.) increase in the exposure, across all the relevant explanatory variables. In other words, the exponentially transformed coefficients means the growth multiple of the odds per one S.D. increase for each variable.

By using the "function" function of the "base" package and the "ggplot" function of the "ggplot2" package both in R, the exponentially transformed standard coefficients of

explanatory variables were computed, as shown in Table 5.17 and Table 5.18 for ***Cas9***, in Table 5.19 and Table 5.20 for **Microbiome**, for Participant and for Exit respectively.

(a) **Cas9**

for Participant (Table 5.17)

PUB_MFP is the most influential at 1.483, next to which FIRST_AUTH_MFP, NATION_VC, and IP_BINARY are similarly influential at 1.244, 1.223 and 1.220 respectively. Both PUB_MFP and FIRST_AUTH_MFP represent MFPs (Multi-variable Fractional Polynomials, See 4.4.4 and 5.1.1.2) transformed from original PUB and FIRST_AUTH respectively. Subsequently, FIRST_AUTH_MFP * IP_BINARY, FIRST_AUTH_MFP * NATION_STARTUP, and IP_NUM follows at 1.149, 1.120 and 1.117 respectively. In contrast, the exponentially transformed coefficients of several variables are less than 1, which means that their increase negatively affects the odds. For example, CITATION_OUTDEG_CENT and FIRST_AUTH_MFP * NATION_VC are negatively influential against the odds with their transformed coefficients at 0.817 and 0.849 respectively, the latter of which seemingly works as weight to some extent against FIRST_AUTH_MFP and NATION_VC, both of which have positive effect individually.

Table 5.17 Effects of Explanatory Variables on Odds for Participant in Cas9

	Explanatory Variables [a, b, c]	exp_coef [d]	p_value [e]	
1	PUB_MFP.c	1.483	0.000	***
2	IP_CITED.c * CORRESP_AUTH.c	1.299	0.218	
3	FIRST_AUTH_MFP.c	1.244	0.000	***
4	NATION_VC.c	1.223	0.000	***
5	IP_BINARY	1.220	0.000	***
6	FIRST_AUTH_MFP.c * IP_BINARY	1.149	0.000	***
7	FIRST_AUTH_MFP.c * NATION_STARTUP.c	1.120	0.000	***
8	IP_NUM.c	1.117	0.007	**
9	PAPER_CITED_NUM_IN_IP.c	1.070	0.175	
10	CORRESP_AUTH.c	1.058	0.170	
11	CORRESP_AUTH.c * IP_BINARY	1.038	0.186	
12	IP_CITED.c	0.935	0.551	
13	FIRST_AUTH_MFP.c * CORRESP_AUTH.c	0.935	0.027	
14	PAPER_CITED_BINARY_IN_IP	0.922	0.078	+
15	NATION_STARTUP.c	0.907	0.006	**
16	FIRST_AUTH_MFP.c * NATION_VC.c	0.849	0.001	***
17	CITATION_OUTDEG_CENT.c	0.817	0.000	***
18	IP_NUM.c * IP_CITED.c	0.755	0.266	

a) "_MFP" indicates that these variables were turned into their multivariable fractional polynomial forms. Regarding Cas9, the same MFPs were applied both to Participant and Exit.

b) ".c" indicates that these variables were centered from their originals, i.e. adjusted so that their means became zero, although centralization does not affect the values of exp_coef and p_value herein. This indication is omitted from the paper.

c) ***Paper-related Features***

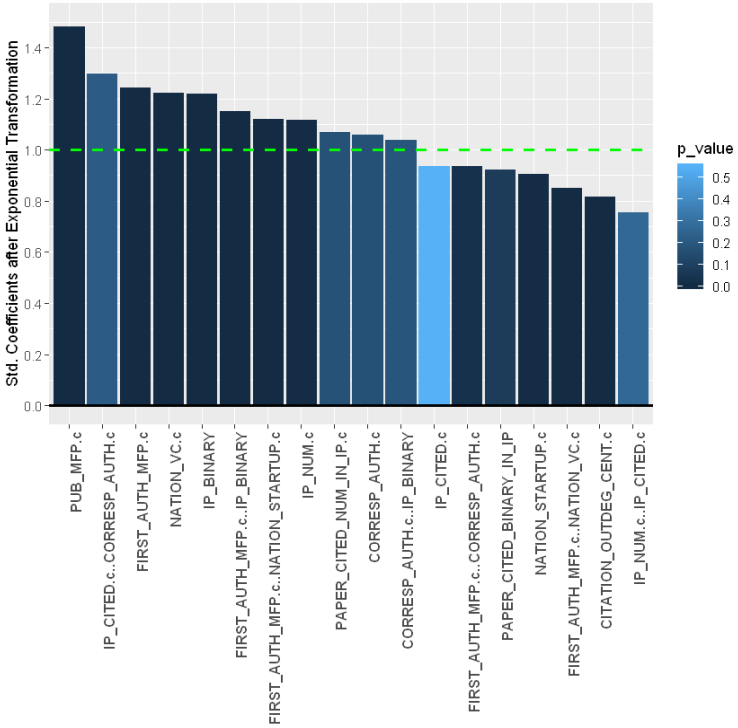
Patent-related Features

Ecosystem Factors

Interaction Terms Factors : Combinations of Above

d) Exponential transformation of standardized β_i, or, raising *e* to the power of standardized βj

e) +,*,,*** respectively denote that the variable is significant at **10%, 5%, 1%** and **0.1%**



for Exit (Table 5.18)

Similar to *for Participant*, PUB_MFP is the most influential at 1.521, next to which FIRST_AUTH_MFP, NATION_TURNOVER and NATION_VC are similarly influential at 1.301, 1.261 and 1.257 respectively, followed by FIRST_AUTH_MFP * IP_BINARY, CORRESP_AUTH * CITATION_OUTDEG_CENT and NATION_STARTUP * FIRST_AUTH_MFP at 1.169, 1.151, and 1.132 respectively. On the other hand, several explanatory variables have their exponentially transformed coefficients less than 1, suggesting their negative effect. For example, similar to the case of *for Participant*, CITATION_OUTDEG_CENT and NATION_STARTUP are negatively influential against the odds, with their transformed coefficients at 0.806 and 0.856 respectively. Interaction Terms Factors with CITATION_OUTDEG_CENT, such as PUB_MFP * CITATION_OUTDEG_CENT, FIRST_AUTH_MFP * CITATION_OUTDEG_CENT and CITATION_OUTDEG_CENT * PAPER_CITED_BINARY_IN_IP are also negatively effective at 0.817, 0.828 and 0.854 respectively. Likewise in *for Participant*, NATION_VC * FIRST_AUTH_MFP has negative effect at 0.810, a seeming weight against FIRST_AUTH_MFP and NATION_VC individually, both of which have positive effect.

Table 5.18 Effects of Explanatory Variables on Odds for Exit in Cas9

	Explanatory Variables [a, b, c]	exp_coef [d]	p_value [e]	
1	PUB_MFP.c	1.521	0.000	***
2	FIRST_AUTH_MFP.c	1.301	0.002	**
3	NATION_TURNOVER.c	1.261	0.003	**
4	NATION_VC.c	1.257	0.001	***
5	FIRST_AUTH_MFP.c * IP_BINARY	1.169	0.004	**
6	CORRESP_AUTH.c * CITATION_OUTDEG_CENT.c	1.151	0.015	*
7	CITATION_OUTDEG_CENT.c * IP_BINARY	1.149	0.131	
8	IP_CITED.c	1.144	0.200	
9	NATION_STARTUP.c * FIRST_AUTH_MFP.c	1.132	0.000	***
10	COAUTH_DEG_CENT.c * NATION_TURNOVER.c	1.114	0.161	
11	PUB_MFP.c * UNIV_RESEARCH.c	1.113	0.020	*
12	COAUTH_DEG_CENT.c * IP_BINARY	1.073	0.089	+
13	IP_NUM.c	1.069	0.164	
14	CORRESP_AUTH.c * IP_BINARY	1.069	0.278	
15	PAPER_CITED_BINARY_IN_IP	1.061	0.478	
16	IP_BINARY	1.051	0.576	
17	CORRESP_AUTH.c	1.029	0.612	
18	COAUTH_DEG_CENT.c	0.994	0.929	
19	IP_BINARY * UNIV_RESEARCH.c	0.932	0.136	
20	CORRESP_AUTH.c * PAPER_CITED_BINARY_IN_IP	0.931	0.225	
21	UNIV_RESEARCH.c	0.907	0.141	
22	NATION_VC.c * IP_CITED.c	0.858	0.101	
23	NATION_STARTUP.c	0.856	0.002	**
24	CITATION_OUTDEG_CENT.c * PAPER_CITED_BINARY_IN_IP	0.854	0.062	+
25	FIRST_AUTH_MFP.c * CITATION_OUTDEG_CENT.c	0.828	0.002	**
26	PUB_MFP.c * CITATION_OUTDEG_CENT.c	0.817	0.023	*
27	NATION_VC.c * FIRST_AUTH_MFP.c	0.810	0.004	**
28	CITATION_OUTDEG_CENT.c	0.806	0.011	*

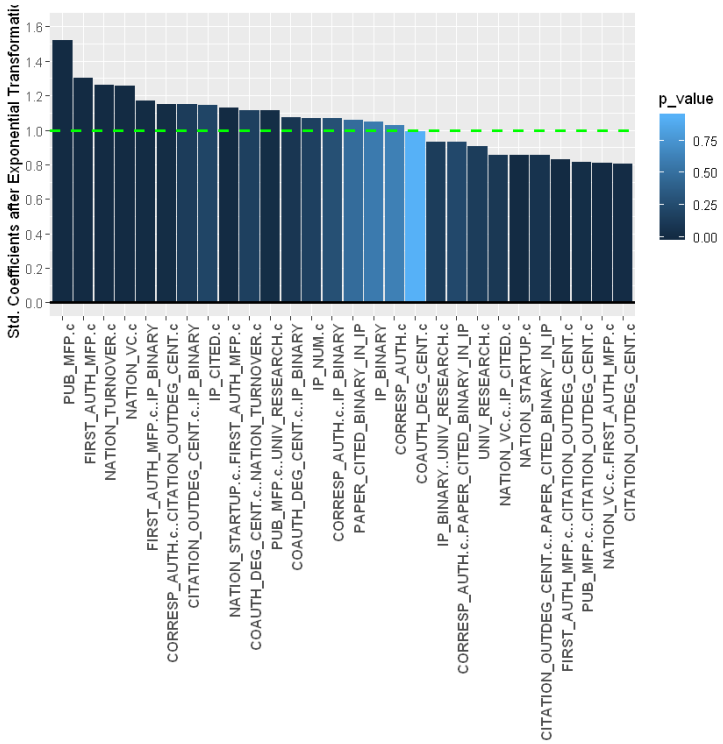
a) "_MFP" indicates that these variables were turned into their multivariable fractional polynomial forms. Regarding Cas9, the same MFPs were applied both to Participant and Exit.

b) ".c" indicates that these variables were centered from their originals, i.e. adjusted so that their means became zero, although centralization does not affect the values of exp_coef and p_value herein. This indication is omitted from the paper.

c) ***Paper-related Features***
Patent-related Features
Ecosystem Factors
Interaction Terms Factors : Combinations of Above

d) Exponential transformation of standardized β_i, or, raising *e* to the power of standardized βj

e) **+,*,**,*** respectively denote that the variable is significant at **10%, 5%, 1%** and **0.1%**



(a) ***Microbiome***

for Participant **(**Table 5.19**)**

CORRESP_AUTH_MFPp is the most influential at 1.846, next to which UNIV_RESEARCH, PAPER_CITED_BINARY_IN_IP, NATION_VC are similarly influential at 1.264, 1.144 and 1.134 respectively, followed by CORRESP_AUTH_MFPp * UNIV_SIZE and IP_NUM * IP_CITED_MFPp at 1.063 and 1.061 respectively. In contrast, several explanatory variables with the exponentially transformed coefficients (exp_coef) less than 1 have negative effect. UNIV_SIZE individually is the most negatively influential at 0.846, while its Interaction Terms Factors such as CORRESP_AUTH_MFPp * UNIV_SIZE has positive effect as seen above. Subsequently, NATION_VC * UNIV_RESEARCH has negative effect at 0.877, which seemingly works as weight to some extent against NATION_VC and UNIV_RESEARCH individually, both of which having positive effect.

for Exit (Table 5.20)

Similar to the case of *for Participant*, CORRESP_AUTH_MFPp is the most influential at 1.984, next to which CITATION_OUTDEG_CENT * CITATION_INDEG_CENT and FIRST_AUTH * COAUTH_DEG_CENT are relatively closely influential at 1.697 and 1.506 respectively, followed by PAPER_CITED_BINARY_IN_IP, UNIV_INNOV * CITATION_INDEG_CENT, CITATION_INDEG_CENT, and FIRST_AUTH * NATION_VC_MFPe at 1.210, 1.181, 1.126 and 1.114 respectively. On the other hand, explanatory variables with less-than-1 exp_coef, having negative effect, are observed. CITATION_INDEG_CENT * COAUTH_DEG_CENT has the most negative effect at 0.483, next to which PAPER_CITED_NUM_IN_IP and CITATION_OUTDEG_CENT * FIRST_AUTH are at 0.662 and 0.672. NATION_VC_MFPe is individually found to be negative for Exit at 0.782. CORRESP_AUTH_MFPe * CITATION_OUTDEG_CENT and CORRESP_AUTH_MFPe * CITATION_INDEG_CENT are found to be negative at 0.796 and 0.841, seemingly working as weight against the top two most effective variables composed of the same components as those of them. COAUTH_DEG_CENT individually has negative effect at 0.863 too.

Table 5.19 Effects of Explanatory Variables on Odds for Participant in Microbiome

	Explanatory Variables [a, b, c]	exp_coef [d]	p_value [e]	
1	CORRESP_AUTH_MFPp.c	1.846	0.000	***
2	UNIV_RESEARCH.c	1.264	0.000	***
3	PAPER_CITED_BINARY_IN_IP	1.144	0.000	***
4	NATION_VC.c	1.134	0.000	***
5	CORRESP_AUTH_MFPp.c * UNIV_SIZE.c	1.063	0.003	**
6	IP_NUM.c * IP_CITED_MFPp.c	1.061	0.073	+
7	PAPER_CITED_NUM_IN_IP.c * UNIV_SIZE.c	1.050	0.057	+
8	CITATION_OUTDEG_CENT.c * FIRST_AUTH.c	1.040	0.214	
9	UNIV_INNOV.c	1.020	0.674	
10	FIRST_AUTH.c * IP_NUM.c	0.976	0.277	
11	CITATION_OUTDEG_CENT.c	0.967	0.396	
12	PAPER_CITED_NUM_IN_IP.c	0.955	0.251	
13	CORRESP_AUTH_MFPp.c * UNIV_INNOV.c	0.944	0.004	**
14	CORRESP_AUTH_MFPp.c * PAPER_CITED_BINARY_IN_IP	0.935	0.007	**
15	FIRST_AUTH.c * UNIV_RESEARCH.c	0.918	0.012	*
16	CITATION_OUTDEG_CENT.c * UNIV_SIZE.c	0.913	0.028	*
17	UNIV_INNOV.c * UNIV_RESEARCH.c	0.912	0.029	*
18	FIRST_AUTH.c	0.905	0.007	**
19	CORRESP_AUTH_MFPp.c * CITATION_OUTDEG_CENT.c	0.905	0.006	**
20	NATION_VC.c * UNIV_RESEARCH.c	0.877	0.000	***
21	UNIV_SIZE.c	0.846	0.001	***

a) "_MFPp" indicates that these variables were turned into their multivariable fractional polynomial forms for analysis of Participant.

b) ".c" indicates that these variables were centered from their originals, i.e. adjusted so that their means became zero, although centralization does not affect the values of exp_coef and p_value herein. This indication is omitted from the paper.

c) ***Paper-related Features***

Patent-related Features

Ecosystem Factors

Interaction Terms Factors : Combinations of Above

d) Exponential transformation of standardized β_j, or, raising e to the power of standardized βj

e) **+,*,**,*** ** respectively denote that the variable is significant at **10%, 5%, 1%** and **0.1%**

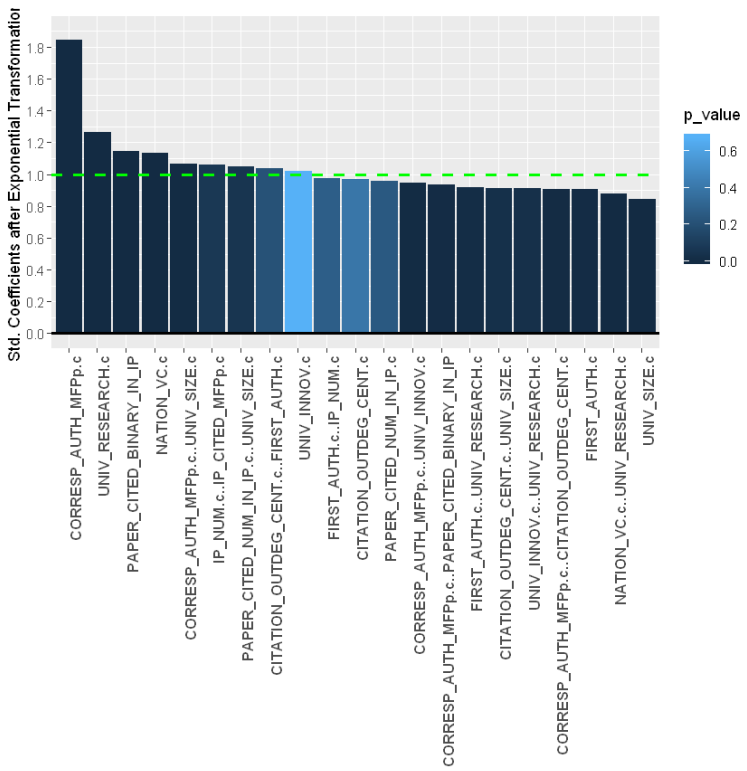


Table 5.20 Effects of Explanatory Variables on Odds for Exit in Microbiome

	Explanatory Variables [a, b, c]	exp_coef [d]	p_value [e]	
1	CORRESP_AUTH_MFPe.c	1.984	0.000	***
2	CITATION_OUTDEG_CENT.c * CITATION_INDEG_CENT.c	1.697	0.005	**
3	FIRST_AUTH.c * COAUTH_DEG_CENT.c	1.506	0.000	***
4	PAPER_CITED_BINARY_IN_IP	1.210	0.000	***
5	UNIV_INNOV.c * CITATION_INDEG_CENT.c	1.181	0.000	***
6	CITATION_INDEG_CENT.c	1.126	0.015	*
7	FIRST_AUTH.c * NATION_VC_MFPe.c	1.114	0.026	*
8	FIRST_AUTH.c	0.953	0.443	
9	CITATION_OUTDEG_CENT.c	0.940	0.423	
10	CORRESP_AUTH_MFPe.c * FIRST_AUTH.c	0.937	0.175	
11	UNIV_INNOV.c	0.911	0.060	+
12	COAUTH_DEG_CENT.c	0.863	0.077	+
13	CORRESP_AUTH_MFPe.c * CITATION_INDEG_CENT.c	0.841	0.064	+
14	CORRESP_AUTH_MFPe.c * CITATION_OUTDEG_CENT.c	0.796	0.010	**
15	NATION_VC_MFPe.c	0.782	0.000	***
16	CITATION_OUTDEG_CENT.c * FIRST_AUTH.c	0.672	0.009	***
17	PAPER_CITED_NUM_IN_IP.c	0.662	0.041	*
18	CITATION_INDEG_CENT.c * COAUTH_DEG_CENT.c	0.483	0.011	*

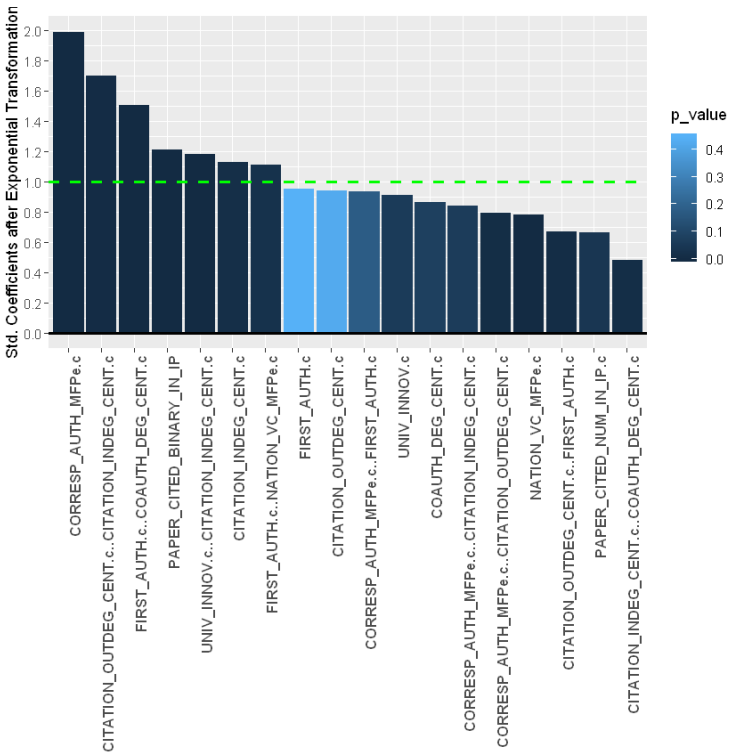
a) "_MFPe" indicates that these variables were turned into their multivariable fractional polynomial forms for analysis of Exit.

b) ".c" indicates that these variables were centered from their originals, i.e. adjusted so that their means became zero, although centralization does not affect the values of exp_coef and p_value herein. This indication is omitted from the paper.

c) ***Paper-related Features***
Patent-related Features
Ecosystem Factors
Interaction Terms Factors : Combinations of Above

d) Exponential transformation of standardized β_i, or, raising *e* to the power of standardized βj

e) **+,*,**,*** ** respectively denote that the variable is significant at **10%, 5%, 1%** and **0.1%**



5.3.2. Assessing Academic Researchers Related to 5-Biopharma-Topics

5.3.2.1. Computation of Startup Readiness and Its Implication for Practical Use

As done for **Cas9** and **Microbiome** in 5.3.1.1, by computing the logistic regression assessment models shown in (5-13) and (5-14), values of each author's startup readiness regarding **5-Biopharma-Topics** that are expressed between 0 and 1, can be attained, with respect to Participant and Exit respectively. Their top 30 startup readiness authors are again presented in Table 5.21 in descending order, including their computed startup readiness values, actual event status and relevant event timing, if any.

The same analysis and implications made in 5.3.1.1 for **Cas9** and **Microbiome** exactly applies here as follows. We observe that, regarding **5-Biopharma-Topics** as well, the lineups of highly ranked academic researchers regarding startup readiness for Participant and Exit, together with those of actually positive ones, are considerably different from each other. The assessment model potentially enables business stakeholders to detect the ones who are likely to be Participant/Exit and the ones who are promising in terms of Exit among researchers whether Participant or not. This model can also be used to measure startup readiness for Exit ever before creating startups, possibly urging academic researchers and their stake holders (e.g., university administrators and policymakers) to understand their limitations of startup readiness regarding Exit early on, and to improve their features of explanatory variables (See 6.2.2 regarding importance of each of them) upon or before engaging in or founding startups.

In this way, this model can help formulate and develop promising academic startups in **5-Biopharma-Topics** too, enabling researchers to intensively enhance their Essential Individual Factors, while letting business stakeholders exert their expertise of financing, management and business in a mutually complementary manner.

Table 5.21 Top 30 Startup Readiness Researchers (for Participant & Exit) in 5-Biopharma-Topics
(Researchers Whose Family and Given Names in This Order Are Identified Only)
(Positive Researchers Highlighted in Yellow for Participant and in Orange for Exit)

5-Biopharma-Topics

for Participant

Rank	Startup Readiness	Author	Research Topic	Participant	Est. Date
1	1.000	Zhang, Feng	CRISPR	1	11/1/2013
2	1.000	Zhang, Feng	Cas9	1	2/11/2004
3	1.000	Doudna, Jennifer	CRISPR	1	11/1/2013
4	1.000	Liu, David	Cas9	1	1/1/1997
5	0.999	Church, George	Cas9	1	1/1/2009
6	0.981	Church, George	CRISPR	1	1/1/2009
7	0.926	Quake, Stephen	Cas9	1	5/9/2003
8	0.899	Apte, Zachary	microbiome	1	10/15/2012
9	0.894	Li, Li	Cas9	1	1/1/2014
10	0.880	Miller, Jeffrey	CRISPR	1	2/1/1998
11	0.876	Joung, Keith	Cas9	1	11/1/2013
12	0.861	Richman, Jessica	microbiome	1	10/15/2012
13	0.854	Doudna, Jennifer	Cas9	1	11/1/2013
14	0.851	Mills, David	microbiome	1	1/1/1998
15	0.849	Charpentier, Emmanuelle	Cas9	1	10/31/2013
16	0.819	Li, Li	CRISPR	1	4/21/1998
17	0.813	Zhao, Yangbing	CRISPR	1	1/1/2015
18	0.809	Sanjana, Neville	CRISPR	0	
19	0.797	Joung, Keith	CRISPR	1	11/1/2013
20	0.761	Kim, Jong, Wook	CRISPR	0	
21	0.749	Poste, George	exosome	1	6/1/1992
22	0.739	Bajaj, Jasmohan	microbiome	0	
23	0.738	Zhang, Lei	CRISPR	0	
24	0.735	Anderson, Daniel	CRISPR	1	1/1/2015
25	0.703	Zhang, Lei	Cas9	1	5/12/2005
26	0.702	Wang, Jianping	microbiome	0	
27	0.698	Lu, Timothy	Cas9	1	1/1/2016
28	0.687	Blaser, Martin	microbiome	0	
29	0.673	Elinav, Eran	microbiome	0	
30	0.667	Wang, Jianbin	CRISPR	1	1/1/2015

for Exit

Rank	Startup Readiness	Author	Research Topic	Exit	IPO Date	M&A Date	Participant	Est. Date
1	1.000	Zhang, Feng	CRISPR	1	2/3/2016		1	11/1/2013
2	1.000	Liu, David	Cas9	1	9/24/2008		1	1/1/1997
3	1.000	Doudna, Jennifer	CRISPR	1	2/3/2016		1	11/1/2013
4	0.947	Zhang, Feng	Cas9	1	2/3/2016		1	2/11/2004
5	0.856	Ochiya, Takahiro	exosome	0			0	
6	0.797	Mills, David	microbiome	0			1	1/1/1998
7	0.785	Church, George	CRISPR	1	2/3/2016	1/20/2017	1	1/1/2009
8	0.745	June, Carl	CAR-T	0			1	1/1/2015
9	0.695	Blaser, Martin	microbiome	0			0	
10	0.685	Church, George	Cas9	1	2/3/2016	1/20/2017	1	1/1/2009
11	0.602	Jensen, Torben, Heick	exosome	0			0	
12	0.562	Kim, Jongmin	CRISPR	0			0	
13	0.529	Xavier, Ramnik	microbiome	0			1	12/23/2016
14	0.457	Schloss, Patrick	microbiome	0			0	
15	0.457	Bajaj, Jasmohan	microbiome	0			0	
16	0.444	Doudna, Jennifer	Cas9	1	2/3/2016		1	11/1/2013
17	0.444	Liu, Yutao	exosome	0			0	
18	0.421	Bajaj, Jasmohan	microbiome	0			0	
19	0.409	Lu, Tse-Yuan	microbiome	0			0	
20	0.384	Cooper, Laurence	CAR-T	0			0	
21	0.378	Wang, Yang	exosome	0			1	10/22/2015
22	0.374	Aagaard, Kjersti	microbiome	0			0	
23	0.365	Whiteside, Theresa	CAR-T	0			0	
24	0.358	Elinav, Eran	microbiome	0			0	
25	0.356	Wang, Liang	CAR-T	0			0	
26	0.349	Li, Jian	CRISPR	0			0	
27	0.347	Wang, Yong	CRISPR	0			0	
28	0.347	Zhang, Chen-Yu	CAR-T	0			0	
29	0.346	Gilbert, Jack	microbiome	0			0	
30	0.343	Sanjana, Neville	CRISPR	0			0	

5.3.2.2. Assessing Importance of Each Explanatory Variable

As discussed in 5.3.1.2, in order to compare the scope of the impact of relevant explanatory variables, we need to standardize their coefficients so that they are based on the same scale and can explain each author's likelihood of being a Participant/Exit. Furthermore, as exhibited in 5.3.1.2, this thesis attained exponential transformation of these standardized coefficients, or, raised e (the base of the natural logarithm) to the power of standardized β_j(for Participant, j = 1, …, 66, and for Exit, j = 1, …, 24), in order to detect and compare the effects of each variable in terms of how many times the odds of each author's being a Participant or Exit (i.e., $\frac{P_i}{1-P_i}$) increase associated with one S.D. increase in the exposure, across all the relevant explanatory variables. Put another way, the exponentially transformed coefficients mean the growth multiple of the odds per one S.D. increase for each variable.

By using the "function" function of the "base" package and the "ggplot" function of the "ggplot2" package in R, likewise as demonstrated in 5.3.1.2, the exponentially transformed standard coefficients of explanatory variables were computed, as shown in Table 5.22 and Table 5.23.

For Participant **(**Table 5.22**)**

CORRESP_AUTH_MFPp * IP_CITED and CORRESP_AUTH_MFPp are outstandingly influential with the coefficients at 1.655 and 1.434, both of which include an MFP-transformed version of CORRESP_AUTH. Subsequently, other substantially influential variables follow such as FIRST_AUTH_MFPp * NATION_STARTUP_MFPp at 1.176, NATION_STARTUP_MFPp at 1.163, COAUTH_DEG_CENT_FPp * NATION_STARTUP_MFPp at 1.153, IP_BINARY at 1.147, FIRST_AUTH_MFPp at 1.143, PAPER_CITED_NUM_IN_IP * UNIV_RESEARCH at 1.140, NATION_VC_MFPp * UNIV_RESEARCH at 1.139, IP_CITED at 1.119, and KW_GROWTH * NATION_TURNOVER_MFPp at 1.112.

In contrast, the coefficients of several variables are less than 1, meaning that their increase negatively affects the odds. The most negatively influential such variable is PAPER_CITED_NUM_IN_IP * IP_CITING at 0.692, followed by other substantially counter-effective variables PUB_MFPp at 0.785, IP_CITED * UNIV_RESEARCH at 0.822, FINANCED_AMOUNT * NATION_TURNOVER_MFPp at 0.840, NATION_VC_MFPp * NATION_STARTUP_MFPp at 0.846, NATION_VC_MFPp at 0.852, NATION_TURNOVER_MFPp * NATION_STARTUP_MFPp at 0.858, NATION_STARTUP_MFPp * UNIV_INNOV at 0.884, PUB_MFPp * KW_GROWTH

at 0.886, IP_CITED * NATION_STARTUP_MFPp at 0.887, and CORRESP_AUTH_MFPp * IP_NUM at 0.899.

Table 5.22 Effects of Variables on Odds for Participant in 5-Biopharma-Topics

	Explanatory Variables [a, b, c]	exp_coef [d]	p_value [e]	
1	CORRESP_AUTH_MFPp.c * IP_CITED.c	1.635	0.004	**
2	CORRESP_AUTH_MFPp.c	1.434	0.000	***
3	CORRESP_AUTH_MFPp.c * IP_CITING.c	1.262	0.108	
4	FIRST_AUTH_MFPp.c * NATION_STARTUP_MFPp.c	1.176	0.000	***
5	NATION_STARTUP_MFPp.c	1.163	0.000	***
6	COAUTH_DEG_CENT_MFPp.c * NATION_STARTUP_MFPp.c	1.153	0.000	***
7	IP_BINARY	1.147	0.000	***
8	FIRST_AUTH_MFPp.c	1.143	0.000	***
9	PAPER_CITED_NUM_IN_IP.c * UNIV_RESEARCH.c	1.140	0.003	**
10	NATION_VC_MFPp.c * UNIV_RESEARCH.c	1.139	0.000	***
11	IP_CITED.c	1.119	0.051	+
12	KW_GROWTH.c * NATION_TURNOVER_MFPp.c	1.112	0.002	**
13	COAUTH_DEG_CENT_MFPp.c * IP_CITING.c	1.094	0.066	+
14	NATION_VC_MFPp.c * FIRST_AUTH_MFPp.c	1.088	0.001	***
15	IP_NUM.c * UNIV_RESEARCH.c	1.086	0.008	**
16	UNIV_RESEARCH.c	1.084	0.004	**
17	FINANCED_FREQ.c * UNIV_RESEARCH.c	1.077	0.001	***
18	COAUTH_DEG_CENT_MFPp.c * PAPER_CITED_BINARY_IN_IP	1.076	0.003	**
19	IP_CITING.c	1.074	0.081	+
20	CORRESP_AUTH_MFPp.c * PUB_MFPp.c	1.069	0.000	***
21	NATION_STARTUP_MFPp.c * UNIV_RESEARCH.c	1.063	0.032	*
22	IP_NUM.c	1.061	0.009	**
23	IP_NUM.c * NATION_STARTUP_MFPp.c	1.055	0.005	**
24	IP_CITING.c * UNIV_INNOV.c	1.053	0.031	*
25	NATION_VC_MFPp.c * PUB_MFPp.c	1.053	0.016	*
26	IP_NUM.c * FIRST_AUTH_MFPp.c	1.049	0.182	
27	IP_GROWTH.c	1.044	0.059	+
28	FIRST_AUTH_MFPp.c * PAPER_CITED_BINARY_IN_IP	1.040	0.043	*
29	FIRST_AUTH_MFPp.c * UNIV_RESEARCH.c	1.038	0.076	+
30	FIRST_AUTH_MFPp.c * KW_GROWTH.c	1.034	0.106	
31	IP_GROWTH.c * UNIV_RESEARCH.c	1.032	0.057	+
32	COAUTH_DEG_CENT_MFPp.c	1.018	0.471	
33	CORRESP_AUTH_MFPp.c * COAUTH_DEG_CENT_MFPp.c	1.003	0.811	
34	FINANCED_AMOUNT.c	0.996	0.869	
35	CORRESP_AUTH_MFPp.c * IP_GROWTH.c	0.985	0.196	
36	IP_BINARY * FINANCED_AMOUNT.c	0.982	0.241	
37	IP_BINARY * NATION_STARTUP_MFPp.c	0.975	0.152	
38	IP_GROWTH.c * PAPER_CITED_NUM_IN_IP.c	0.971	0.064	+
39	FIRST_AUTH_MFPp.c * IP_CITED.c	0.971	0.633	
40	CORRESP_AUTH_MFPp.c * PAPER_CITED_BINARY_IN_IP	0.971	0.052	+
41	IP_GROWTH.c * PUB_MFPp.c	0.968	0.136	
42	IP_BINARY * UNIV_RESEARCH.c	0.959	0.024	*
43	COAUTH_DEG_CENT_MFPp.c * UNIV_RESEARCH.c	0.949	0.014	*
44	NATION_TURNOVER_MFPp.c	0.943	0.039	*
45	PAPER_CITED_NUM_IN_IP.c * UNIV_INNOV.c	0.939	0.069	+
46	PUB_MFPp.c * UNIV_RESEARCH.c	0.939	0.009	**
47	UNIV_INNOV.c	0.932	0.003	**
48	PUB_MFPp.c * UNIV_INNOV.c	0.931	0.002	**
49	IP_NUM.c * COAUTH_DEG_CENT_MFPp.c	0.931	0.077	+
50	NATION_VC_MFPp.c * UNIV_INNOV.c	0.929	0.004	**
51	IP_GROWTH.c * UNIV_INNOV.c	0.923	0.000	***
52	IP_GROWTH.c * NATION_TURNOVER_MFPp.c	0.921	0.000	***
53	CORRESP_AUTH_MFPp.c * FINANCED_AMOUNT.c	0.916	0.000	***
54	IP_BINARY * COAUTH_DEG_CENT_MFPp.c	0.914	0.002	**
55	CORRESP_AUTH_MFPp.c * IP_NUM.c	0.899	0.086	+
56	IP_CITED.c * NATION_STARTUP_MFPp.c	0.887	0.000	***
57	PUB_MFPp.c * KW_GROWTH.c	0.886	0.000	***
58	NATION_STARTUP_MFPp.c * UNIV_INNOV.c	0.884	0.000	***
59	NATION_TURNOVER_MFPp.c * NATION_STARTUP_MFPp.c	0.858	0.000	***
60	NATION_VC_MFPp.c	0.852	0.000	***
61	NATION_VC_MFPp.c * NATION_STARTUP_MFPp.c	0.846	0.000	***
62	FINANCED_AMOUNT.c * NATION_TURNOVER_MFPp.c	0.840	0.000	***
63	CORRESP_AUTH_MFPp.c * PAPER_CITED_NUM_IN_IP.c	0.829	0.250	
64	IP_CITED.c * UNIV_RESEARCH.c	0.822	0.006	**
65	PUB_MFPp.c	0.785	0.000	***
66	PAPER_CITED_NUM_IN_IP.c * IP_CITING.c	0.692	0.089	+

a) "_MFPp" indicates that these variables were turned into their multivariable fractional polynomial forms for analysis of Participant.

b) ".c" indicates that these variables were centered from their originals, i.e. adjusted so that their means became zero, although centralization does not affect the values of exp_coef and p_value herein. This indication is omitted from the paper.

c) ***Paper-related Features***

Patent-related Features

Hot Topic Factors

Ecosystem Factors

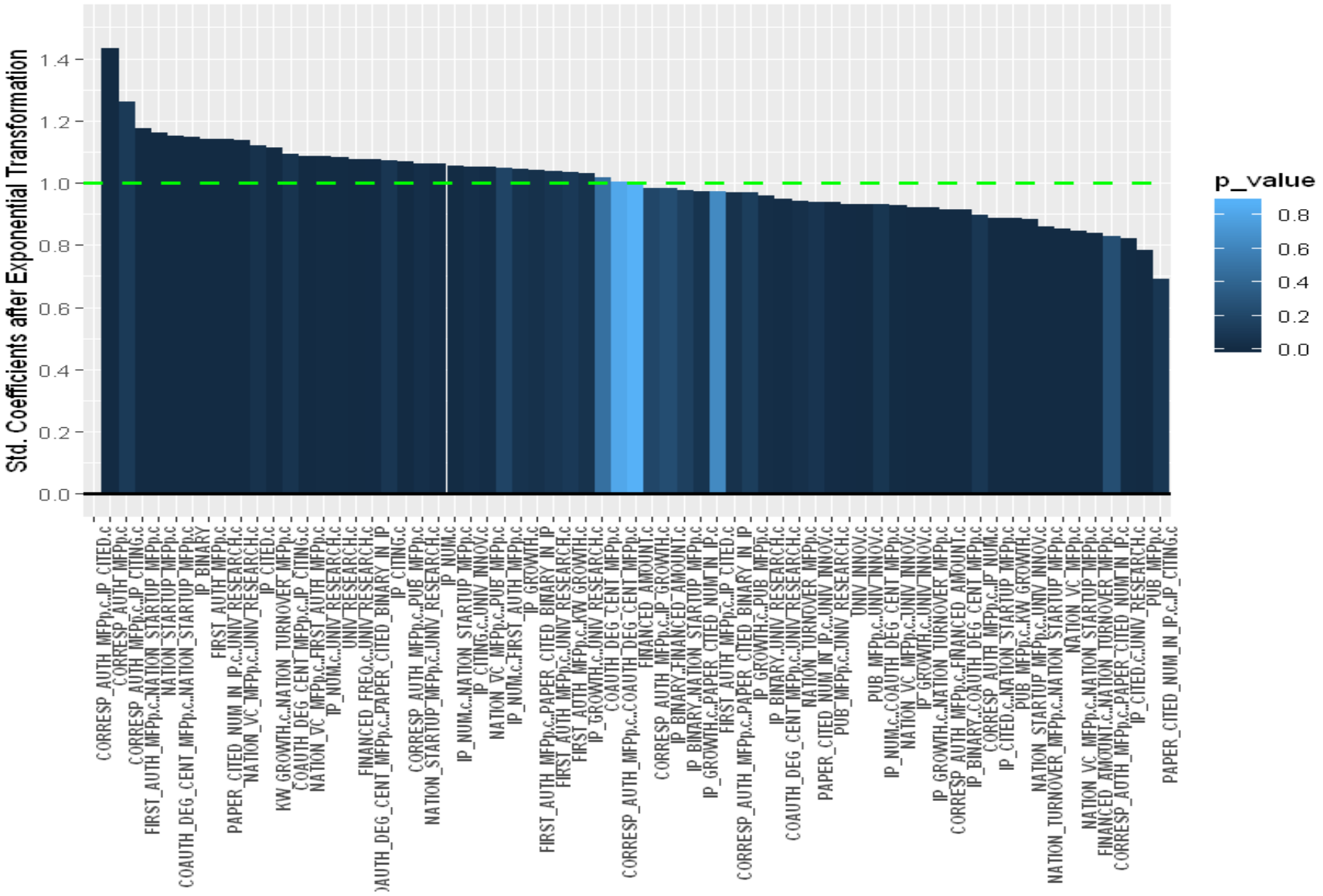
Interaction Terms Factors : Combinations of Above

d) Exponential transformation of standardized β_j, or, raising e to the power of standardized βj

e) **+,*,**,*** ** respectively denote that the variable is significant at **10%, 5%, 1%** and **0.1%**

(…CONTINUED ON NEXT PAGE)

(CONTINUED FROM PREVIOUS PAGE)



For Exit **(**Table 5.23)

CORRESP_AUTH_MFPe is outstandingly the most influential with the coefficient at 1.542, second to which NATION_VC is influential at 1.246. Following influential variables include FIRST_AUTH_MFPe * NATION_STARTUP_MFPe at 1.205, IP_BINARY at 1.172, FIRST_AUTH_MFPe at 1.139, CORRESP_AUTH_MFPe * IP_NUM at 1.123, and COAUTH_DEG_CENT * FINANCED_AMOUNT at 1.103.

On the other hand, CORRESP_AUTH_MFPe * KW_GROWTH, PUB_MFPe, and IP_CITING * IP_NUM are negatively influential against the odds, whose coefficients are 0.826, 0.844, and 0.849 respectively.

Table 5.23 Effects of Variables on Odds for Exit in 5-Biopharma-Topics

	Explanatory Variables [a, b, c]	exp_coef [d]	p_value [e]	
1	CORRESP_AUTH_MFPe.c	1.542	0.000	***
2	NATION_VC.c	1.246	0.000	***
3	FIRST_AUTH_MFPe.c * NATION_STARTUP_MFPe.c	1.205	0.000	***
4	IP_BINARY	1.172	0.000	***
5	FIRST_AUTH_MFPe.c	1.139	0.000	***
6	CORRESP_AUTH_MFPe.c * IP_NUM.c	1.123	0.013	*
7	CORRESP_AUTH_MFPe.c * IP_CITING.c	1.121	0.134	
8	COAUTH_DEG_CENT.c * FINANCED_AMOUNT.c	1.103	0.000	***
9	IP_CITING.c	1.050	0.015	*
10	IP_BINARY * FIRST_AUTH_MFPe.c	1.047	0.017	*
11	CORRESP_AUTH_MFPe.c * FINANCED_AMOUNT.c	1.043	0.004	**
12	FINANCED_AMOUNT.c	1.036	0.227	
13	IP_NUM.c	1.031	0.133	
14	UNIV_SIZE.c	0.970	0.298	
15	CORRESP_AUTH_MFPe.c * UNIV_SIZE.c	0.966	0.004	**
16	FINANCED_AMOUNT.c * PUB_MFPe.c	0.965	0.248	
17	COAUTH_DEG_CENT.c * NATION_STARTUP_MFPe.c	0.958	0.025	*
18	PUB_MFPe.c * UNIV_SIZE.c	0.925	0.010	**
19	CORRESP_AUTH_MFPe.c * PAPER_CITED_BINARY_IN_IP	0.919	0.000	***
20	NATION_VC.c * FIRST_AUTH_MFPe.c	0.910	0.002	**
21	FINANCED_AMOUNT.c * NATION_STARTUP_MFPe.c	0.905	0.001	***
22	IP_CITING.c * IP_NUM.c	0.849	0.012	*
23	PUB_MFPe.c	0.844	0.000	***
24	CORRESP_AUTH_MFPe.c * KW_GROWTH.c	0.826	0.000	***

a) "_MFPe" indicates that these variables were turned into their multivariable fractional polynomial forms for analysis of Exit.

b) ".c" indicates that these variables were centered from their originals, i.e. adjusted so that their means became zero, although centralization does not affect the values of exp_coef and p_value herein. This indication is omitted from the paper.

c) ***Paper-related Features***

Patent-related Features

Hot Topic Factors

Ecosystem Factors

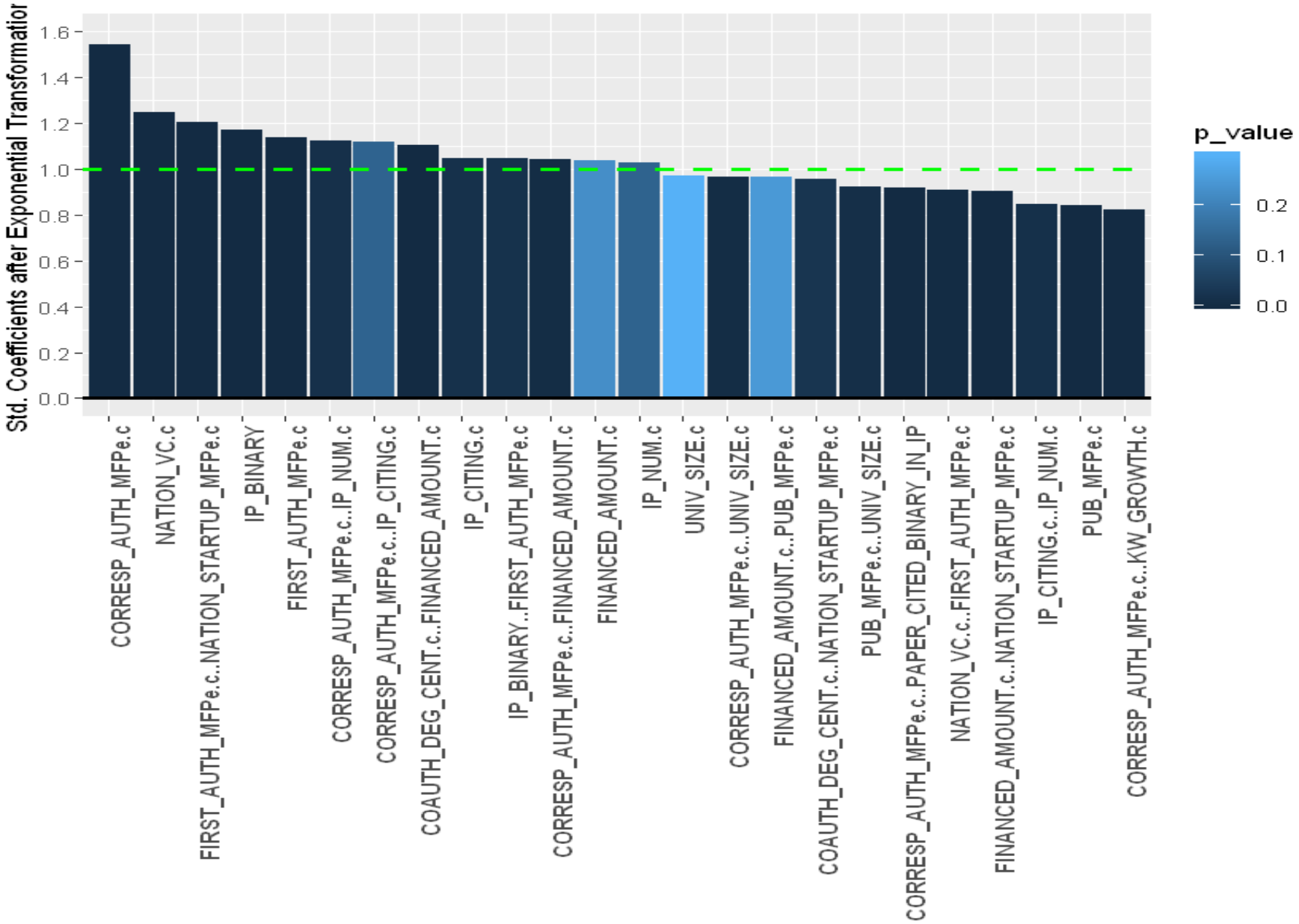
Interaction Terms Factors : Combinations of Above

d) Exponential transformation of standardized β_j, or, raising *e* to the power of standardized βj

e) **+,*,**,*** ** respectively denote that the variable is significant at **10%, 5%, 1%** and **0.1%**

(…CONTINUED ON NEXT PAGE)

(CONTINUED FROM PREVIOUS PAGE)



Chapter 6. Discussion

Similar to the results presented in earlier reports of the literature, the results of this research suggest that the resource-based theory is also applicable in assuming the concept of startup readiness of this thesis. Several determinants of each and every Individual (both Paper-related and Patent-related), Hot Topic and Ecosystem Factor worked, in addition to their Interacting Terms Factors. One result contradictory to that of earlier literature, however, was that, although Landry et al. argue that "publication assets" were found to have no impact on spin-off creation by researchers [24], the results of this thesis demonstrate that various Paper-related Features were key influential determinants of startup readiness in the biopharmaceutical domain, as shown in Table 5.17, Table 5.18, Table 5.19, Table 5.20, Table 5.22 and Table 5.23, and as discussed later.

In recent years, because of the marked emergence of industrial and academic interest related to the biopharmaceutical domain such as Cas9 hereto, it is hypothesized that, for researchers' startup readiness, their academic capabilities such as their profile in research communities and responsibility/initiative of research, and their intellectual property-wise capabilities to invent and build patents on research, as well as their national/regional and academic environments, can matter. Therefore, this thesis built the hypothesis that, the "hotter" the topics of research by scientists and the greater such variables of papers, patents, academic organizations and nations become, the higher the startup readiness by researchers, and that startup readiness also depends on several interaction terms composed of these individual and ecosystem main effect variables.

6.1. Evaluating the Assessment Model

6.1.1. The Model's Performance

In order to measure the performance of the logistic regression model of this thesis as a whole, as shown in the four ROC Curves that were plotted in Figure 5-1, Figure 5-2, Figure 5-3, Figure 5-4, Figure 5-5, Figure 5-6, the values of their AUCs (Area Under The Curve) were obtained herein. AUC represents how much the classifier model is capable of distinguishing between classes, and as a general rule shown by Hosmer and Lemeshow [91], if $0.7 \leq \mathrm{AUC} < 0.8$, such AUC is considered acceptable discrimination. The AUC values were; (i) **Cas9**: 0.6629 for Participant and 0.7029 for Exit, (ii) **Microbiome**: 0.7407 for Participant and 0.7728 for Exit, (iii) **5-Biopharma-Topics**: 0.6898 for Participant and 0.7228 for Exit, in which the values of Exit outperform those of Participant in all topics and the values of **5-Biopharma-Topics** surpass those of solo **Cas9** while falling behind those of solo **Microbiome**.

Regarding the characteristics of startup participation compared to their exit, that typically accompany securities markets for IPO or corporate acquirers for M&A, (i) participation in startups is an invisible act, which leads to lack of registration in the database VentureSource, thus undermining our coverage of Participant, and (ii) participation in startups is more of a personal act than an exit deal evaluated by third parties. Thus, variables regarding Participant, which can be evaluated, tend to be incomplete. Furthermore, with respect to the difference in the model's AUC values between **Cas9** and **Microbiome**, it should be inferably attributable to the stage of the startup activities per each research topic. In fact, it is found that in 2018, while **Microbiome** appeared as a keyword in startups actively financed in VentureSource again, **Cas9** did not. This suggests a possibility that the **Cas9** research field was already saturated for startup participation and exit as of the end of 2018. As seen in Table 5.9 regarding correlations of explanatory variables for **5-Biopharma-Topics**, since KW_GROWTH and FINANCED_FREQ are inversely correlated, it could be suggested that when we assess academic researchers across different biopharmaceutical research topics (e.g., **Cas9**, **CAR-T**, **CRISPR**, **Microbiome** and **Exosome**), we can cater to Hot Topic Factors related to these topics: KW_GROWTH, IP_GROWTH, FINANCED_AMOUNT, and FINANCED_FREQ as shown in Table 6.1 (See 4.4.2), in order to enhance the assessment model's stability, with an appropriate range of research topics rather than with only one topic.

Table 6.1 Hot Topic Factors for Top 5 Biopharmaceutical Research Topics

	KW_GROWTH	IP_GROWTH	Industry Code
Cas9	5.321	6.520	Gene Therapy
CAR-T	5.000	18.375	Gene Therapy
CRISPR	4.522	6.026	Gene Therapy
Microbiome	2.542	5.440	Biotechnology Therapeutics
Exosome	2.200	2.810	Pharmaceuticals
	FINANCED_AMOUNT	**FINANCED_FREQ**	
Gene Therapy	27.512	67	
Biotechnology Therapeutics	23.940	140	
Pharmaceuticals	32.166	121	

6.1.2. The Model's Assessment of Academic Researchers

To inspect how well the logistic regression assessment model of this thesis classified the relevant academic researchers as Participants/Exits and Non-Participants/Non-Exits at each researcher level, this section visualized distributions of estimated probability of startup readiness and compared it to the actual results of whether the relevant researchers became Participants/Exits or Non-Participants/Non-Exits, respectively, for authors related to both **Cas9** and **Microbiome**, as well as the **5-Biorpharma-Topics**, as shown in Figure 6-1, Figure 6-2 and Figure 6-3.

Cas9

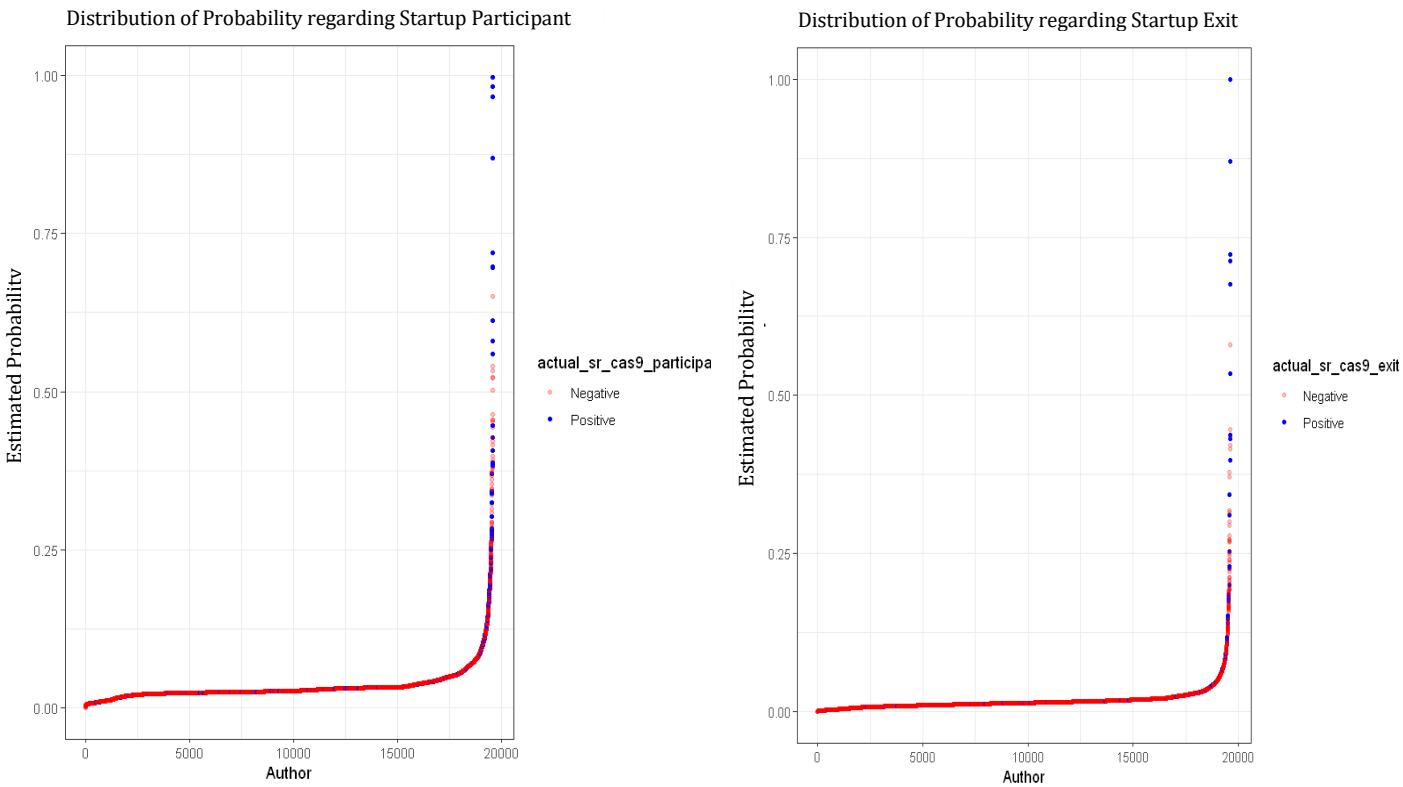


Figure 6-1 Distribution of Startup Participants/Exits Plotted on the Predicted Probability Curve in Cas9

Microbiome

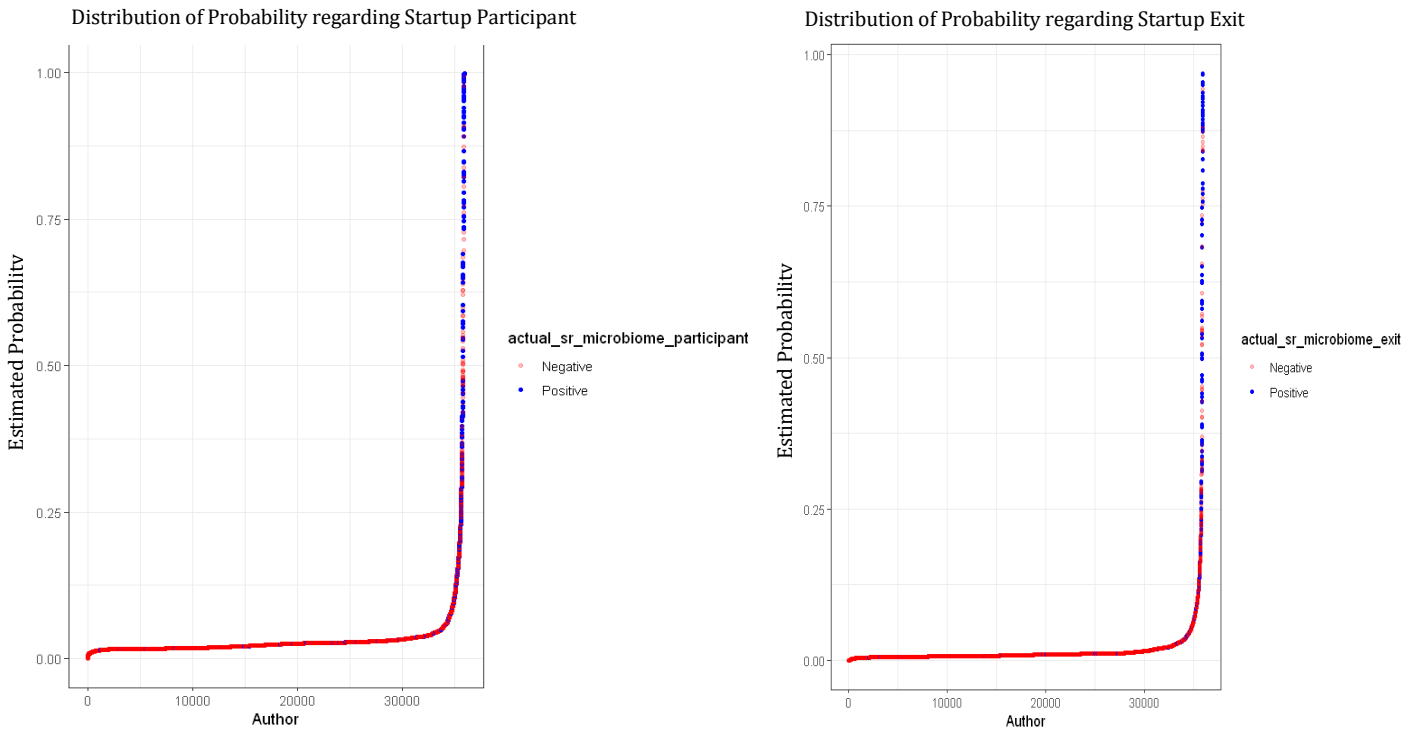


Figure 6-2 Distribution of Startup Participants/Exits Plotted on Estimated Probability Curve in Microbiome

5-Biopharma- Topics

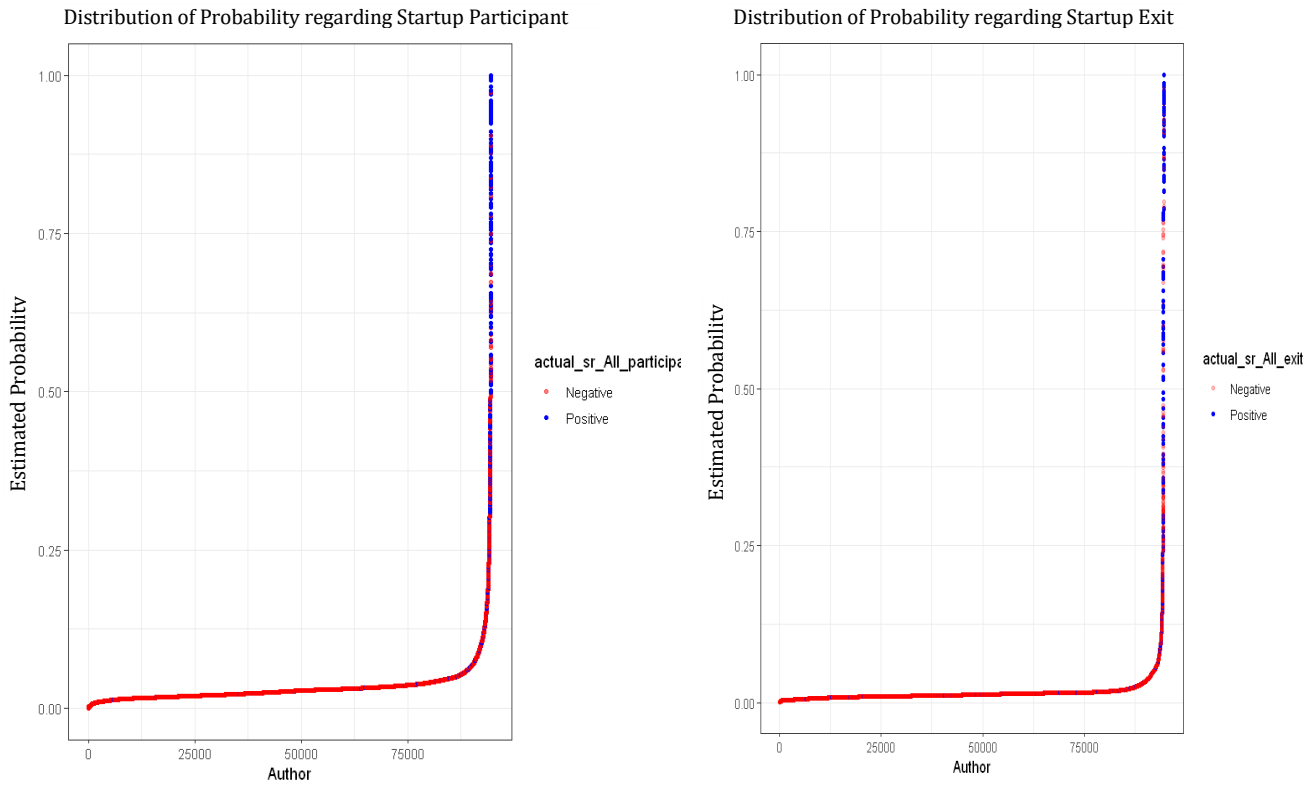


Figure 6-3 Distribution of Startup Participants/Exits Plotted on Estimated Prob. Curve in 5-Biopharma-Topics

As observed in these figures, authors (researchers) with higher probability of startup readiness tend to more likely be Participants/Exits (blue dots), than those with lower probability who tend to be Non-Participants/Non-Exits (diluted red dots). As visualized, we observe that authors belonging to solo **Microbiome** and **5-Biopharma-Topics** are well-classified as Participants/Exits, compared to authors belonging to solo **Cas9**.

Additionally, to compare the distribution of estimated probabilities regarding startup readiness between the two researcher groups: positives (Participants/Exits) and negatives (Non-Participants/Non-Exits) who belong to **Cas9**, **Microbiome** and **5-Biopharma-Topics**, this section visualized distributions of predicted probabilities of both groups and compared them in Figure 6-4, Figure 6-5 and Figure 6-6.

As observed in these figures, authors (researchers) who are positives (Participants/Exits, blue dots) have wider range of estimated probability of startup readiness, whereas those who are negatives (Non-Participants/Non-Exits, diluted red dots) tend to have very skewed distribution to lower predicted probability.

Cas9

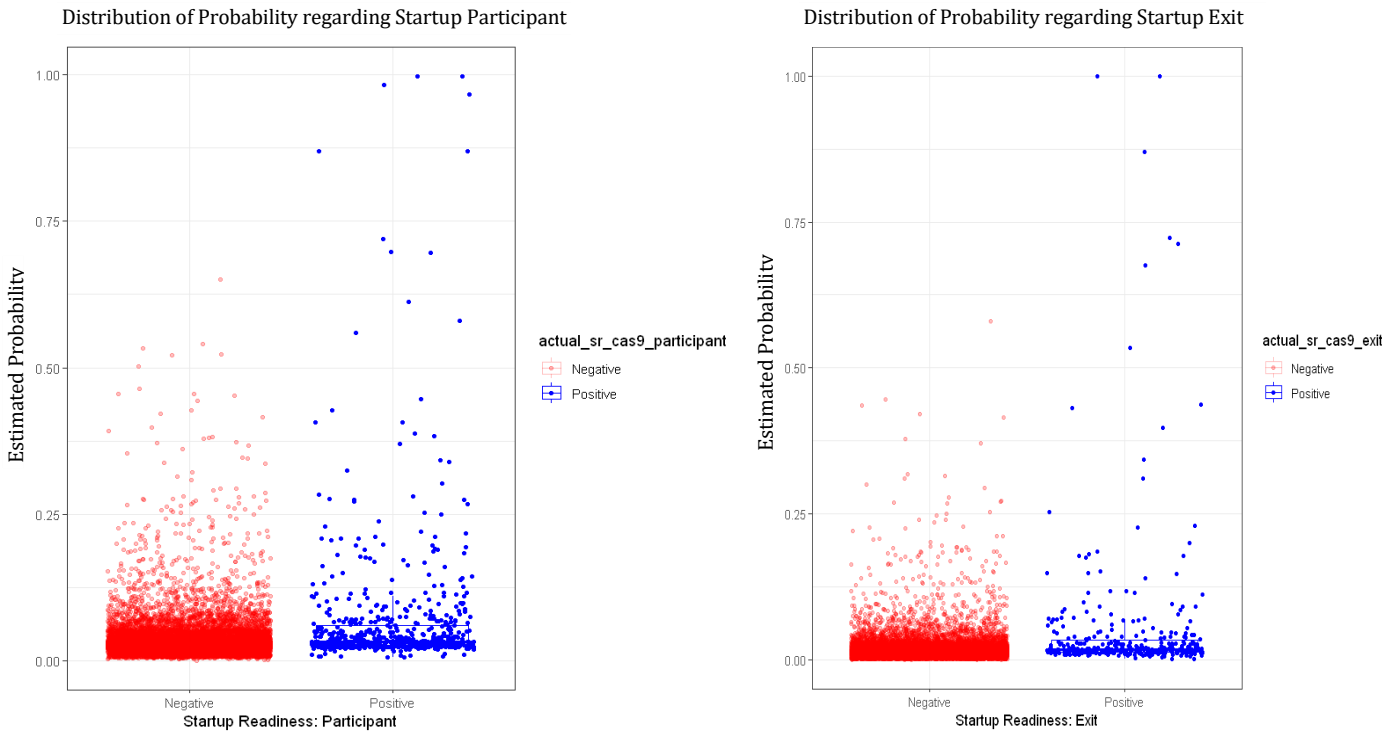


Figure 6-4 Jittered Estimated Probabilities per Cas9 Researcher Group (Negatives vs Positives)

Microbiome

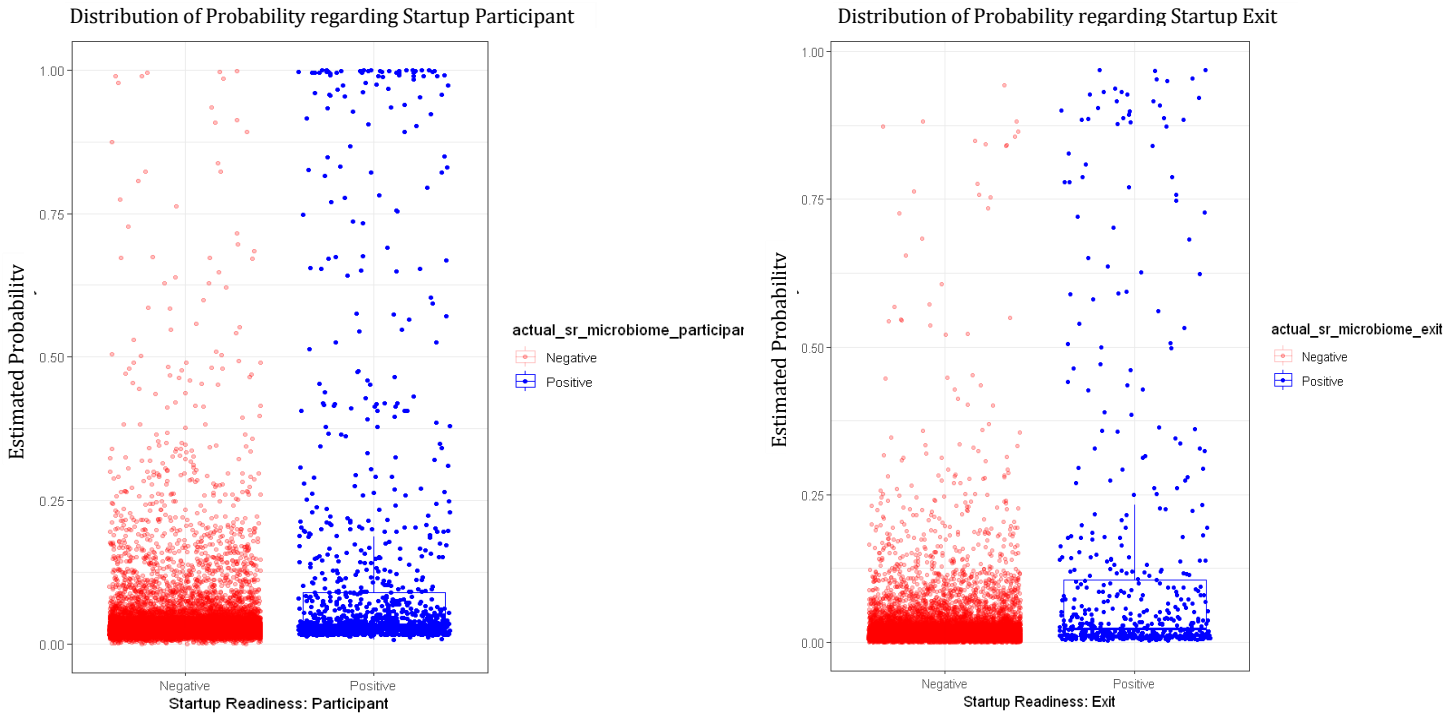


Figure 6-5 Jittered Estimated Probabilities per Microbiome Researcher Group (Negatives vs Positives)

5-Biopharma- Topics

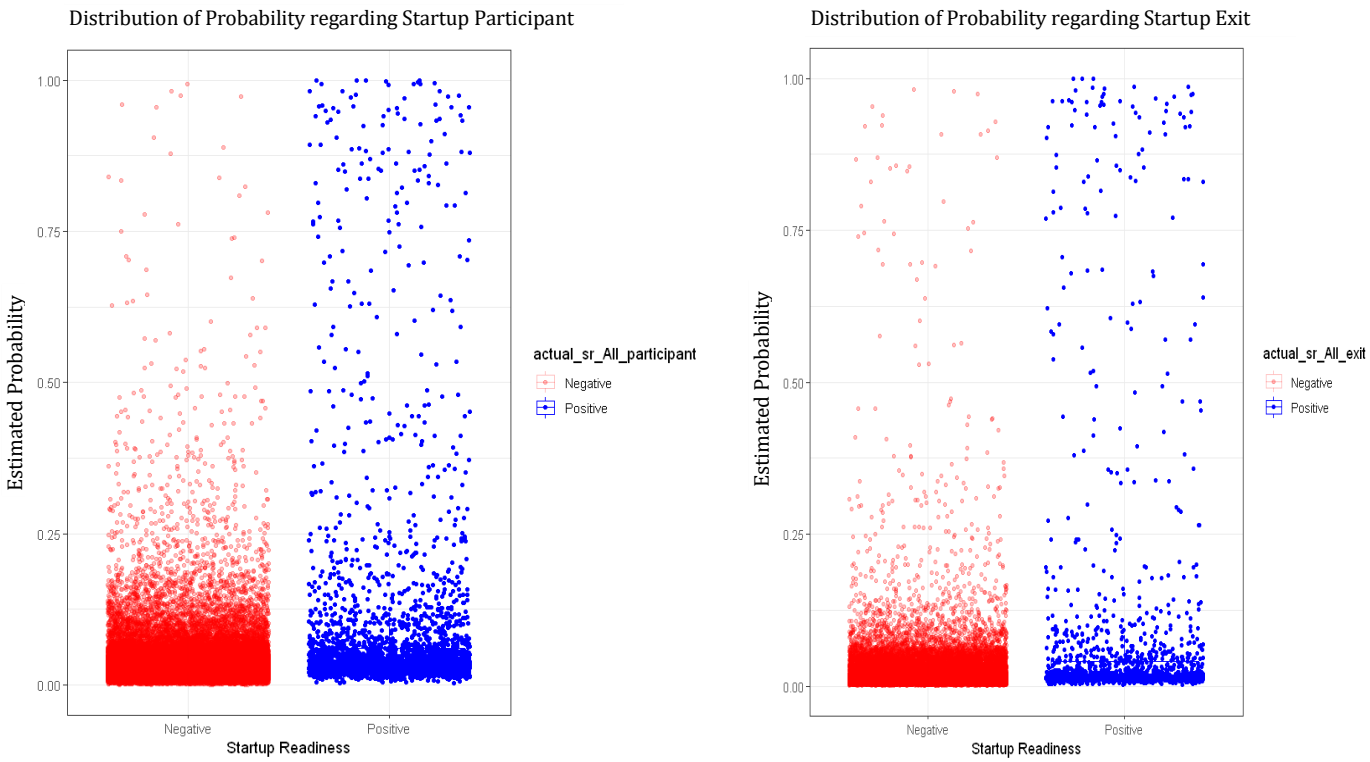


Figure 6-6 Jittered Estimated Probabilities per 5-Biopharma-Topics Researcher Group (Negatives vs Positives)

6.2. Interpreting Explanatory Variables per Each Researcher Group

The logistic regression model shows that a complementary set of explanatory variables are mobilized by individual authors as well as the hot topic trend and the relevant ecosystem when startup readiness is activated. As developed in 4.4, explanatory variables for the assessment model of this thesis is composed of a complementary set of (i) Individual Factors composed of Paper-related Features and Patent-related Features, (ii) Hot Topic Factors/Features, (iii) Ecosystem Factors composed of Academic Organization-related Features and Nation-related Features, and (iv) Interaction Terms Factors which are combinations of independent explanatory variables belonging to different or the same type(s) of features/variables. They are implemented in the model to reflect externalities and spillovers among variables in order to achieve higher explanatory performance.

In order to interpret these explanatory variables on a multidimensional-analysis basis, this section will discuss (i) characteristics of each variable's mean, standard deviation and distribution, (ii) effects of each variable, and (iii) importance of each set of variables, per each researcher group corresponding to the positive and negative groups regarding Participant and Exit.

Researchers are grouped in two dimensions: (i) three-fold research topics: **Cas9**, **Microbiome** and **5-Biopharma-Topics** (again, the top five biopharmaceutical topics combined: **Exosome**, **Microbiome**, **CRISPR**, **Cas9**, and **CAR-T**, as discussed in 5.1.1.6), and, (ii) dichotomous event status expressed by target variables regarding startup readiness: Participant/Exit (positive) and Non-Participant/Non-Exit (negative). Refer to 4.4 for definitions of each variable.

6.2.1. Characteristics of Each Explanatory Variable's Mean, SD and Distribution

Characteristics of each variable in light of the summary statistics composed of its mean, standard deviation (SD) and distribution are compared herein, across different researcher groups in terms of their research topics (**Cas9**, **Microbiome** and **5-Biopharma-Topics**) and event status (the negatives such as non-Participant and non-Exit researchers, the positives such as Participant and Exit researchers, and a mixed state that is Participant but has not yet achieved Exit (referred to as "Participant-Non-Exit" hereinafter)).

As a result, in Individual Factors composed of Paper-related Features and Patent-related Features, remarkable differences were found between the positives and the negatives across research topics. Moreover, sequential or non-sequential differences regarding each variable's such characteristics across research topics were observed among the following three event status groups: Non-Participant, Participant-Non-Exit, and Exit (in other words, Participant who falls into Exit).

For visual comparison, by using the "ggplot" function of R's "ggplot2" package, this section plotted the summary statistics (the mean and plus/minus the SD) and created violin plots to visualize their distributions allowing for a deeper understanding of the density. Detailed descriptive statistics of these variables are presented in Appendix D.

For Paper-related Features

As apparent in Figure 6-7, compared to negative groups (i.e., Non-Participant/Non-Exit researchers), positive groups (i.e., Participant/Exit researchers) across **Cas9**, **Microbiome** and **5-Biopharma-Topics** have significantly larger means and SDs for most of the Paper-related Features universally: PUB, PAPER_CITED, PAPER_CITING, CORRESP_AUTH and COAUTH_ DEG_CENT. FIRST_AUTH is an only exception, in which the means are close across researcher groups, whereas SDs of positive groups are larger than those of negative groups. It is also observed that density distributions between the positive and the negative groups are clearly different.

However, when we compare means and SDs of researchers who are either Non-Participant, Participant-Non-Exit, or Exit, notable differences are observed across **Cas9**, **Microbiome** and **5-Biopharma-Topics**, as depicted in Figure 6-8. In **Cas9**, means and SDs of each Paper-related Feature increase in incremental steps from Non-Participant to Participant-Non-Exit to Exit, except for FIRST_AUTH. On the other hand, in **Microbiome**, such incremental increase is not observed in that researchers who achieved Exit do not necessarily tend to have higher counts than Participants without Exit, among Paper-related Features except for CORRESP_AUTH. This difference could

be due to **Microbiome**'s scientific taxonomy concept and its nascent application phase, as discussed in 6.2.2.1. **5-Biopharma-Topics** is similar to **Cas9** in that means and SDs increase from Non-Participant to Participant-Non-Exit to Exit except for FIRST_AUTH assumedly because its combined topics mitigates **Microbiome**'s peculiarity.

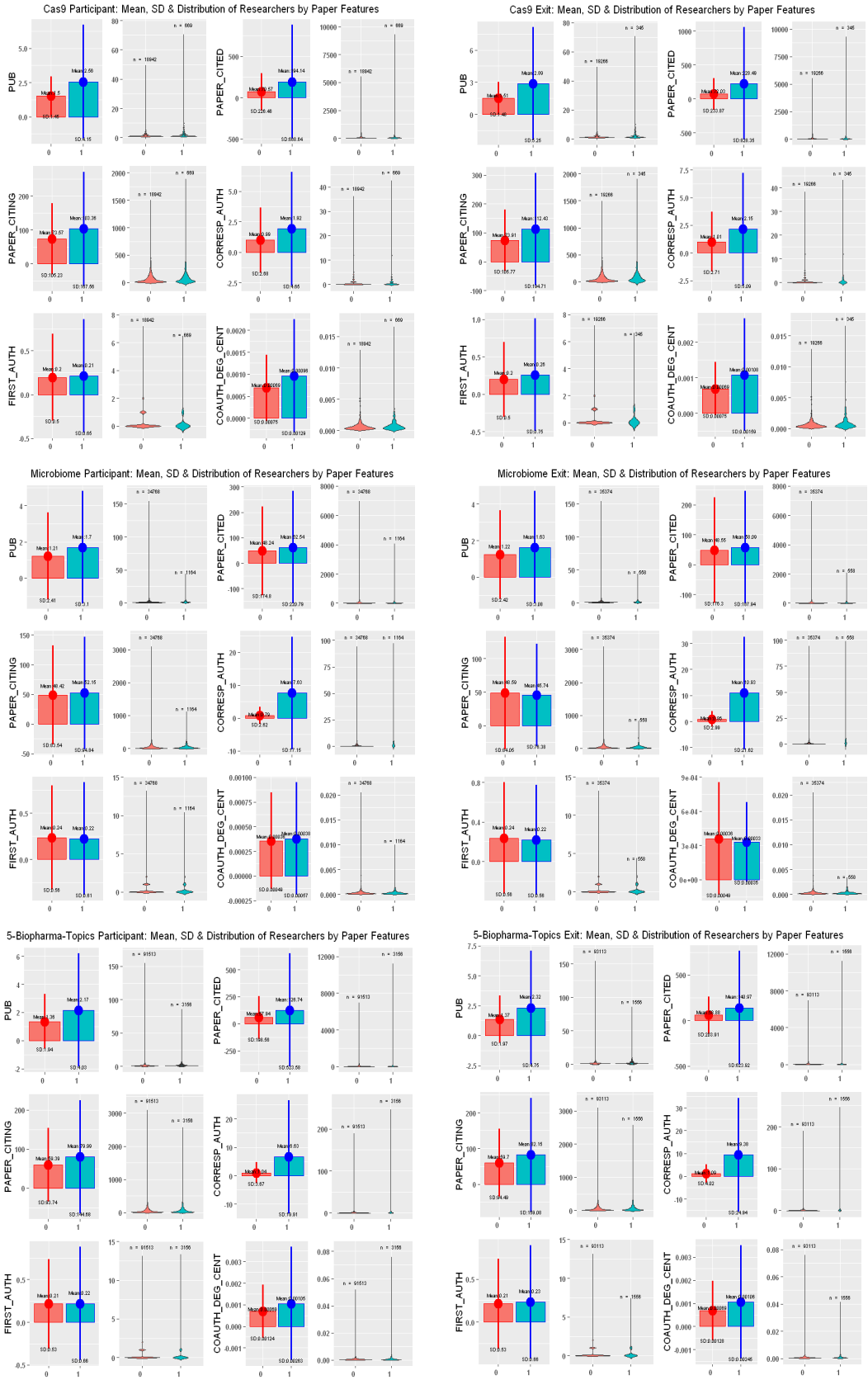
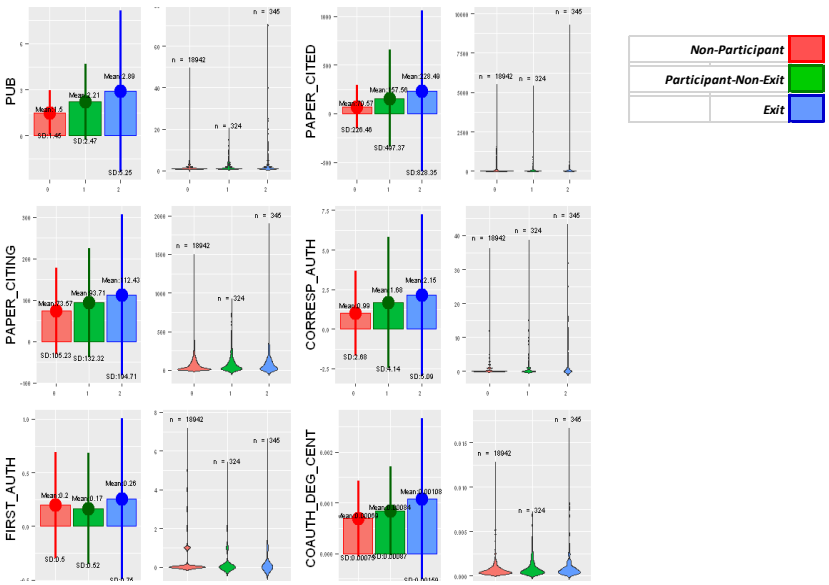
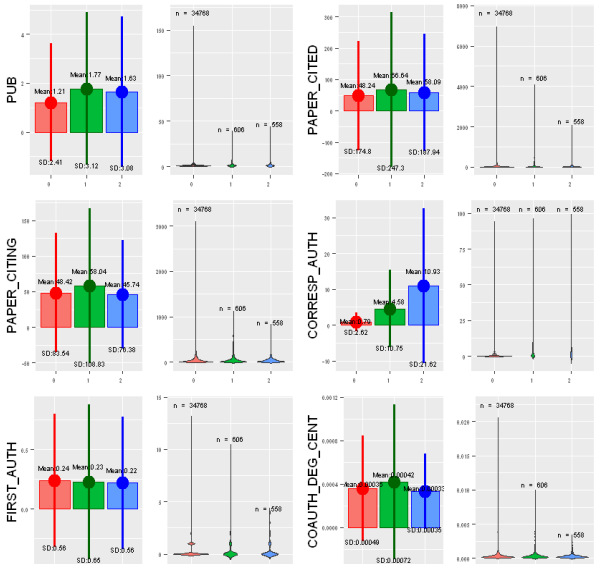


Figure 6-7 Mean, SD & Distribution of Paper-related Features per Researcher Group (Negatives vs Positives)

Cas9 Summary Stats among Non-Participant, Participant-Non-Exit & Exit Researchers



Microbiome Summary Stats among Non-Participant, Participant-Non-Exit & Exit Researchers



5-Biopharma-Topics Summary Stats among Non-Participant, Participant-Non-Exit & Exit

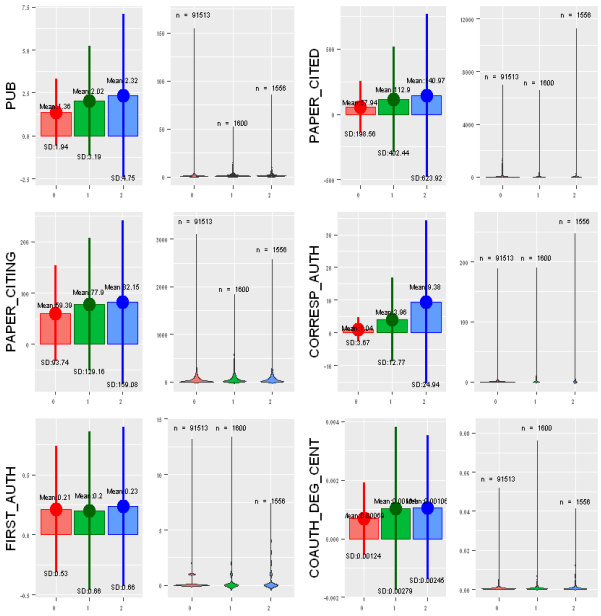


Figure 6-8 Mean, SD & Distribution of Paper-related Features per Researcher Group (Non-Participant, Participant-Non-Exit & Exit Researchers)

For Patent-related Features

As shown in Figure 6-9, we observe stark differences regarding Patent-related Features between the positives (Participant/Exit) and the negatives (Non-Participant/Non-Exit) belonging to **Cas9**, **Microbiome** and **5-Biopharma-Topics**. In all relevant research topics, positive group researchers have overwhelmingly larger means and SDs than negative group researches for all of the Patent-related Features: IP_NUM, PAPER_CITED_NUM_IN_IP, IP_CITED, and IP_CITING.

On the other hand, in comparison of means and SDs among researchers who are either Non-Participant, Participant-Non-Exit, or Exit, it is again notable that differences are observed across **Cas9**, **Microbiome** and **5-Biopharma-Topics** as presented in Figure 6-10. It is observable that, in **Cas9**, means and SDs of each Patent-related Feature increase in an incremental fashion from Non-Participant to Participant-Non-Exit to Exit, whereas, in **Microbiome**, such incremental increase is not observed; on the contrary researchers who achieved Exit have significantly smaller Patent-related Features with a narrower range than Participants without Exit, in all features. This difference could be attributable to importance of scientific achievement itself, rather than the degree of intellectual property development, when it comes to Exit in the **Microbiome** field as opposed to the **Cas9** field, as can be inferred from 6.2.2.1 later. **5-Biopharma-Topics** is overall similar to **Cas9,** in that means and SDs of Patent-related Features except for IP_NUM increase in a phased manner, from Non-Participant to Participant-Non-Exit to Exit. It is presumably because of its combined topics alleviating each topic's peculiarity like **Microbiome**'s.

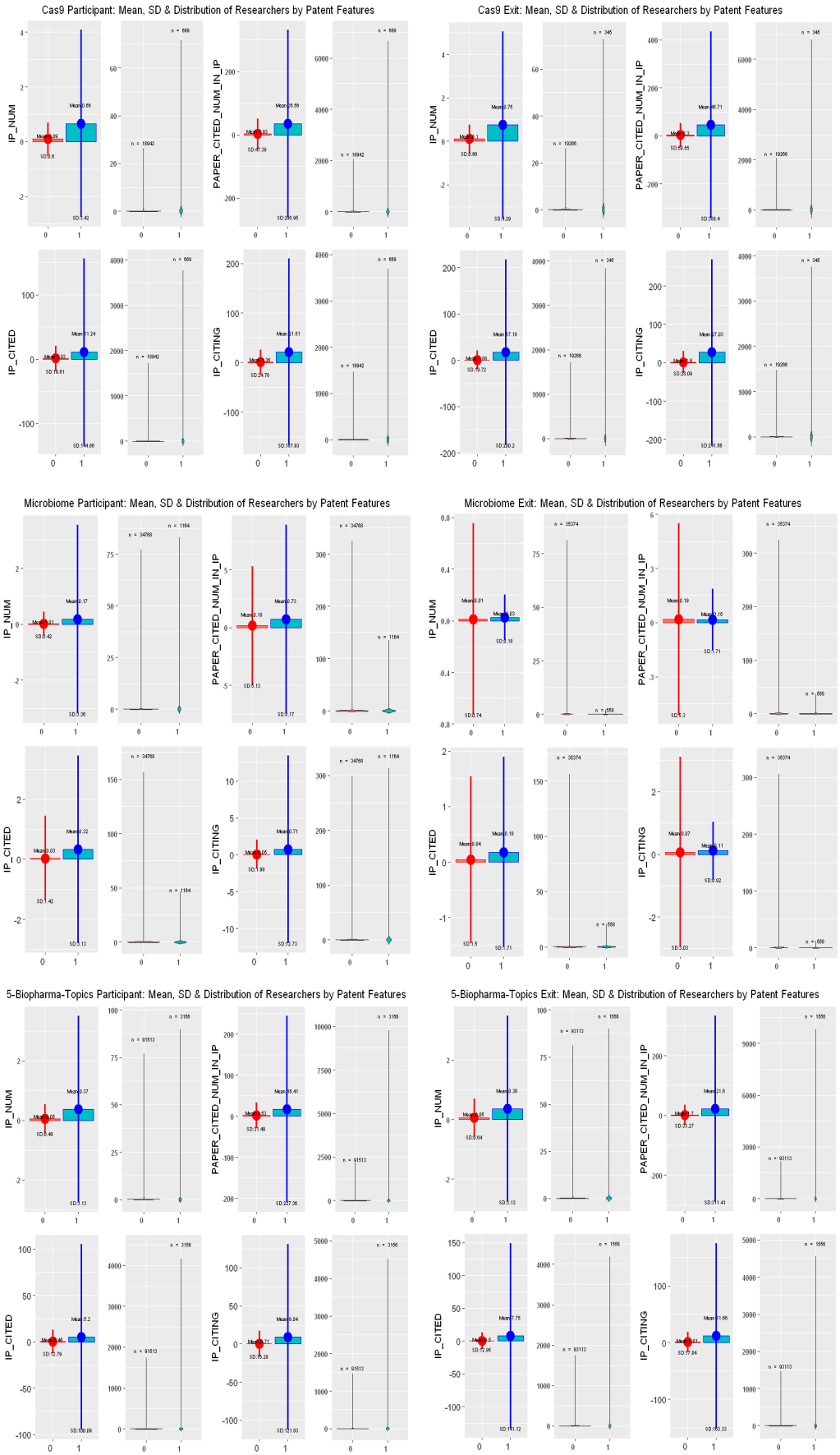
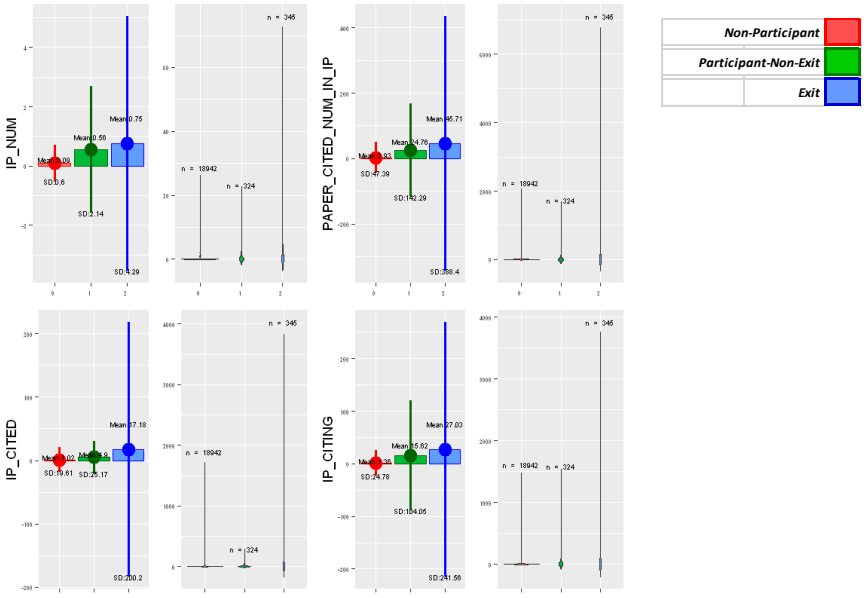
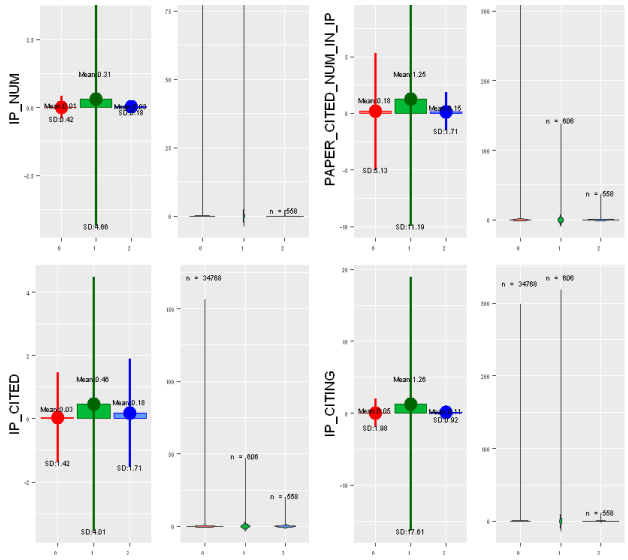


Figure 6-9 Mean, SD & Distrib. of Patent-related Features per Researcher Group (Negatives vs Positives)

Cas9 Summary Stats among Non-Participant, Participant-Non-Exit & Exit Researchers



Microbiome Summary Stats among Non-Participant, Participant-Non-Exit & Exit Researchers



5-Biopharma-Topics Summary Stats among Non-Participant, Participant-Non-Exit & Exit Researchers

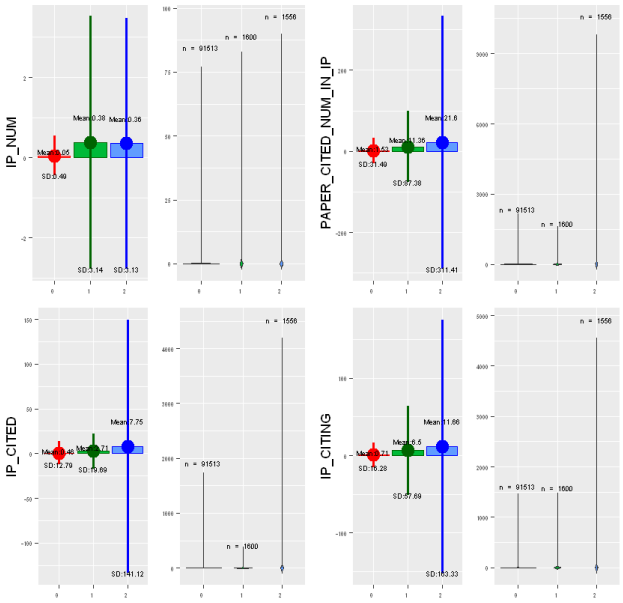


Figure 6-10 Mean, SD & Distribution of Patent-related Features per Researcher Group (Non-Participant, Participant-Non-Exit & Exit Researchers)

For Hot Topic Factors/Features

We can compare the summary statistics and the distribution of Hot Topics Features regarding each research topic: FINANCED_AMOUNT, FINANCED_FREQ, KW_GROWTH and IP_GROWTH between positive and negative researcher groups across **5-Biopharma-Topics**. Regarding solo **Cas9** and solo **Microbiome** researchers, Hot Topic Features cannot be taken into account, since the relevant researchers belong to the same research topics; preventing us from comparing researchers in different topics.

As shown in Figure 6-12, positive and negative researcher groups are very close in terms of the means and SDs regarding FINANCED_AMOUNT, FINANCED FREQ, and KW_GROWTH, whereas their density distributions are different from each other, which could produce different effects among these features as explanatory variables. With respect to IP_GROWTH, however, a difference between positive and negative researchers is recognized, in which positive researcher groups have larger means and SDs relative to negative researcher groups, albeit with smaller margins.

When it comes to comparison of means and SDs among researchers who are either Non-Participant, Participant-Non-Exit, or Exit, we do not observe any significant differences, albeit a small difference in IP_GROWTH between researcher groups belonging to Non-Participant and those belonging to Participant (including both Non-Exit and Exit equally), as shown in Figure 6-11

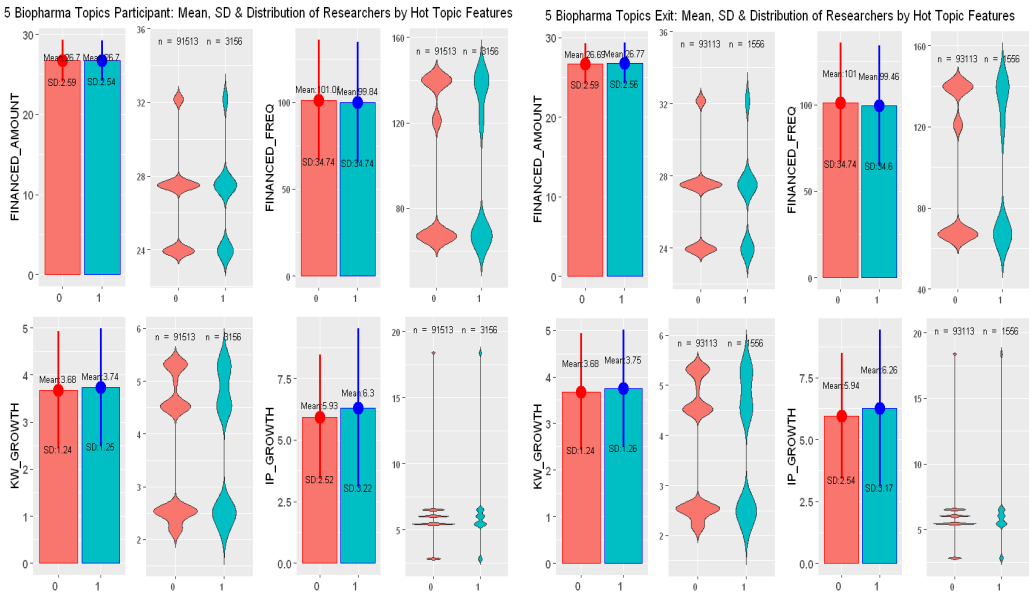


Figure 6-12 Mean, SD & Distrib. of Hot Topic Features per Researcher Group (Negatives vs Positives)

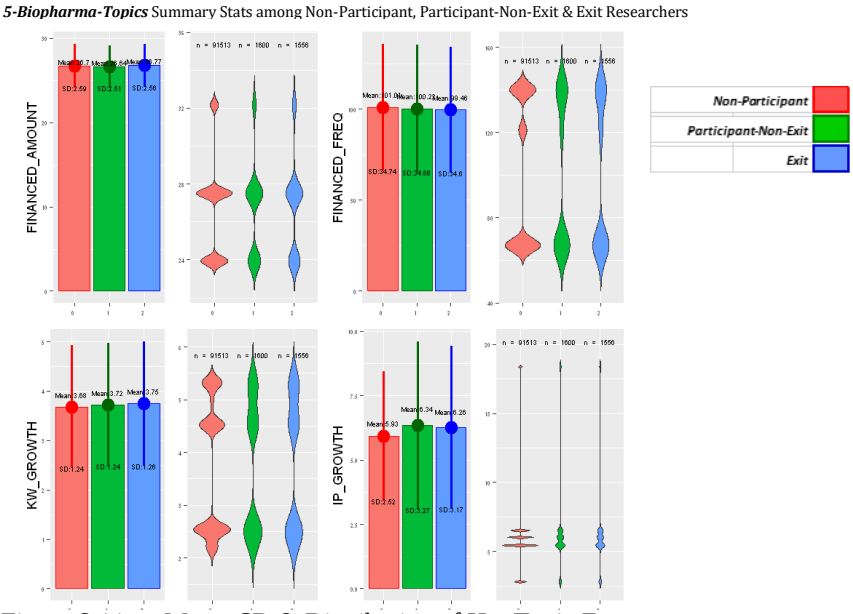


Figure 6-11 Mean, SD & Distribution of Hot Topic Factors per Researcher Group (Non-Participant, Participant-Non-Exit & Exit Researchers)

For Ecosystem Factors

As seen in Figure 6-13, among **Cas9**, **Microbiome** and **5-Biopharma-Topics**, positive researcher groups have larger means and SDs for most of the Ecosystem Factors: UNIV_SIZE, UNIV_RESEARCH, UNIV_INNOV, NATION_VC and NATION_TURNOVER, compared to negative researcher groups, albeit moderately. Their density distributions between the two researcher groups are also different. However, it is observed that NATION_STARTUP is the only exception, in which the means are closer across researcher groups, whereas SDs of positive researchers groups are discreetly larger compared to negative researcher groups. It is observable that the positive and negative groups have different density distributions, too.

Secondly, when comparing means and SDs among Non-Participant, Participant-Non-Exit, and Exit researchers, it is noteworthy that, **Cas9** and **Microbiome** relatively have similar characteristics in most of the Ecosystem Factors, as depicted in Figure 6-14. Both in **Cas9** and **Microbiome,** it is particularly observable that means and SDs of NATION_TURNOVER increase in an incremental fashion from Non-Participant to Participant-Non-Exit to Exit, whereas such incremental increase is not observed in UNIV_SIZE, UNIV_RESEARCH and NATION_STARTUP; on the contrary researchers who achieved Exit have moderately smaller values of the former two factors. This lets us infer that national professional voluntary turnover matters both for Participant and Exit, and that the size and the research level of academic organizations could be more beneficial to Participant, than to Exit, irrespective of the relevant research topics' characteristics as discussed in ***For Paper-related Features*** of this section. On the other hand, it is observable that, in **Cas9**, means of UNIV_INNOV and NATION_VC increase in an incremental fashion from Non-Participant to Participant-Non-Exit to Exit, whereas, in **Microbiome**, such incremental increase is not observed; on the contrary researchers who achieved Exit have moderately smaller values of UNIV_INNOV than Participants without Exit. **5-Biopharma-Topics** is similar to **Cas9,** in that means and SDs of Ecosystem Factors overall change in a similar manner, from Non-Participant to Participant-Non-Exit to Exit, assumedly because it is combined topics that mitigates each topic's peculiarity such as **Microbiome**'s.

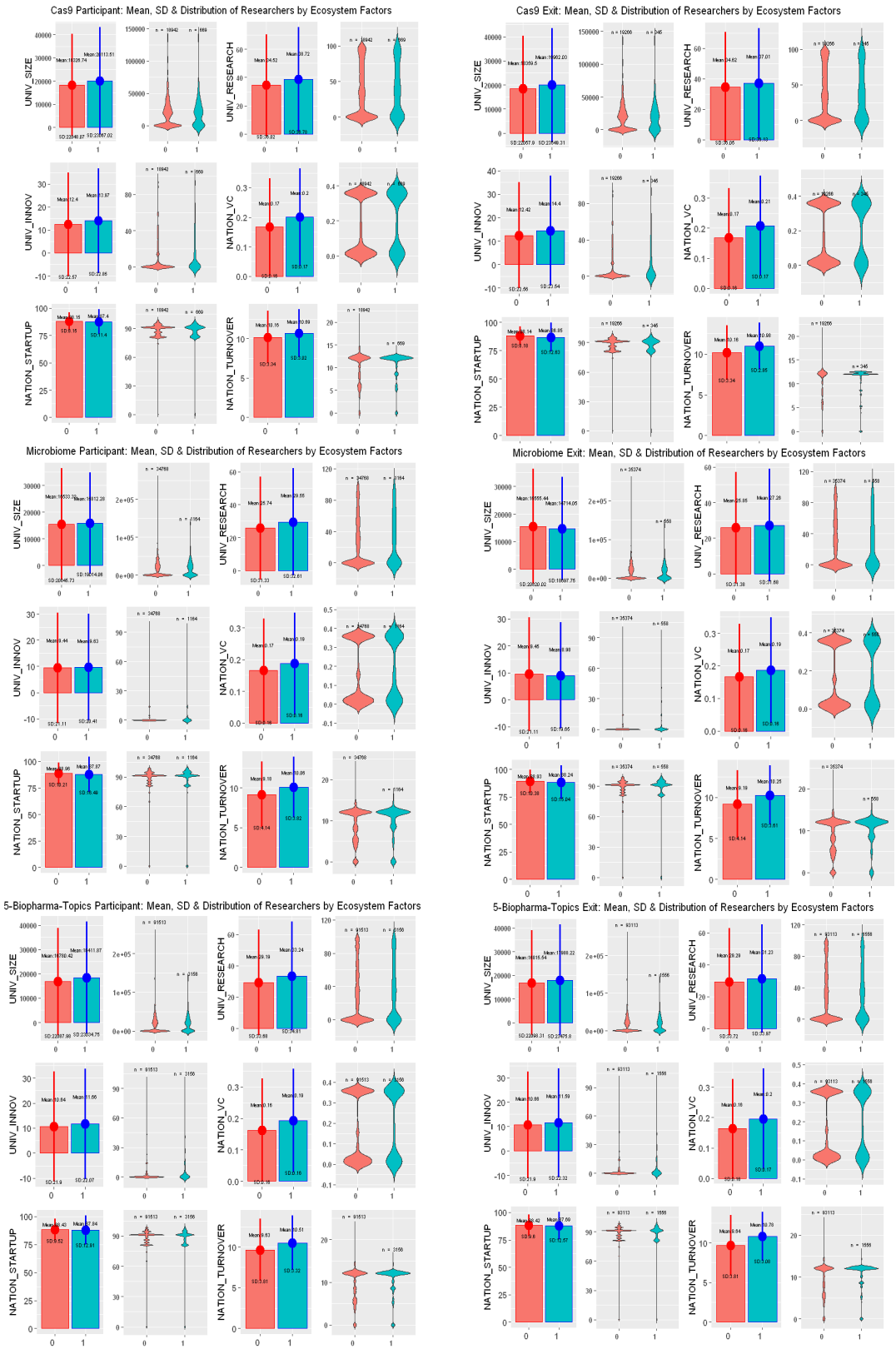
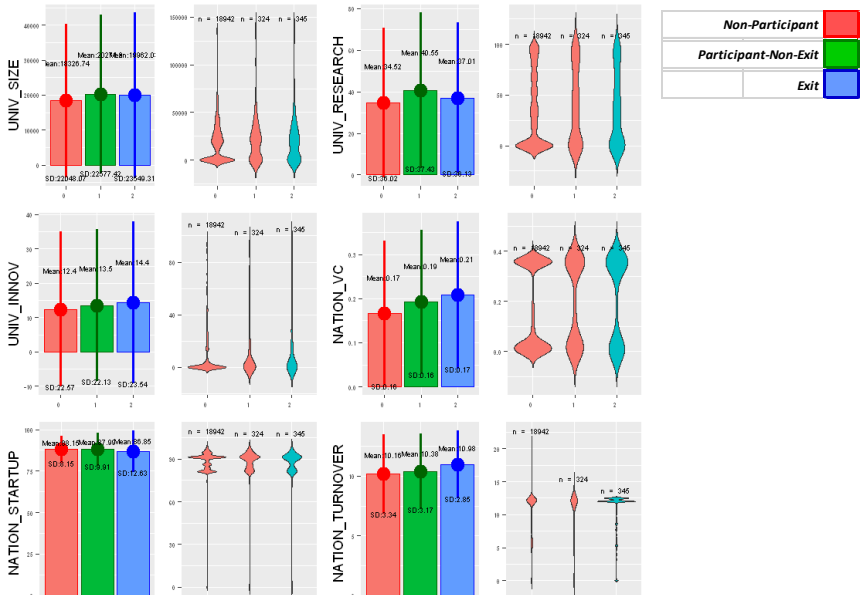
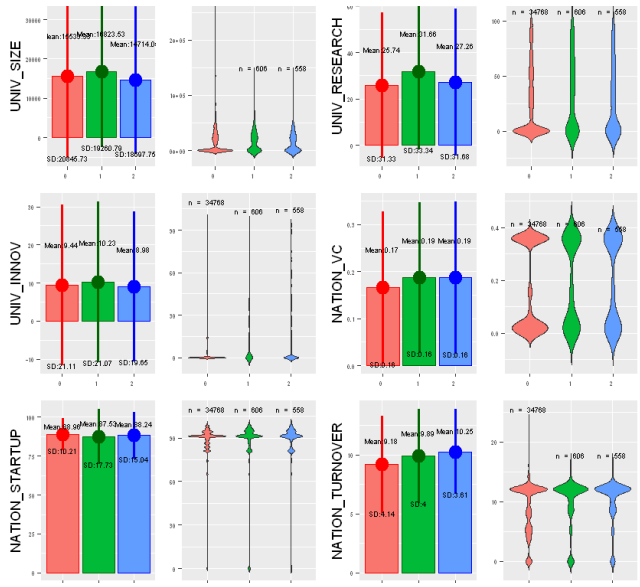


Figure 6-13 Mean, SD & Distrib. of Ecosystem Factors per Researcher Group (Negatives vs Positives)

Cas9 Summary Stats among Non-Participant, Participant-Non-Exit & Exit Researchers



Microbiome Summary Stats among Non-Participant, Participant-Non-Exit & Exit Researchers



5-Biopharma-Topics Summary Stats among Non-Participant, Participant-Non-Exit & Exit

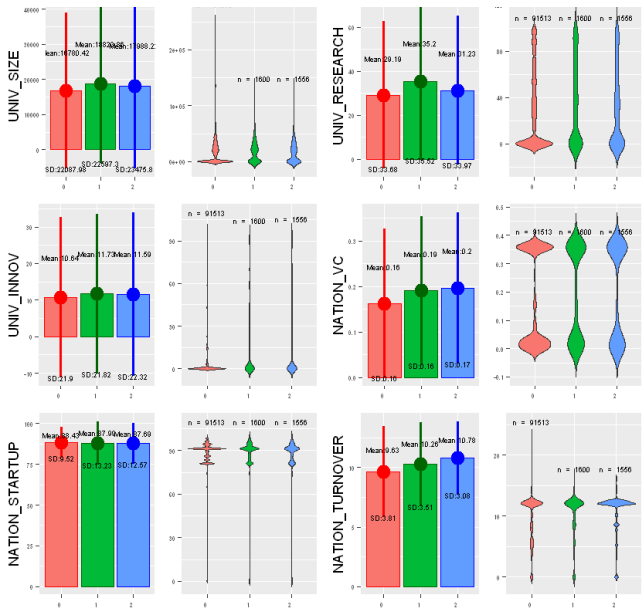


Figure 6-14 Mean, SD & Distribution of Ecosystem Features per Researcher Group (Non-Participant, Participant-Non-Exit & Exit Researchers)

6.2.2. Effective Explanatory Variables and Their Effects Across Researcher Groups' Assessment Models

Tables regarding explanatory variables and their effects on odds of startup readiness of Participant and Exit, for authors related to solo **Cas9** (Table 5.17 and Table 5.18), solo **Microbiome** (Table 5.19 and Table 5.20) and **5-Biopharma-Topics** (Table 5.22 and Table 5.23) show how a variety of groups of features work together to influence startup readiness of academic researchers, indicating effects of explanatory variables per each research researcher group. Given the descriptions in 5.3.1.2 and 5.3.2.2, comparison of these explanatory variables and their effects across various researcher groups' assessment models, is discussed herein.

Implications extracted from the results and the comparisons described in subsequent sections 6.2.2.1 and 6.2.2.2 are as follows.

Across Factors/Features

- Between Participant and Exit for each one of **Cas9**, **Microbiome** and **5-Biopharma-Topics**, a number of influential explanatory solo variables are common. This suggests that each researcher group has common influential explanatory solo variables regardless of Participant or Exit, albeit with a few peculiar influential explanatory variables for each of Participant and Exit.

For Paper-related Features

- For **5-Biopharma-Topics**, both CORRESP_AUTH (its MFP-transformed version; hereinafter the same applies to all cases of CORRESP_AUTH in this paragraph) and FIRST_AUTH (its MFP-transformed version; hereinafter the same applies to all mentions of FIRST_AUTH in this paragraph) are positively influential explanatory variables regardless of Participant or Exit. On the other hand, only CORRESP_AUTH is positively influential for Microbiome and so is the case of FIRST_AUTH for Cas9. It suggests that FIRST_AUTH could work more effectively for a research topic on a technology tool in practical application, whereas CORRESP_AUTH could work more effectively for a research topic on a genuinely scientific taxonomy concept.
- For **Cas9**, PUB (its MFP-transformed version; hereinafter the same applies to all instances of PUB in this paragraph) is a positively influential explanatory variable regardless of Participant or Exit, whereas for **Microbiome**, it does not work effectively. Rather, it does work against Exit for **5-Biopharma-Topics**. This suggests that PUB could work effectively for a research topic on a technology tool in practical application, whereas it could work against or could be in vain toward a research topic on a technology tool on more scientific concepts.

- For **Cas9**, CITATION_OUTDEG_CENT is a negatively influential explanatory variable regardless of Participant or Exit. This suggests that the more the academic researchers cite other researchers' papers, the lesser is the ***inventive step*** (inventiveness) achieved by them, because the research topic is on a technology tool in practical application and the ratio of inventors among relevant researchers is high.
- For **Microbiome**, FIRST_AUTH and CITATION_INDEG_CENT are positively influential explanatory variables for Exit. On the other hand, COAUTH_DEG_CENT is a negatively influential explanatory variable against Exit. It suggests that COAUTH_DEG_CENT could signal a lack of inventive step in a research topic on a genuinely scientific taxonomy concept, as CITATION_OUTDEG_CENT does so for **Cas9**.

For Patent-related Features

- For **Cas9** and **5-Biopharma-Topics**, IP_BINARY is a positively influential explanatory variable both for Participant and Exit. On the other hand, for **Microbiome**, IP_BINARY is not positively influential, whereas PAPER_CITED_BINARY_IN_IP is a positively influential explanatory variable regardless of Participant or Exit. This suggests that IP_BINARY could work effectively for startup readiness in (a) research topic(s) on technology with somewhat practical application, whereas PAPER_CITED_BINARY_IN_IP could do so for a research topic on genuinely scientific concept with less practical application.

For Ecosystem Factors

- For **Cas9, NATION_TURNOVER** is a positively influential explanatory variable for Exit. This suggests that nations' workforce voluntary turnover could effectively work for Exit in a research topic on technology with practical application.
- While **NATION_TURNOVER** per se is not a positively influential explanatory variable for Participant in either **Cas9**, **Microbiome** or **5-Biopharma-Topics,** the Interaction Terms Factor composed of NATION_TURNOVER and KW_GROWTH are positively influential for Participant in **5-Biopharma-Topics**. This suggests that, nations' workforce voluntary turnover, in combination with a certain range of research topics with growing appearance frequency, attracts human talent for startup creation.
- For **5-Biopharma-Topics,** NATION_STARTUP (its MFP-transformed version; hereinafter the same applies in this paragraph) is a positively influential explanatory variable for Participant, while so is NATION_VC (its MFP-transformed version; hereinafter the same applies in this paragraph) for Exit. This suggests that, across the highest growth biopharmaceutical research topics overall, nations' ease of starting a

business could effectively work for startup participation, while so does nations' ease of raising venture capital funding for exit.

- Any Academic Organization-related Features per se are not positively influential explanatory variables regardless of Participant or Exit in either **Cas9**, **Microbiome** or **5-Biopharma-Topics**, while the Interaction Terms Factor composed of UNIV_RESEARCH and PAPER_CITED_NUM_IN_IP and that composed of UNIV_RESEARCH and NATION_VC become positively influential explanatory variables for Participant in **5-Biopharma-Topics**, although UNIV_RESEARCH does not even become a component of Interaction Terms Factors for Exit in either **Cas9**, **Microbiome**, or **5-Biopharma-Topics**. This implies that, the third parties assessing academic researchers, when evaluating their potential of Exit, would rather pay attention to those researchers' individual scientific prominence and innovation capability, than to the research score of the academic organizations that they belong to. In other words, it is suggested that the research score of academic organizations could not effectively work for exit, whereas it could for startup participation.

6.2.2.1. Effective Explanatory Variables and Their Effects in the Assessment Models of Solo Cas9 and Solo Microbiome

Although both **Cas9** and **Microbiome** belong to the biopharmaceutical domain in the broad sense, the two research topics are clearly different from each other as follows.

First of all, types of research topic concept of these two are not the same as already seen in 3.2.1 (ii): **Cas9** refers to a specific technology tool that is the CRISPR-associated protein 9 that specifies the critical cleavage site of DNA sequence, whereas **Microbiome** means communities of microorganisms in and on organisms, not a technology tool. Moreover, the two topics are contrastive in terms of the phase and the degree of practical, industrial application too: in the **Cas9** field 5.56% of the relevant authors are inventors of patents, while in the **Microbiome** field only 0.55% of the relevant authors are inventors. (Appendix C-1 and Appendix C-2)

Since such differences assumingly lead to the difference of explanatory variables and their effects between the assessment models of **Cas9** and those of **Microbiome**, explanatory variables with high and low effect for each research topic in terms of the exponentially transformed coefficients over 1.100 (defined as "effective" herein) and below 0.900 (defined as "counter-effective" herein) are analyzed as substantially effective and counter-effective respectively, as follows.

(i) Paper-related Features

Participant

In **Cas9,** PUB_MFP and FIRST_AUTH_MFP are effective at 1.483 and at 1.244 respectively, while in **Microbiome** CORRESP_AUTH_MFPp is overwhelmingly effective at 1.846. Because **Microbiome** is a broad, non-technical taxonomy concept with a lesser degree of practical application as opposed to **Cas9**, authors who participate in startups in this field should arguably be able to fulfil academic responsibility to back the venture, which makes CORRESP_AUTH_MFPp matter a lot. In the life sciences field, researchers who supervise first authors, who are typically established researchers with more academic responsibilities (such as so-called Principal Investigators), tend to be corresponding authors. On the other hand, since **Cas9** is a specifically technical research topic with high degree of practical and industrial application, academic productivity, creativity and proactiveness matter compared to academic responsibility, supposedly leading to PUB_MFP and FIRST_AUTH_MFP as effective variables.

Simultaneously, in **Cas9,** CITATION_OUTDEG_CENT (index of an author's proactiveness to refer to prior research) is found counter-effective at 0.817, while in **Microbiome** no Paper-related Features are found substantially counter-effective. This indicates that, in a specifically technical research topic with advanced application level like **Cas9**, high tendency to refer to others' work rather than relying on own creativity, could signal counter-effect against startup orientation.

Exit

Similarly, in **Cas9,** PUB_MFP and FIRST_AUTH_MFP are effective at 1.521 and at 1.301 respectively, while in **Microbiome,** CORRESP_AUTH_MFPe and CITATION_INDEG_CENT are effective at overwhelming 1.984 and at 1.126. Similar to the reason seen in Participant above, in **Microbiome**, authors participating in startups are expected to be able to execute academic responsibility, which makes CORRESP_AUTH_MFPe highly effective. Additionally, since Exit is an event that will not occur without evaluation of third parties, a higher profile in the academic community is important, thereby supposedly making CITATION_INDEG_CENT effective. On another front, regarding **Cas9**, exactly as seen in Participant above, academic productivity, creativity and proactiveness are important, thus causing PUB_MFP and FIRST_AUTH_MFP to be substantially effective.

Parallelly, in **Cas9,** CITATION_OUTDEG_CENT is coincidentally found counter-effective again at 0.817, while in **Microbiome,** COAUTH_DEG_CENT is found counter-effective at 0.863. The former result suggests that, just as seen in Participant, in a specifically technical research topic with advanced application level like **Cas9,** higher tendency to refer to prior research could work against Exit too. The latter result presumably shows that for researchers with a genuine taxonomy research topic with lower application level like **Microbiome**, high centralities in co-authorship

networks could signal researchers' strong presence in the research community at the expense of commercialization and/or entrepreneurship, thus suppressing Exit.

(ii) Patent-related Features

Participant

In **Cas9**, IP_BINARY and IP_NUM are effective at 1.220 and at 1.117 respectively, whereas in **Microbiome**, PAPER_CITED_BINARY_IN_IP (index of whether (an) academic paper(s) is (are) cited in an author's invented patent) is effective at 1.144 without IP_BINARY and IP_NUM. This difference can be attributable to the contrast between **Cas9** and **Microbiome**, in a sense that existence and strength of patent protection are effective for startup readiness to commercialize a tool, whereas a bridge from scientific research to application is important for startup readiness based on a non-technical taxonomy concept.

No solo Patent-related Features are found counter-effective in **Cas9** or **Microbiome**.

Exit

There are no Patent-related Features that are effective solely in **Cas9**, whereas in **Microbiome**, PAPER_CITED_BINARY_IN_IP is effective at 1.210. The same explanation can be applied to **Microbiome** as seen in Participant. As opposed to Participant, however, IP_BINARY and IP_NUM do not matter for **Cas9**, assumingly because, for Exit, third parties evaluate qualitative features of patents rather than IP_BINARY and IP_NUM.

Concurrently, while no solo Patent-related Features are found counter-effective in **Cas9**, in **Microbiome**, PAPER_CITED_NUM_IN_IP (index of frequency of academic papers being cited in an author's invented patent) is found counter-effective at 0.662. Although patent itself is a form of protection that provides researchers exclusive rights to commercialize, for researchers with a genuine taxonomy research topic with lower application level like **Microbiome**, it is inferred that frequency of academic papers being cited in patent could represent traditional values of scientific research, discouraging commercialization and/or entrepreneurship of relevant researchers.

(iii) Ecosystem Factors

Participant

While **Cas9** and **Microbiome** share the same variable NATION_VC (index of venture capital investment relative to their GDP of relevant countries) being effective at 1.223 and at 1.134 respectively, **Microbiome** has another variable UNIV_RESEARCH (index of research score of relevant academic organizations) being more effective at 1.264 additionally. It is quite natural that NATION_VC is effective both in Cas9 and

Microbiome. Moreover, since **Microbiome** is assumedly more basic research-oriented than **Cas9**, it seems natural that UNIV_RESEARCH matters for startup readiness for **Microbiome.**

Simultaneously, while no Ecosystem Factors are found substantially counter-effective in **Cas9,** UNIV_SIZE (index of the size of relevant academic organizations in terms of the number of students) is found counter-effective at 0.846 in **Microbiome.** This indicates that, for researchers with less application like **Microbiome**, size of academic organizations they belong to could be counter-effective against startup orientation.

Exit

In **Cas9,** NATION_TURNOVER (index of life science workforce voluntary turnover in relevant countries) and NATION_VC are effective at 1.261 and 1.257 respectively, whereas in **Microbiome** there are no solo Ecosystem Factors being substantially effective. As opposed to **Microbiome**, in **Cas9**, values of NATION_TURNOVER and NATION_VC arguably have immediate effect on startup readiness, since **Cas9** is not only a research topic but also already a technical tool for commercialization. NATION_TURNOVER matters only in Exit as opposed to Participant, presumably because voluntary mobility of life science workforce is the key to make biopharmaceutical startups grow smoothly enough to be acquired

Parallelly, in **Cas9,** NATION_STARTUP (index of score for starting business in relevant countries) is found counter-effective at 0.856, while in **Microbiome,** NATION_VC_MFPe is found counter-effective at 0.782, both of which should be counterintuitive. Inferably, the former is because **Cas9** startups tend to need more capital than regularly founded startups can afford due to their capital intensiveness, thus making values of NATION_STARTUP and Participant counter-effective. The latter is also inferably because, in nations with high values of NATION_VC, venture capitalists and third parties are especially selective in choosing **Microbiome** startups for potential Exit, thus making values of NATION_VC_MFPe and Exit counter-effective.

(iv) Interaction Terms Factors

Participant

Cas9 has the following effective Interaction Terms Factors (ITFs): FIRST_AUTH_MFP * IP_BINARY at 1.149 and FIRST_AUTH_MFP * NATION_STARTUP at 1.120, both of which include FIRST_AUTH_MFP that is an effective feature solely as a Paper-related Feature too as described in (i). On the other hand, **Microbiome** has no substantially effective Interaction Terms Factors.

Simultaneously, FIRST_AUTH_MFP * NATION_VC is found counter-effective at 0.849 in **Cas9**, while in **Microbiome** NATION_VC * UNIV_RESEARCH is found counter-effective at 0.877 too, both ITFs of which are composed of two effective variables that are already seen in (i) Paper-related Features and (iii) Ecosystem Factors. These ITFs seemingly work as weight against their component solo variables which are substantially effective.

Exit

In **Cas9,** only two ITFs: FIRST_AUTH_MFP * IP_BINARY and CORRESP_AUTH * CITATION_OUTDEG_CENT are effective at 1.169 and 1.151 respectively, whereas in **Microbiome**, two ITFs which are combinations of two different Paper-related Features (CITATION_OUTDEG_CENT * CITATION_INDEG_CENT and FIRST_AUTH * COAUTH_DEG_CENT), and, another two ITFs composed of a Paper-related Feature and an Ecosystem Factor (CITATION_INDEG_CENT * UNIV_INNOV and FIRST_AUTH * NATION_VC_MFPe), are effective at 1.697 and 1.506, as well as at 1.181 and 1.114, respectively. Regarding **Cas9**, while the former component of the first ITF (i.e. FIRST_AUTH_MFP) is itself a substantially effective variable and the latter component of the second ITF (i.e. CITATION_OUTDEG_CENT) is also itself a substantially counter-effective variable, both of which are Paper-related Features, the other components of these two ITFs (i.e., IP_BINARY and CORRESP_AUTH) are not statistically significant. Regarding **Microbiome**, since the research topic is a broad non-technical scientific taxonomy concept with a lower degree of practical application, the relevant ITFs are mostly comprised of several Paper-related Features: CITATION_OUTDEG_CENT (index of an author's proactiveness to introduce prior research), CITATION_INDEG_CENT (index of scientific attention an author receives), FIRST_AUTH (index of an author's activeness and creativeness) and COAUTH_DEG_CENT (index of overall centrality among co-authorship networks) (See 4.4.1.1), all of which are factors of academic researchers' scientific prominence, while only two Ecosystem factors included herein are indexes indicating how much academic researchers' ecosystem is favorable for their startup activities: UNIV_INOV (an Academic Organization-related Feature, index of innovativeness of relevant academic organizations) and NATION_VC_MFPe. Among these components of ITFs, only CITATION_INDEG_CENT is a substantially effective variable.

Simultaneously, in **Cas9**, four ITFs are found counter-effective at 0.854, 0.828, 0.817 and 0.810, three of which include CITATION_OUTDEG_CENT, the most counter-effective Paper-related Feature as a component as described in (i): CITATION_OUTDEG_CENT * PAPER_CITED_BINARY_IN_IP, FIRST_AUTH_MFP * CITATION_OUTDEG_CENT, PUB_MFP * CITATION_OUTDEG_CENT and

NATION_VC * FIRST_AUTH_MFP respectively. Likewise, in **Microbiome**, four ITFs are found counter-effective, all of which are combinations of two Paper-related Features: CORRESP_AUTH_MFPe * CITATION_INDEG_CENT, CORRESP_AUTH_MFPe * CITATION_OUTDEG_CENT, CITATION_OUTDEG_CENT * FIRST_AUTH, and CITATION_INDEG_CENT * COAUTH_DEG_CENT, at 0.841, 0.796, 0.672 and 0.483 respectively. Inferably, these ITFs are substantially negatively effective because of less self-reliant CITATION_OUTDEG_CENT as described in (i), as well as the rest of the Paper-related Features emphasizing strong academic commitment that could discourage Exit herein.

It is observed that in Exit, a higher number of ITFs work compared to those in Participant, presumably because Exit occurs based on evaluation of third parties who take into account, and are influenced by, more objective features than academic researchers themselves, thereby making a higher number of Paper-related and Ecosystem Features effective.

6.2.2.2. Effective Explanatory Variables and Their Effects in the Assessment Models of 5-Biopharma-Topics

As seen in the above **Cas9** and **Microbiome** cases, each growing research topic in the biopharmaceutical domain, could have different types of concepts in terms of whether it is relevant to a technology tool and how much their degree of application is advanced, causing effective variables of their assessment models to diversify. Thus, the assessment model regarding one specific research topic cannot be expected to well-explain startup readiness and effective variables of academic researchers in another research topic. For example, the assessment model of **Cas9** will not be able to explain the mechanism of startup readiness and factors regarding **Microbiome** researchers. Therefore, in order to mitigate such limitations, building the assessment model based on a wider range of growing biopharmaceutical research topics is needed, which can work in a more effective manner across broad-based biopharmaceutical research topics. The assessment model regarding **5-Biopharma-Topics** was constructed for this purpose. Regarding the degree of practical, industrial application of relevant research topics overall, the ratio of inventors of patents over all relevant authors is 3.05% (Appendix C-3), which is between 5.55% of **Cas9** and 0.55% of **Microbiome** as seen before.

Explanatory variables with higher effect in terms of the exponentially transformed coefficients over 1.100 are analyzed as follows.

(i) Paper-related Features

Participant

CORRESP_AUTH_MFPp and FIRST_AUTH_MFPp are effective at 1.434 and 1.143 respectively, while another MFP-transformed version of CORRESP_AUTH is also found substantially effective in **Microbiome** Participant variables as well, as described in 6.2.2.1.

Concurrently, PUB_MFPp is found counter-effective at 0.785, whereas another MFP-transformed version of PUB is not found counter-effective in either **Cas9** Participant or **Microbiome** Participant variables.

Exit

CORRESP_AUTH_MFPe and FIRST_AUTH_MFPe are substantially effective at 1.542 and 1.139 respectively, while another MFP-transformed version of CORRESP_AUTH is found in **Microbiome** Exit variables and another MFP-transformed version of FIRST_AUTH in **Cas9** Exit variables as substantially effective ones, no matter what MFP-transformation, as seen in 6.2.2.1. Coincidentally, other MFP-transformed versions of CORRESP_AUTH and FIRST_AUTH are also seen in the substantially effective variables of Participant above.

Simultaneously, PUB_MFPe is found counter-effective at 0.844, while another MFP-transformed version of PUB is found counter-effective in the former component of PUB_MFP * CITATION_OUTDEG_CENT of **Cas9** Exit variables. Coincidentally, it is observed that another MFP-transformed version of PUB is also substantially counter-effective in Participant above.

(ii) Patent-related Features

Participant

IP_BINARY and IP_CITED are effective at 1.147 and 1.119 respectively. Although the former IP_BINARY is found substantially effective in **Cas9** Participant variables as described in 6.2.2.1, IP_CITED is not found effective in Participant variables of either **Cas9** or **Microbiome**.

No solo Patent-related Features are found counter-effective herein.

Exit

IP_BINARY is effective at 1.172, which is also seen as a component of an effective Interaction Terms Factor: FIRST_AUTH_MFP * IP_BINARY of **Cas9** Exit variables. IP_BINARY is also shared as an effective variable with Participant as described above.

No solo Patent-related Features are found counter-effective herein as in Participant.

(iii) Hot Topic Factors

There are no solo Hot Topic Factors found substantially effective in particular for either Participant or Exit.

(iv) Ecosystem Factors

Participant

NATION_STARTUP_MFPp (an MFP-transformed Nation-related Feature, index of easiness to start business of relevant countries) is effective at 1.163, while another MFP-transformed version of NATION_STARTUP is also found as a component of an effective Interaction Terms Factor: FIRST_AUTH_MFP * NATION_STARTUP among **Cas9** Participant variables.

Simultaneously, NATION_VC_MFPp is found counter-effective at 0.852. Actually, NATION_VC is a component of a counter-effective Interaction Terms Factor NATION_VC * UNIV_RESEARCH of **Microbiome** Participant variables seen in 6.2.2.1.

Exit

NATION_VC is effective at 1.246, which is also found as an effective variable of **Cas9** Exit, and as a component of an effective Interaction Terms Factor among **Microbiome** Exit variables: FIRST_AUTH * NATION_VC_MFPe in its MFP-transformed form. In contrast to Participant in which the ease of starting a business (i.e. NATION_STARTUP_MFPp) works as an effective variable for startup readiness as shown above, it is suggested that, in order to expedite Exit, national venture capital environment is quite important.

No solo Ecosystem Factors are found counter-effective herein.

(v) Interaction Terms Factors

Participant

CORRESP_AUTH_MFPp * IP_CITED, FIRST_AUTH_MFPp * NATION_STARTUP_MFPp, COAUTH_DEG_CENT_MFPp * NATION_STARTUP_MFPp, PAPER_CITED_NUM_IN_IP * UNIV_RESESARCH, NATION_VC_MFPp * UNIV_RESEARCH, and KW_GROWTH * NATION_TURNOVER_MFPp are effective as Interaction Terms Factors ("ITFs" hereinafter), at 1.635, 1.176, 1.153, 1.140, 1.139, and 1.112 respectively. Among the components of these ITFs in their pre-MFP-transformed forms, CORRESP_AUTH, FIRST_AUTH, NATION_STARTUP, PAPER_CITED_NUM_IN_IP, UNIV_RESESARCH, and NATION_VC are found as substantially effective Participant variables or their components in **Cas9** and **Microbiome**, no matter what MFP-transformed forms or not they take, as observed in 6.2.2.1.

Simultaneously, there are 10 counter-effective ITFs: CORRESP_AUTH_MFPp * IP_NUM at 0.899, IP_CITED * NATION_STARTUP_MFP at 0.887, PUB_MFPp * KW_GROWTH at 0.886, NATION_STARTUP_MFPp * UNIV_INNOV at 0.884,

NATION_TURNOVER_MFPp * NATION_STARTUP_MFPp at 0.858, NATION_VC_MFPp * NATION_STARTUP_MFPp at 0.846, FINANCED_AMOUNT * NATION_TURNOVER_MFPp at 0.840, IP_CITED * UNIV_RESEARCH at 0.822, PAPER_CITED_NUM_IN_IP * IP_CITING at 0.692. Regarding components of these ITFs, while PUB_MFPp and NATION_VC_MFPp are the sole counter-effective variables as seen in (i) and (iv), NATION_VC and UNIV_RESEARCH are found as a component of counter-effective ITFs irrespective of their MFP-transformation either in **Cas** or **Microbiome** as seen in 6.2.2.1.

Exit

FIRST_AUTH_MFPe * NATION_STARTUP_MFPe, CORRESP_AUTH_MFPe * IP_NUM, and COAUTH_DEG_CENT * FINANCED_AMOUNT are effective as ITFs at 1.205, 1.123 and 1.103. No matter what MFP-transformed forms taken, FIRST_AUTH, NATION_STARTUP, CORRESP_AUTH, and COAUTH_DEG_CENT are also found as substantially effective Exit variables or their components in **Cas9** and **Microbiome** as seen in 6.2.2.1. They are overlapped in substantially effective variables of Participant as above.

In parallel, CORRESP_AUTH_MFPe * KW_GROWTH is found counter-effective at 0.826. Another MFP-transformed version of CORRESP_AUTH is also found as a component of counter-effective ITFs in **Microbiome** as seen in 6.2.2.1.

6.2.3. Importance of Each Set of Factors/Features for Assessment

To validate the features employed in this thesis, this section compares the AUC's (areas under the curve) of ROC (receiver operating characteristic) curves related to different factor/feature sets, in terms of the types of factors/features, for the purpose of assessing startup readiness regarding Participant and Exit by the assessment model of this thesis, as follows: (i) "Whole Set": a whole set of Paper-related Features and Patent-related Features of Individual Factors, Hot Topic Factors, Ecosystem Factors (composed of Academic Organization-related Features and Nation-related Features, combined so due to the insufficiency of either type of features), and Interaction Terms Factors, as described in 4.4, (ii) "Except Paper": Whole Set minus Paper-related Features, (iii) "Except Patent": Whole Set minus Patent-related Features, (iv) "Except Hot Topic": Whole Set minus Hot Topic Factors, (v) "Except Ecosystem": Whole Set minus Ecosystem Factors, and (vi) "Except Interaction": Whole Set minus Interaction Terms Factors. All the above feature sets are considered for academic researchers across the five biopharmaceutical topics combined. For academic researchers of solo **Cas9** and solo **Microbiome**, however, due to its lack of hot topic features for comparison since all

researchers are in the same **Cas9** or **Microbiome** category, Hot Topic Features are removed from Whole Set, thus Except Hot Topic is not necessary. Following are the results of AUC's for each feature set, regarding Participant and Exit, for academic researchers of solo **Cas9** and solo **Microbiome** as well as those of **5-Biopharma-Topics**.

Firstly, as shown in Table 6.2, Whole Set achieves the highest AUC among all feature sets in all instances (0.663 for **Cas9** Participant researchers, 0.703 for **Cas9** Exit researchers, 0.741 for **Microbiome** Participant researchers, 0.773 for **Microbiome** Exit researchers, 0.690 for **5-Biopharma-Topics** Participant researchers, and 0.723 for **5-Biopharma-Topics** Exit researchers). Except Paper, Except Patent, Except Hot Topic, Except Ecosystem and Except Interaction render smaller AUC values than that of Whole Set in any event. Thus, we can infer that using all the types of explanatory variables as proposed in 4.4 is more valid to assess startup readiness of academic researchers, than using just parts of them.

Among all sets of variables/features, Paper-related Features play a critical role to assess both Participant and Exit, as the AUC's values of Except Paper are valued at or close to the lowest for all groups of researchers (0.607 for **Cas9** Participant researchers, 0.632 for **Cas9** Exit researchers, 0.564 for **Microbiome** Participant researchers, 0.536 for **Microbiome** Exit researchers, 0.598 **for 5-Biopharma-Topics** Participant researchers, and 0.582 for **5-Biopharma-Topics** Exit researchers), among Except Paper, Except Patent, Except Hot Topic, Except Ecosystem and Except Interaction. In particular, it is noteworthy that Except Paper renders remarkably low AUC values for assessing Exit among **Microbiome** and **5-Biopharma-Topics** researchers, of which value (0.536 and 0.582 respectively) are considered almost meaningless as a classifier.

It is obviously worth noting that, **Microbiome** researchers are, whether Participant or Exit, assessed with the best performance (i.e., the best AUC values) by the assessment model's explanatory variables (other than Except Paper), followed by **5-Biopharma-Topics** researchers and **Cas9** researchers in this order. **Microbiome** is a genuinely scientific taxonomy concept that is not as much advanced in practical application as **Cas9**, and is not a technology concept like **Cas9** (See 6.2.2.1). This enables us to assess startup readiness more appropriately by calculating Paper-related Features above all features, compared to **Cas9** that is a concept of technology tool. Furthermore, although **5-Biopharma-Topics**'s AUC values are somewhat behind those of **Microbiome**, they considerably surpass those of **Cas9** whether Participant or Exit, because **5-Biopharma-Topics** cover a wider range of academic researchers with richer features including Hot Topic Factors and Interaction Terms Factors than **Cas9** does, thereby making the assessment model perform better.

Comparatively speaking, between Participant researchers and Exit researchers, for all solo **Cas9**, solo **Microbiome** and **5-Biopharma-Topics**, the AUC values for Exit researchers are higher than Participant researchers, with only exception for the AUC value for Except Paper of Exit researchers regarding solo **Microbiome** and **5-Biopharma-Topics**. We can infer that it is because (i) more complete data could be available for Exit researchers than for Participant researchers, which could enable Exit researchers to be assessed more precisely than Participant researchers in most cases, since being a Participant by one's own will is not easily recognized by third parties, and because (ii) more standardized and consistent data could be attained for Exit researchers than for Participant researchers, since exits are reviewed and executed by third parties who could have standardized, consistent evaluation standards that are scalable.

Table 6.2 AUC's of Each Set of Features to Assess Startup Readiness for Researchers of Cas9, Microbiome and 5-Biopharma-Topics

	Cas9		Microbiome		5-Biopharma-Topics	
	Participant	Exit	Participant	Exit	Participant	Exit
(i) Whole Set	0.663	0.703	0.741	0.773	0.690	0.723
(ii) Except Paper	0.607	0.632	0.564	0.536	0.598	0.582
(iii) Except Patent	0.603	0.691	0.736	0.769	0.679	0.717
(iv) Except Hot Topic					0.678	0.709
(v) Except Ecosystem	0.645	0.651	0.723	0.761	0.654	0.717
(vi) Except Interaction Terms	0.654	0.675	0.731	0.749	0.659	0.710

6.2.4. Influential Values That Could Affect the Assessment Model

This section discusses whether there are observations that have significant impact on the model coefficient and specification in the datasets used. Observations of ***leverage***, ***outlier***, and ***influence*** that may have significant impact on model building are in question. Leverage is defined as an observation with covariate pattern that is far away from the regressor space. Outlier is defined as such an observation that its response value is unusually conditional on covariate pattern. Influence is the product of leverage and outlier. Since there could be a significant shift of the coefficient when influential observation is dropped from the model, potential influential values are checked in this section. Summary statistics for leverage, outlier and influence are considered to be hat values, studentized residuals and Cook's distance, which are used for this section. [93, 94, 95, 96]

6.2.4.1. Computation and Diagnosis of Influential Observations

In order to check whether the fit is supported over the entire set of covariate patterns, regression diagnostics is used. In other words, regression diagnostics is to detect influential observations that have significant impact on the model. The following descriptions will focus on single or subgroup of observations and introduce how to perform computation and analysis on leverage, outliers and influence.

Leverage [97]

The predicted responses of the regression model can be obtained by pre-multiplying the $n \times 1$ column vector, $\boldsymbol{y}$, containing the observed responses by the $n \times n$ matrix $\boldsymbol{H}$:

$$\hat{y} = Hy = X(X'X)^{-1}X'y \tag{6-1}$$

Where the regression model can be written succinctly by using the matrix formulation as:

$$Y = X\beta + \epsilon \tag{6-2}$$

The $n \times n$ matrix $\boldsymbol{H}$ is called **the hat matrix.** It is the matrix that puts the hat on the observed response vector $\boldsymbol{y}$ to get the predicted response vector $\hat{y}$, which contains the leverages that help identify extreme x values.

The predicted response can be written as:

$$\hat{y}_i = h_{i1}y_1 + h_{i2}y_2 + \ldots + h_{ii}y_i + \ldots + h_{in}y_n \quad for\ \ i = 1, \ldots, n \tag{6-3}$$

where n = the number of observations.

The **leverage**, h_{ii}, quantifies the influence that the observed response y_i has on its estimated value $\hat{y}_i$. That is, if h_{ii} is small, then the observed response y_i plays only a small role in the value of the predicted response $\hat{y}_i$. On the other hand, if h_{ii} is large, then the observed response y_i plays a large role in the value of the estimated response $\hat{y}_i$. For this reason the h_{ii} are called the ***leverages***.

Observations that are far from the average covariate pattern (or regressor space) are considered to have high leverage. Leverage is expressed as hat value. Hat values of each observation can be obtained using hatvalues() function from *car* package.

Outlier [98]

Studentized residuals are based on the concept of ***deleted residuals***. The basic idea of deleted residuals is to delete the observations one at a time, each time refitting the regression model on the remaining *n*–1 observations. Then, we compare the observed response values to their fitted values based on the models with the ith observation deleted. This produces deleted residuals**.** Standardizing the deleted residuals produces ***studentized residuals***.

If we let:

- y_i denote the observed response for the ith observation, and
- $\hat{y}_{(i)}$ denote the estimated response for the ith observation based on the estimated model with the ith observation deleted

then the ith (unstandardized) deleted residual is defined as:

$$d_i = y_i - \hat{y}_{(i)} \tag{6-4}$$

A studentized residual (sometimes referred to as an "externally studentized residual") is:

$$t_i = \frac{d_i}{\mathrm{s}(d_i)} = \frac{e_i}{\sqrt{MSE_i(1 - h_{ii})}} \tag{6-5}$$

where e_i = the residual.

That is, a studentized residual is just a deleted residual divided by its estimated standard deviation (first formula). This turns out to be equivalent to the ordinary residual divided by a factor that includes the mean square error based on the estimated model with the ith observation deleted, $MSE_{(i)}$, and the leverage, h_{ii} (second formula).

Another formula for studentized residuals allows them to be calculated using only the results for the model fit to all the observations:

$$t_i = r_i \left(\frac{n - k - 2}{n - k - 1 - r_i^2} \right)^{1/2} \tag{6-6}$$

where r_i is the ith standardized residual, *n* = the number of observations, and *k* = the number of explanatory variables.

Outlier is defined as an observation with a response value that is unusually conditional on covariate patterns. If the one with these characteristics survives, it is an outlier. Such an outlier may have significant impact on model fitting, i.e., producing unusual *y* values. Outlier can be formally examined using studentized residuals. Studentized residuals can be calculated using studres() function from *MASS* package.

Influence [99]

Finally, if removal of an observation causes substantial change in the estimates of coefficient, it is called influential observation. Influence can be thought of as the product of leverage and outlier (e.g., it has high hat value and response value is unusual conditional on covariate pattern). Cook's distance is a summary measure of influence. A large value of Cook's distance indicates an influential observation.

Cook's distance measure, denoted D_i, is defined as:

$$D_i = \frac{(y_i - \hat{y}_i)^2}{(k+1) \times MSE}\left[\frac{h_{ii}}{(1-h_{ii})^2}\right] \tag{6-7}$$

Cook's D_i depends on both the residual, e_i (in the first term), and the leverage, h_{ii} (in the second term). That is, both the x value and the y value of the data point play a role in the calculation of Cook's distance.

Cook's distance can be examined by using influencePlot() function provided by *car* package, providing both studentized residuals and hat values as well, as shown in Figure 6-15 and Table 6.3. Herein, all relevant authors across the topics are arranged in ascending order corresponding to their values of startup readiness regarding Participant/Exit per each researcher group, and are numbered accordingly.

In order to decide when a Cook's distance measure is large enough to warrant treating an observation as influential, this thesis complies with the guidelines commonly used as follows.

- If D_i is greater than 0.5, then the ith data point is worthy of further investigation as it **may be influential**.
- If D_i is greater than 1, then the ith data point is **quite likely to be influential**.
- Or, if D_i sticks out like a sore thumb from the other D_i values, it is **almost certainly influential**. Some literature advises the use of graphics and to examine, in closer details, the points with D_i that are substantially larger than the rest, suggesting that thresholds should just be used to enhance graphical displays. [100]

As a result, the authors diagnosed as potentially influential observations are: (i) for **Cas9**, No. 19610, 19600, 19598 and 19103 for Participant (Cook's distances: 0.169, 0.048, 0.023 and 0.021 respectively), and No. 19537, 17473, 19599, 19570, 17712 and 19564 for Exit (Cook's distances: 0.199, 0.053, 0.051, 0.031, 0.031 and 0.028 respectively); (ii) for **Microbiome**, No. 34966, 35621 and 35583 for Participant (Cook's distances: 2.350, 0.099 and 0.075 respectively), and No. 35857, 22244, 31755 and 35847 for Exit (Cook's distances: 0.070, 0.066, 0.041 and 0.039 respectively); and (iii) for **5-Biopharma-Topics**, No. 94576, 94487, 94422 and 94333 for Participant (Cook's distances: 0.059, 0.042, 0.035 and 0.031 respectively), and No. 94640 (Cook's distance:

1.285) for Exit, as shown in Figure 6-16. Values of their Cook's distance with their corresponding values of studentized residuals and hat values are shown in Table 6.3.

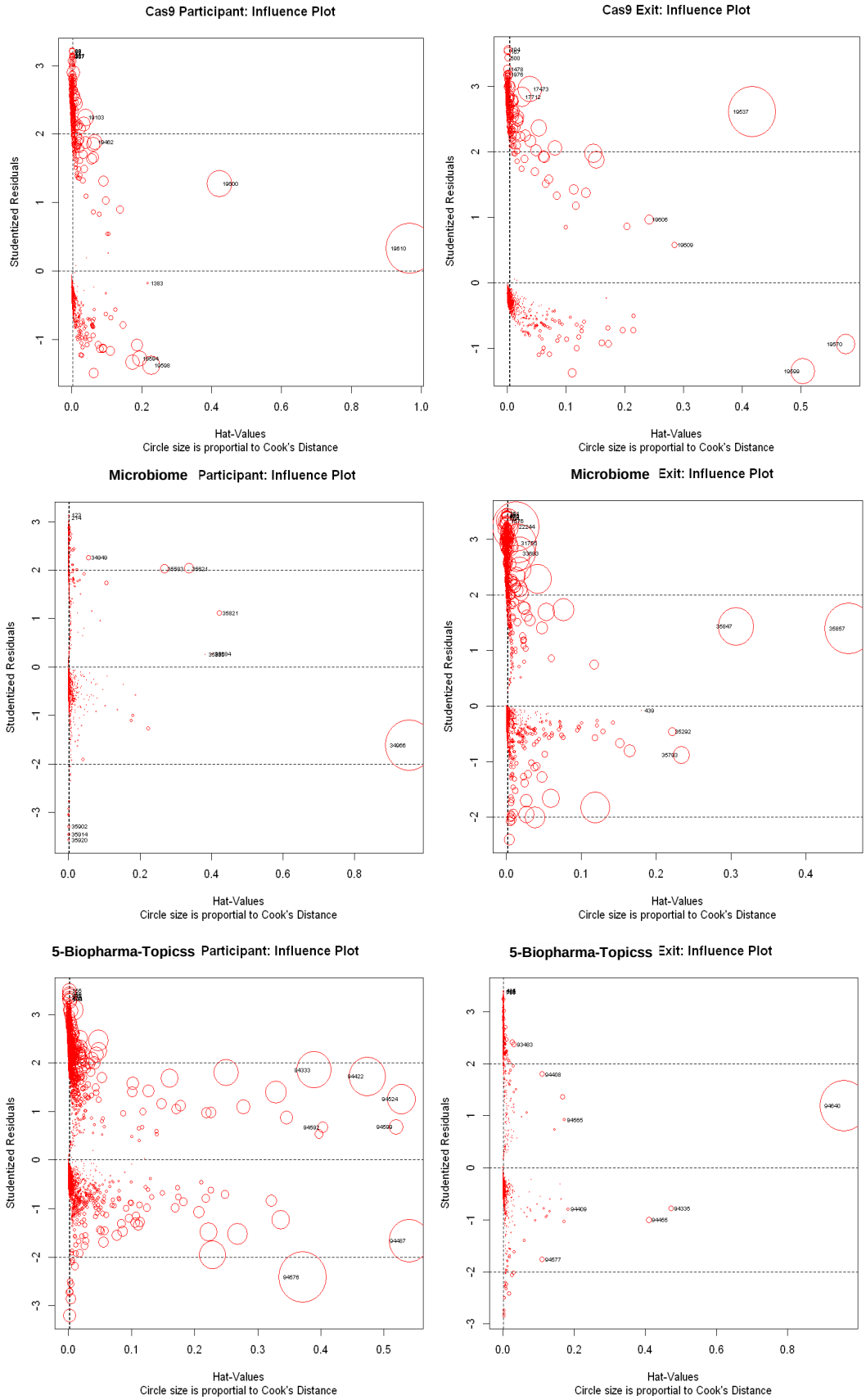


Figure 6-15 Influence Plots per Researcher

Table 6.3 Author No. with Top 20 CookD and Their StudRes & Hat Values per Researcher Group

Cas9 Participant

Rank	Author No.	CookD	StudRes	Hat
1	19610	0.169	0.332	0.968
2	19600	0.048	1.273	0.423
3	19598	0.023	-1.377	0.227
4	19103	0.021	2.242	0.039
5	19462	0.016	1.881	0.067
6	19594	0.016	-1.280	0.195
7	19597	0.015	-1.329	0.174
8	19230	0.014	2.143	0.033
9	19473	0.014	1.847	0.062
10	1667	0.013	2.901	0.004
11	18111	0.012	2.438	0.013
12	17492	0.011	2.495	0.011
13	19450	0.010	1.873	0.040
14	19581	0.010	-1.080	0.186
15	19544	0.010	1.659	0.061
16	19545	0.010	1.658	0.061
17	19548	0.010	1.656	0.061
18	16383	0.009	2.572	0.007
19	16459	0.009	2.568	0.007
20	16461	0.009	2.568	0.007

Cas9 Exit

Rank	Author No.	CookD	StudRes	Hat
1	19537	0.199	2.617	0.417
2	17473	0.053	2.959	0.038
3	19599	0.051	-1.353	0.504
4	19570	0.031	-0.940	0.577
5	17712	0.031	2.838	0.025
6	19564	0.028	1.973	0.147
7	19575	0.024	1.882	0.151
8	19369	0.023	2.369	0.053
9	19520	0.018	2.061	0.081
10	18968	0.013	2.534	0.018
11	17397	0.012	2.769	0.009
12	17230	0.012	2.782	0.009
13	17023	0.012	2.795	0.008
14	19464	0.012	2.170	0.040
15	19050	0.012	2.498	0.018
16	9042	0.011	2.991	0.004
17	19094	0.011	2.477	0.018
18	19541	0.011	1.938	0.061
19	19546	0.011	1.927	0.062
20	19410	0.011	2.254	0.030

Microbiome Participant

Rank	Author No.	CookD	StudRes	Hat
1	34966	2.350	-1.612	0.954
2	35621	0.099	2.030	0.337
3	35583	0.075	2.027	0.269
4	35821	0.030	1.112	0.422
5	34949	0.025	2.260	0.056
6	35625	0.017	1.732	0.106
7	35784	0.016	-1.261	0.222
8	21161	0.012	2.745	0.007
9	34192	0.012	2.433	0.015
10	35414	0.010	1.922	0.044
11	35852	0.009	-1.914	0.041
12	34232	0.009	2.411	0.013
13	35747	0.008	-1.102	0.175
14	35914	0.008	-3.453	0.000
15	28905	0.007	2.667	0.005
16	35902	0.007	-3.293	0.001
17	35238	0.007	2.010	0.024
18	2984	0.007	2.895	0.002
19	35719	0.007	-0.991	0.180
20	14457	0.006	2.811	0.003

Microbiome Exit

Rank	Author No.	CookD	StudRes	Hat
1	35857	0.070	1.397	0.457
2	22244	0.066	3.224	0.013
3	31755	0.041	2.927	0.015
4	35847	0.039	1.431	0.306
5	33693	0.030	2.742	0.018
6	35886	0.026	-1.819	0.118
7	35331	0.024	2.278	0.042
8	34122	0.020	2.649	0.014
9	1576	0.017	3.330	0.001
10	34815	0.016	2.465	0.017
11	35776	0.014	1.733	0.076
12	9498	0.013	3.202	0.002
13	35899	0.012	-2.002	0.038
14	1428	0.011	3.319	0.001
15	31224	0.010	2.854	0.004
16	30580	0.010	2.898	0.003
17	35103	0.009	2.331	0.013
18	35879	0.009	-1.661	0.059
19	35780	0.009	1.686	0.053
20	27975	0.008	2.999	0.002

5-Biopharma-Topics Participant

Rank	Author No.	CookD	StudRes	Hat
1	94576	0.059	-2.410	0.371
2	94487	0.042	-1.665	0.541
3	94422	0.035	1.722	0.474
4	94333	0.031	1.858	0.389
5	94524	0.021	1.257	0.528
6	94561	0.018	-1.952	0.229
7	94261	0.016	1.808	0.250
8	94475	0.012	1.399	0.329
9	94512	0.011	-1.522	0.268
10	90223	0.011	2.456	0.048
11	4460	0.010	3.085	0.008
12	94427	0.009	-1.227	0.336
13	94513	0.008	-1.483	0.222
14	94281	0.008	1.685	0.160
15	92237	0.007	2.247	0.048
16	544	0.006	3.316	0.002
17	92749	0.006	2.169	0.044
18	2773	0.006	3.080	0.004
19	415	0.005	3.384	0.001
20	356	0.005	3.413	0.001

5-Biopharma-Topics Exit

Rank	Author No.	CookD	StudRes	Hat
1	94640	1.285	1.191	0.959
2	94455	0.020	-1.001	0.409
3	94408	0.017	1.797	0.110
4	93483	0.016	2.367	0.031
5	94577	0.016	-1.759	0.110
6	93270	0.016	2.411	0.027
7	94335	0.015	-0.782	0.472
8	94522	0.012	1.370	0.167
9	94634	0.011	-2.415	0.017
10	88904	0.009	2.705	0.006
11	39047	0.009	3.026	0.002
12	5876	0.008	3.244	0.001
13	92408	0.007	2.465	0.010
14	94605	0.007	-2.011	0.029
15	94610	0.007	-2.060	0.025
16	94287	0.007	1.953	0.030
17	94500	0.006	-1.029	0.171
18	91292	0.006	2.544	0.006
19	89359	0.005	2.667	0.004
20	94211	0.005	2.016	0.019

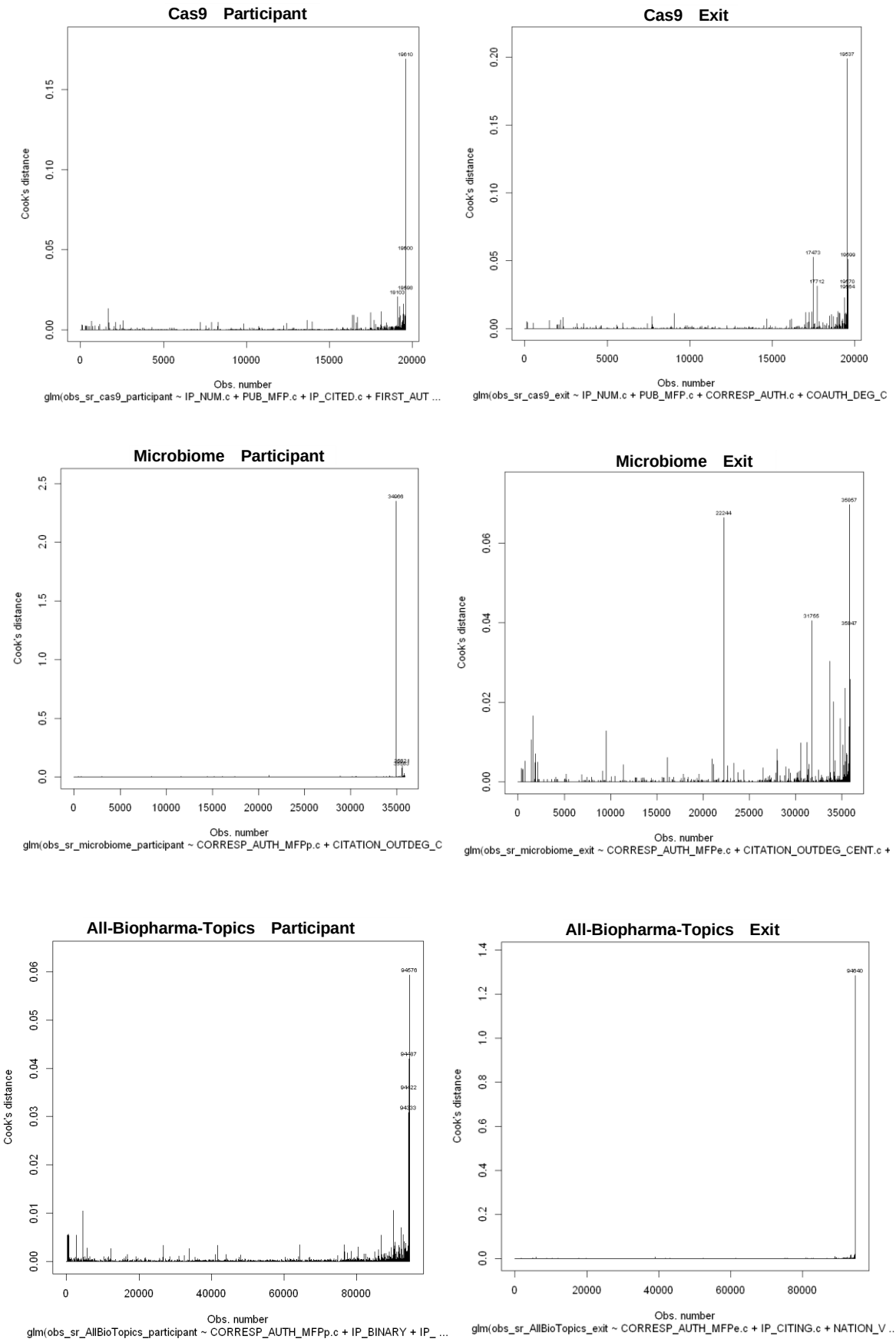


Figure 6-16 Diagnostic Plots to Identify Influential Plots Based on Cook's Distance

6.2.4.2. Distribution of Academic Researchers' Over-Cutoff Cook's Distances and Their Associated Startup Readiness (Probabilities) & Variables

Prior to delving deeper into the extent to which the aforementioned potentially influential observations influence outcomes, this section analyzes distribution of Cook's distances over a recently widely accepted threshold value (i.e., $4 / (n - k - 1)$ [101]) to flag them as possibly being influential, according to relevant researcher groups in question. Moreover, distributions of values of startup readiness (estimated probabilities regarding Participant/Exit by the assessment model of this thesis) and those of solo explanatory variables' difference values from their means, both of which are related to academic researchers with Cook's distances over the threshold, are examined together.

As a result, as seen in Figure 6-17, the following observations are concluded among all academic researchers with Cook's distances over the above threshold, across all research topics.

- Characteristic differences in means and SD's of Cook's distances between positive groups and negative groups are not consistently observed.
- Regardless of Participant/Exit, positive researcher groups herein have greatly larger numbers of observations than negative groups: (i) for **Cas9**, 652 to 88 for Participant and 343 to 79 for Exit; (ii) for **Microbiome**, 1087 to 198 for Participant and 527 to 134 for Exit, and (iii) for **5-Biopharma-Topics**, 2884 to 364 for Participant and 1515 to 240 for Exit. This is a contrasting finding in comparison to the overall descriptive statistics (Appendix C-1 – C-3) in which observations in positive groups are obviously way fewer than in negative groups.
- Regardless of Participant/Exit, negative researcher groups have greatly larger means of startup readiness than positive groups: (i) for **Cas9**, 0.29 to 0.07 for Participant and 0.21 to 0.05 for Exit; (ii) for **Microbiome**, 0.36 to 0.08 for Participant and 0.28 to 0.09 for Exit, and (iii) for **5-Biopharma-Topics**, 0.34 to 0.08 for Participant and 0.32 to 0.06 for Exit. This is a striking finding since in the overall academic researchers' negative groups obviously should have smaller means of startup readiness than positive groups.
- Regardless of Participant/Exit, across all solo explanatory variables, absolute difference values of their means in negative groups are evidently larger than those in positive groups consistently.

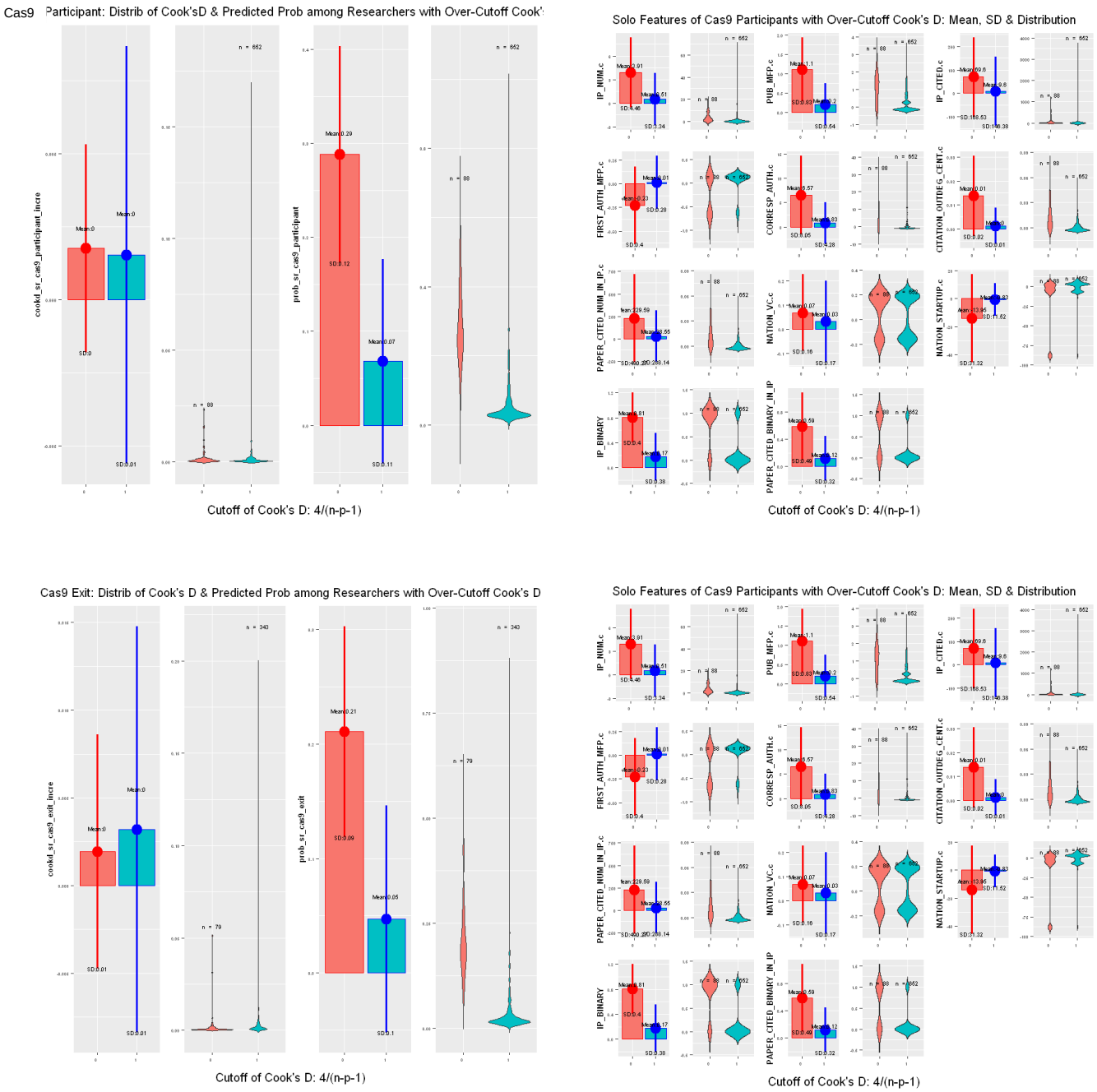
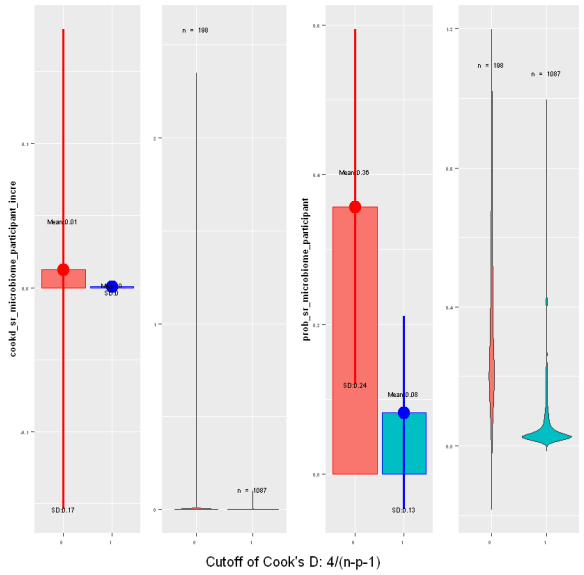


Figure 6-17 Distribution of Researchers' Over-Cutoff CookD and Their SR (Probabilities) and Variables

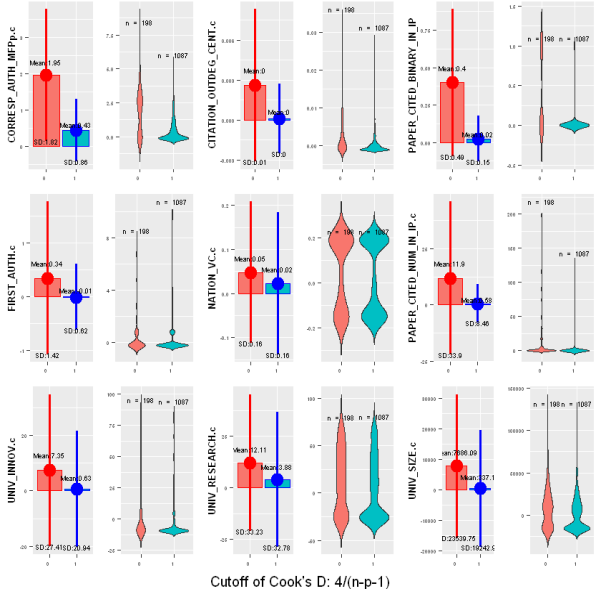
(…CONTINUED ON NEXT PAGE)

(CONTINUED FROM PREVIOUS PAGE & CONTINUED ON NEXT PAGE)

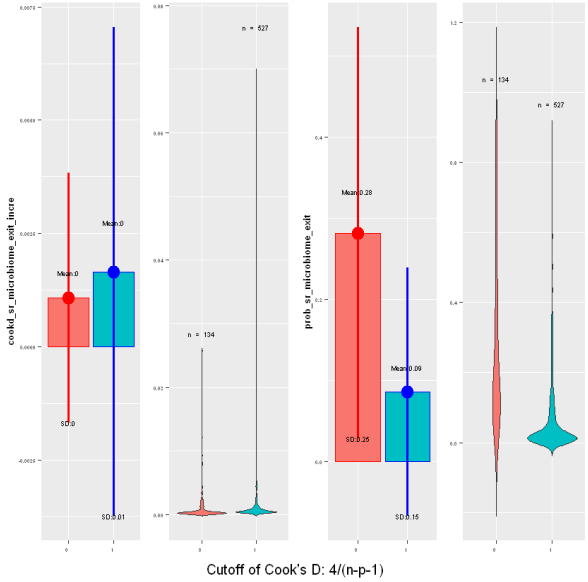
Microbiome Participant: Distrib of CookD & Prob. among Researchers with Over-Cutoff CookD



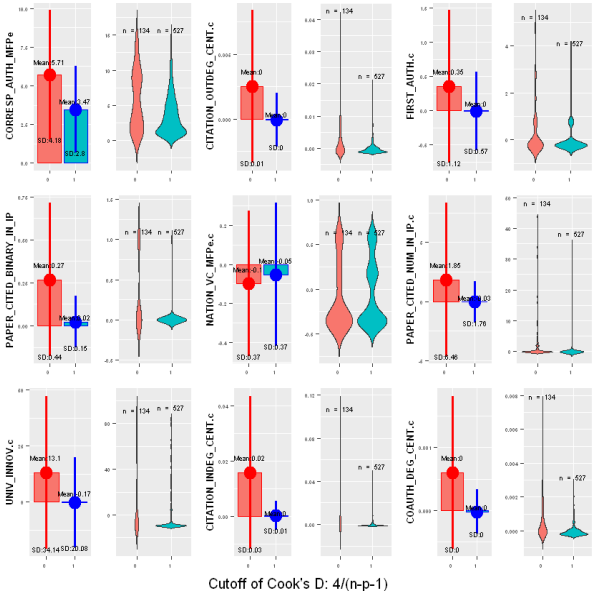
Solo Features of Microbiome Participants with Over-Cutoff Cook's D: Mean, SD & Distribution



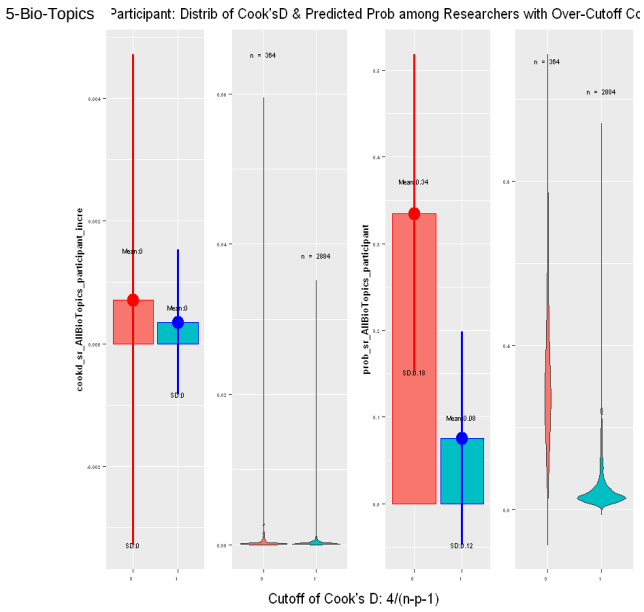
Microbiome Exit: Distrib of CookD & Exit Prob. among Researchers with Over-Cutoff CookD

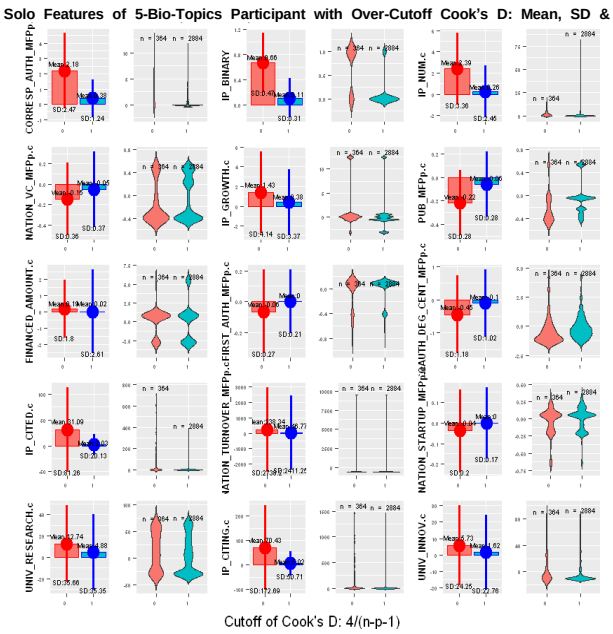


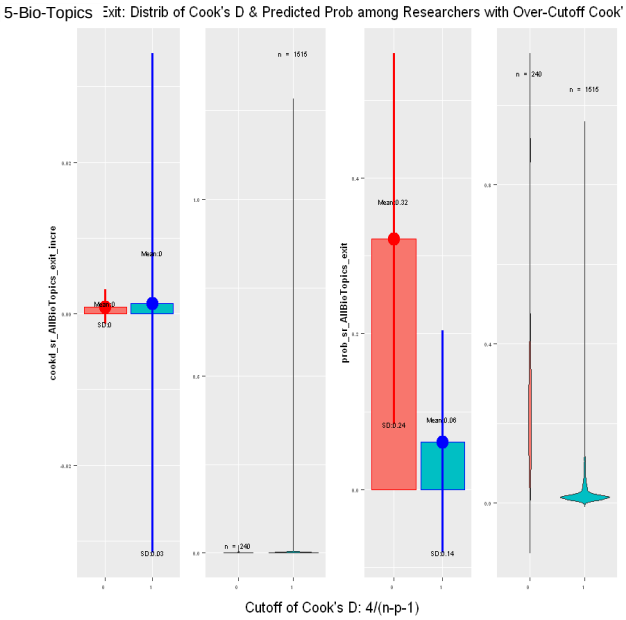
Solo Features of Microbiome Exit with Over-Cutoff Cook's D: Mean, SD & Distribution

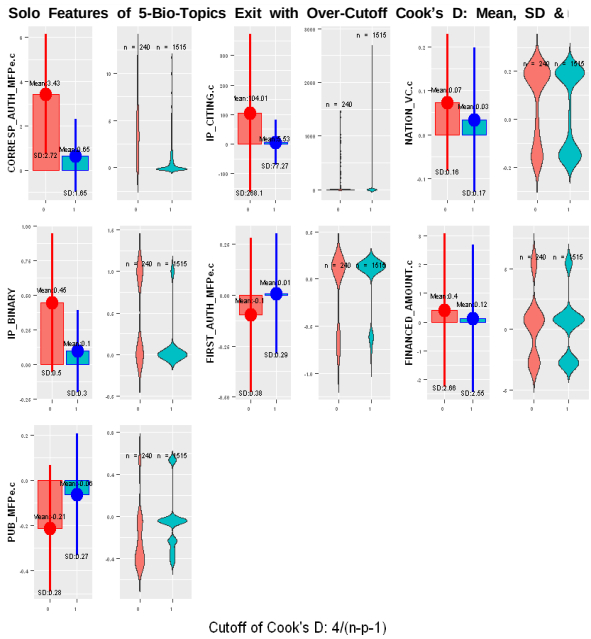


(CONTINUED FROM PREVIOUS PAGE)









6.2.4.3. Examination of Coefficients Affected by Potential Influential Observations and Relevant Variables' Statistical Significance

As described in the introduction of 6.2.4, influential observations are such observations without which there will be a significant impact on the outcome of the model, caused by a significant shift of (a) coefficient(s).

Thus, this section examines change of coefficients by removing potentially influential observations diagnosed in 6.2.4.1 (See Figure 6-16), and whether the effect of

relevant variables are statistically significant on the outcome of the assessment model (Table 5.17, Table 5.18, Table 5.19, Table 5.20, Table 5.22 and Table 5.23), per researchers goups.

(a) ***Cas9***

For Participant, we can see from the output in Appendix E-1 that coefficients are changed minimally, and thus the authors 19610, 19600, 19598 and 19103 are, even when combined, not influential.

On the other hand, for Exit, it is observed in Appendix E-1 that removal of the authors 19537, 17473, 19599, 19570, 17712 and 19564 leads to ***significant change*** in the coefficients of COAUTH_DEG_CENT, IP_CITED and NATION_VC * IP_CITED, as highlighted in yellow. Throughout this section, ***significant change*** refers to either an increase more than double (200%) or a decrease less than a half (50%) (including negative values). From the output in Table 5.18, however, since p_values associated with COAUTH_DEG_CENT (No. 18), IP_CITED (No. 8) and NATION_VC * IP_CITED (No. 22) are 0.929, 0.200 and 0.101 respectively, the effect of those explanatory variables are found not statistically significant at the 5% level, it is not concluded that these observations are influential. Thus, it is concluded that retaining these authors in the model is not necessarily inappropriate.

(b) ***Microbiome***

For Participant, it is observable in Appendix E-2 that removal of the authors 34966, 35621 and 35583 causes significant change in the coefficients of FIRST_AUTH * IP_NUM, IP_NUM * IP_CITED_MFPp and PAPER_CITED_NUM_IN_IP * UNIV_SIZE, as highlighted in yellow. From the output in Table 5.19, however, since p_values associated with FIRST_AUTH * IP_NUM (No. 10), IP_NUM * IP_CITED_MFPp (No. 6) and PAPER_CITED_NUM_IN_IP * UNIV_SIZE (No. 7) are 0.277, 0.073 and 0.057 respectively, the effect of those explanatory variables are found not statistically significant at the 5% level, it is not concluded that these observations are influential. Thus, it can be concluded that retaining these authors in the model is not necessarily inappropriate.

On the other hand, for Exit, Appendix E-2 shows that removal of the authors 35857, 22244, 31755 and 35847 causes coefficients to change just minimally, and thus these authors are not influential even when combined.

(c) ***5-Biopharma-Topics***

For Participant, we can see in Appendix E-3 that removal of the authors 94576, 94487, 94422, and 94333 triggers significant change in the coefficients of CORRESP_AUTH_MFPp * PAPER_CITED_NUM_IN_IP and UNIV_INNOV *

PAPER_CITED_NUM_IN_IP, as highlighted in yellow. Table 5.22, however, shows that p_values associated with CORRESP_AUTH_MFPp * PAPER_CITED_NUM_IN_IP (No. 63) and UNIV_INNOV * PAPER_CITED_NUM_IN_IP (No. 45) are 0.250 and 0.069 respectively, not statistically significant at the 5% level. Thus, it is not concluded that these observations are influential and that retaining these authors in the model is inappropriate.

On the other hand, for Exit, Appendix E-3 shows that removal of the author 94640 leads to just a minimal change in coefficients, thus the author is not influential.

6.3. Expert Interview

This section introduces interviews that were conducted with three experts regarding the research questions (See 1.3) and findings of this thesis, composed of a university-administrator-cum-entrepreneurship-researcher Professor Shigeo Kagami, an entrepreneur-cum-patent-attorney Dr. Yoshihito Daimon, and a venture capitalist Dr. Atsushi Usami, all of whom have been closely involved with academic startups in the biopharmaceutical domain. They have been collaborators with the author at The University of Tokyo Edge Capital Co., Ltd. (UTEC) to invest in and nurture academic startups in fields such as the biopharmaceutical domain.

(i) Professor Shigeo Kagami, The University of Tokyo

As the first interviewee to set the tone for my writing on academic entrepreneurship, I interviewed Professor Shigeo Kagami, General Manager of the Office of Innovation and Entrepreneurship, at the Division of University Corporate Relations (DUCR) in The University of Tokyo. At DUCR, he has led initiatives to foster an environment of entrepreneurship since 2004, building startup ecosystem nurtured through activities such as the EDGE program and the Entrepreneurship Dojo. One of the most successful startups incubated through his initiatives with UTEC in a facility called The University of Tokyo Entrepreneur Plaza run by his office, is PeptiDream Inc. PeptiDream is a biopharmaceutical startup listed in the First Section of the Tokyo Stock Exchange since December 2015, which employs its proprietary Peptide Discovery Platform System for the discovery and development of constrained peptides, small molecule, and peptide-drug conjugate therapeutics.

Kagami started the interview with positive feedback regarding Paper-related Features used in this thesis, saying that it is especially unique to treat such paper-related features related to first authorship and authors' citation/co-authorship centralities as effective factors for startup readiness in the biopharmaceutical domain, referring to research achievement on technology entrepreneurship by Professor Scott Andrew Shane

at Case Western Reserve University, who has published ten books such as “Academic Entrepreneurship: University Spinoffs and Wealth Creation” and over 60 scholarly articles. Professor Kagami completed his doctoral work at this university. Shane provides analysis of the four major factors that jointly influence spinoff activity: the university and societal environment, the technology developed at universities, the industries in which spinoffs operate, and the people involved in his aforementioned book [102].

Professor Kagami opined that this dissertation’s approach to illustrate the importance of academic researchers’ scientific activeness and prominence matter for startup readiness deserves to be evaluated as a novel methodology, because of its attention to several paper-related factors related to scientific activeness and prominence and its in-depth data-driven factor analysis, both of which he believed had been unseen for this purpose. According to Professor Kagami, these potential scientific factors have not been discussed so profoundly in the context of academic entrepreneurship. “This approach could be one answer to explain why some biotechnology startups with many ‘strong’ patents to commercialize do not perform well. In order to be successful as biopharmaceutical startups with high scientific linkage as well as R&D funding need for a longer time horizon until commercialization, it is an eye-opening and convincing finding that the aggressiveness and the quality of scientific research by researchers are critical in the first place, rather than enhancing the evaluation of patents to a disproportionate extent,” said Kagami.

On the other hand, he advised that I conduct further research on ecosystem factors, to elucidate how venture capital firms and technology license offices that can evaluate the potential of science work, in order for this dissertation’s assessment model to effectively function. Another aspect that he advised me to research on was team building, saying that the recent literature on academic entrepreneurship pays much attention to how academic researchers build teams to commercialize their research output effectively.

(ii) Dr. Yoshihito Daimon, Co-Founder and CEO of bitBiome

bitBiome, Inc. is a biopharmaceutical startup co-founded in November 2018 by researchers, entrepreneurs and UTEC related to **Microbiome** that provides microbiome analysis using its proprietary Single-Cell Genomics Technology, which enables us to precisely obtain whole-genome sequence data from just a single microbial for a wide range of microbial species. Their technology allows us to develop potential diagnostics and biological technologies based on newly-found functions of microorganisms. The second interviewee Dr. Yoshihito Daimon researched microbial engineering and genetic resources engineering at Kyushu University’s Graduate School of Bioresource & Bioenvironmental Sciences, and engaged in intellectual property (IP) and legal at Astellas Pharma Inc. until he joined bitBiome.

“The research finding of this dissertation that I found the most interesting is the pivotal role of Paper-related Features for startup readiness, despite a piece of earlier research contradictorily concluding that publications’ assets have little impact on researchers’ spinoff creation.” Daimon said. He alluded that this dissertation’s approach could pave the way for more enriched analysis of highly scientific academic researchers’ startup potential based on such in-depth data of papers. According to him, another interesting finding in particular was that Interaction Terms Factors, or combinations of factors, can effectively influence startup readiness, than just each component of the Interaction Terms Factors could possibly do separately.

While at the same time, as a patent-attorney who had practiced in the pharmaceutical field for years both from mega pharma and startup perspectives, Daimon provided several insightful pieces of advice regarding Patent-related Features for caveats and improvements, as follows.

Firstly, there is time lag between the invention and its surfacing to the public, typically due to the so-called 18-month publication rule. Thus, consideration of this limitation is needed to take account of Patent-related Features, whose time lag could affect the result of computation of startup readiness especially when Patent-related Features are in the early phase of emerging. In the 18-month publication rule, patent applications are confidential to the patent office for generally 18 months after the earliest priority date of the application until the patent office’s publication of it. Daimon pointed out one significant milestone example regarding the **CRISPR/Cas9** field that UC Berkeley Professor Jennifer Doudna achieved in both paper publication and patent application: Her milestone paper coauthored with Emmanuelle Charpentier entitled “A Programmable Dual-RNA–Guided DNA Endonuclease in Adaptive Bacterial Immunity” appeared in *Science* online on June 28, 2012 [103, 104], while her invented patent that pairs up with the paper, with the title of “METHODS AND COMPOSITIONS FOR RNA-DIRECTED TARGET DNA MODIFICATION AND FOR RNA-DIRECTED MODULATION OF TRANSCRIPTION” (Publication No. WO/2013/176772) was published on November 28, 2013. Although it is assumed that Doudna submitted her paper to *Science* shortly after she applied to the U.S. Patent and Trademark Office, her patent publication took one year and five months after her paper appeared online.

Secondly, important patents tend to have the following characteristics compared to others: larger number of citations by patent examiners in patent examination procedures for relevant patent applications, larger number of countries where patents are applied for, and quicker measures to secure valid patents on application. Therefore, implementation of data regarding these characteristics could improve the importance and effectiveness of Patent-related Features in the model of this dissertation.

Finally, he suggested that frequency of first inventor be counted for researchers too, as one of Patent-related Features, since inventors listed first in patent applications often play a leading role in their process of conversion to patent rights, as first authors in research papers do in their paper publication process.

(iii) Dr. Atsushi Usami, Partner of UTEC

UTEC is a venture capital firm that nurtures and invests in academic startups from their seed/early stages that utilize science and technology emanating from universities and research institutes globally, including the University of Tokyo, managing JPY 54.3 billion (approximately USD 500 million) so far. The author has been involved with UTEC since its founding back in April, 2004. Atsushi Usami received his Ph.D. in pharmacology and neuroscience from the University of Tokyo and since October, 2013 led several investments in biopharmaceutical startups in the fields such as **Microbiome** and genome editing neighboring **CRISPR/Cas9** at UTEC.

"I found exceptionally interesting that, overall, Paper-related Features that indicates researchers' creativeness and prominence such as first authorship, work more effectively in Exit than in Participant. The reason is presumably that Exit is an event that needs objective evaluation by third parties, which questions researchers' fundamental credentials as scientists with research output. On the other hand, interestingly for Participant, Ecosystem Factors and Hot Topic Factors that are related to environment surrounding researchers are more notable as factors for startup readiness, as determining factors are more subjective than for Exit," said Usami.

According to Usami who has research background in life sciences including the biopharmaceutical domain, researchers prior to becoming established scientists, who are relatively young researchers in most cases, are expected to assume the role of first authorship in research planning and data collection, whereas established authors are expected to be more of corresponding authors than first authors in general, although some senior researchers do both. After expressing his curiosity in the time lag between researchers' first publication and Participation/Exit of startups, Usami argued, "From the findings of this thesis, it is suggested that, in certain biopharmaceutical research fields, the frequency of an author becoming a first author [FIRST_AUTH] could contribute more in Exit than in Participant, when estimating startup readiness. Thus, I assume that there is a possibility that relatively young researchers have more potential to contribute to Exit, than do researchers with higher positions. As this dissertation suggests, if it is understood that researchers prior to senior positions tend to contribute more effectively to Exit of startups, the finding could make an impact on governmental science and technology policy, which recently has had a trend in which researchers other than already established ones have difficulty in getting funded."

Another aspect we discussed in the interview was potential features about the timing of Exit. Usami pointed out that Exit tends to occur frequently when either excitement/expectation or popularity/penetration of the relevant science and technology is very high. In other words, the period in between is considered to be tough for Exit, which led to his suggestion that relationship between Hot Topic Factors and Exit could be explored to develop features. According to Usami, creation of Hot Topic Factors regarding papers such as growth of Paper-related Features is worth considering, in order to address the timing issue.

6.4. Influence of Exit on Paper- and Patent-Related Features of Academic Researchers

Although available data is limited, influence of Exit on academic researchers' Essential Individual Factors (Paper-related Features and Patent-related Features, See 2.1) is surveyed in this section.

As described in 4.1.6, explanatory variables including Paper- and Patent-related Features are collected during the common periods per each variable across all relevant authors, irrespective of whether and when the authors experience an event (Participant/Exit). This is for the purpose of measuring each feature consistently: in case of Paper- and Patent-related Features, measuring features signaling authors' individual scientific prominence and innovation capability with consistency. The observation periods of these features are 2013-2017 in a uniform way regardless of researcher groups. In explanatory modeling that was conducted in this thesis, as opposed to predictive modeling, it is considered to be preferable that we analyze all subjects in a consistent manner irrespective of subjects' actual state regarding an event (See Chapter 5's introductory description).

In the VentureSource database, it is practically possible to extract the timing of Exit, and to recount Paper-related and Patent-related Features from 2013 until the year before Exit, as long as the Exit occurred in and after 2014 by 2017, which enabled this appendant survey (See 4.3). The database lets us neither extract the timing of Participant, nor discern which counts of explanatory variables precede an event such as Exit on a monthly or daily basis, though.

To put it another way, it is possible to observe differences between the average counts of the authors' Paper- and Patent-related Features prior to the Exit year and those in and after the year, which enables us to understand the influence of Exit on academic researchers' Paper- and Patent-related descriptive statistics beyond the year, albeit for a limited period of time. This section conducts the analysis among all the 94669 authors in 5-Biopharma-Topics and it turns out that, only 466 authors are extracted from the 1556

authors who have experienced Exit. This partiality is due to the fact that most of the relevant authors experienced Exit in either before 2013 or after 2017, which makes it impossible, given the limited observation years 2013-2017 herein, to calculate the change of the average counts from the pre-exit year(s) to the following year(s).

Figure 6-18 demonstrates the change of annual average counts or the analysis per Paper- and Patent-related Features beyond the Exit year, with respect to their means, SDs and distributions. Findings include:

For Paper-related Features

- No significant change in the mean relative to the SD is observed in PUB and FIRST_AUTH. This means that the Exit event does not influence the productivity of relevant authors' publications and first authorship much, suggesting that the event does not affect relevant authors' individual scientific prominence that much.
- Striking increase in the mean relative to the SD is observed in PAPER_CITED and PAPER_CITING. This shows that the Exit event enhances the profile of relevant authors as well as their attention to other authors' research output, presumably because the Exit event attracts others' interest in relevant authors' papers and also forces the authors to pay more attention to relevant research outcomes.
- Considerable decrease in the mean relative to the SD is observed in CORRESP_AUTH. This indicates that the Exit event suppresses relevant authors' corresponding authorship for papers he/she co-authors, suggesting that the Exit event makes authors less keen on responsibility for their manuscript during the paper submission.

For Patent-related Features

- Compared to Paper-related Features, significant change in the mean relative to the SD is not found across all Patent-related Features, while moderate increase is observed in IP_NUM. This suggests that, the Exit event does not influence the profile of patents and the attention to other authors' patents, while moderately improving the productivity of patent publications.

In summary, Exit does not influence authors' academic activities and their intellectual property creation activities in a discontinuous manner, since major changes in their academic and IP-related productivity and prominence are not observed, albeit some decrease in their responsibility for their co-authored papers. In fact, striking increase of citations to and from relevant authors can be expected to occur because of the Exit event, which fosters recognition and development of the relevant research community.

Based on the author's experience, such positive spiral between academic advancement and entrepreneurial success is being seen not only in the biopharmaceutical domain but also on a larger scale, which could be part of the reason that academic startups are increasingly attracting attention globally among various stakeholders in academia, business and government (See Chapter 1's preceding sentences).

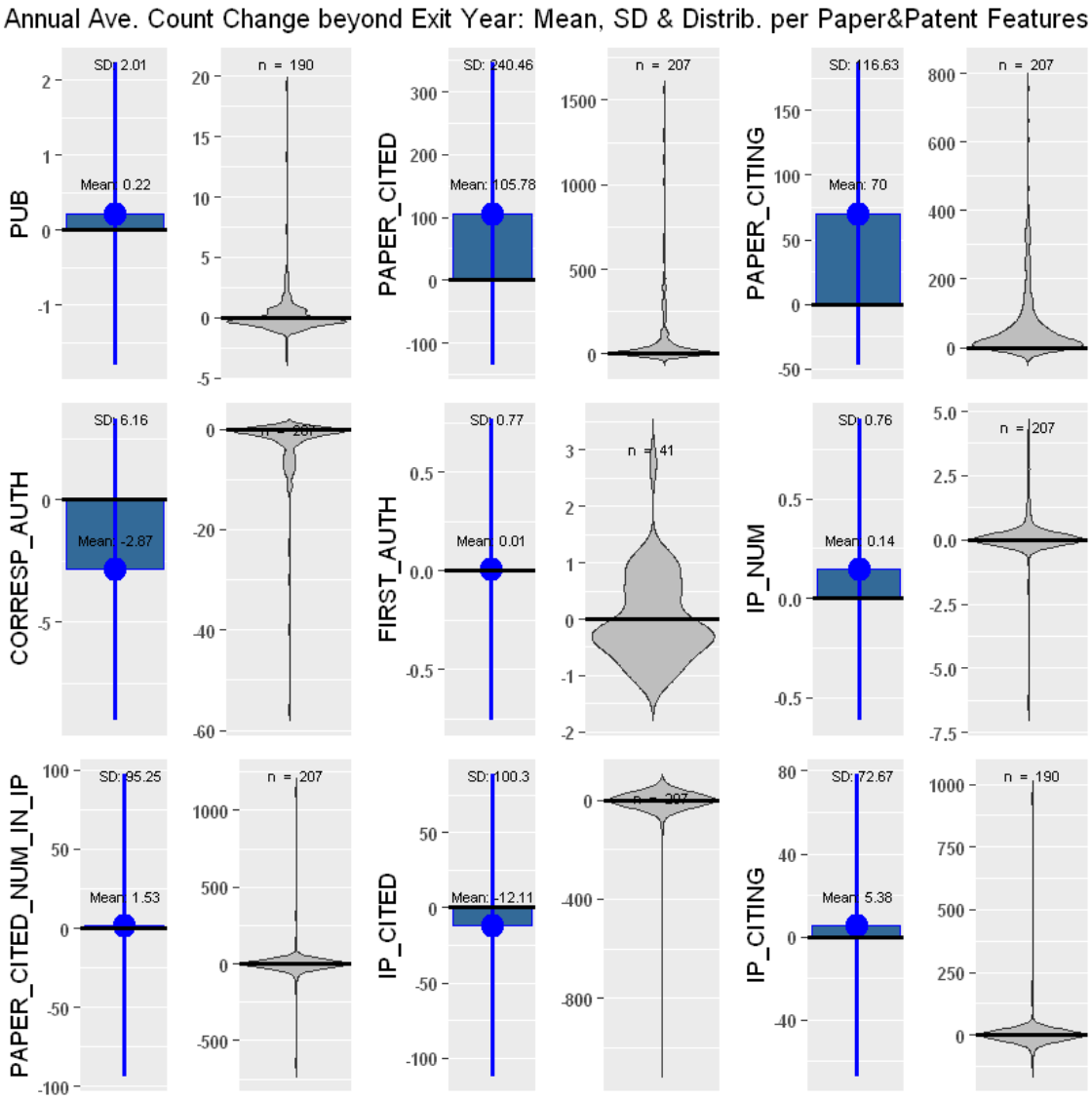


Figure 6-18 Annual Average Count Change beyond Exit Year: Mean, SD & Distrib. per Paper & Patent Features

Chapter 7. Conclusions and Perspectives

7.1. Summary of Findings and Research Questions Revisited

This dissertation shows that, the conceptual framework that is proposed using Essential Individual Factors (composed of Paper-related Features and Patent-related Features), Hot Topic Factors/Features, Ecosystem Factors (composed of Academic Organization-related Features and Nation-related Features) and their Interaction Terms Factors, all of which are built from the selected data sources, is validated as a testable, practical model to assess startup readiness of academic researchers in the biopharmaceutical domain, in terms of startup participation and exit. It also shows that this model is useful to identify promising scientists with high startup readiness and to assess the key factors/features that are important as explanatory variables of startup readiness. The implication is that this model possibly helps enable promising formulation and development of academic startups in the biopharmaceutical domain, such that researchers focus on enhancing their Essential Individual Factors, while business stakeholders exercise their expertise such as financing, management and business, in a mutually complementary fashion.

By implementing the assessment model, it is found that (i) through the construction process of explanatory variables all-embracing groups of factors/features are selected to build the assessment model, while several features are discarded in each group, (ii) Individual Factors composed of Paper-related Features and Patent-related Features are found to be remarkably different between Participant/Exit researchers and non-Participant/non-Exit researchers, and (iii) encompassing groups of factors/features are validated, in that a whole set of feature sets, whether Individual Factors or other factors such as Hot Topic Factors and Ecosystem Factors that relevant researchers are associated with, are found to achieve the best classifying performance for academic researchers in all instances. Ultimately, we found that (iv) the assessment model shows higher performance when we assess startup readiness of Exit, relative to that of Participant, inferably due to the availability of more complete, standardized, and consistent data for Exit, (v) the model shows the best performance to assess startup readiness of researchers regarding **Microbiome**, followed by those of **5-Biopharma-Topics** and **Cas9** in this order, presumably thanks to **Microbiome**'s current status as a more genuine scientific concept, and lesser advancement in practical application despite high keyword growth, and, 5-Biopharma-Topics' wider range of academic researchers with richer features including Hot Topic Factors and Interaction Terms Factors, and (vi) Paper-related Features are found to be the most important

determinants of startup readiness; unique features include those regarding first authorship [FIRST_AUTH] (as pointed out by the interviews of two experts) and corresponding authorship [CORRESP_AUTH], which are found to be substantially important to academic researchers' startup readiness in several cases addressed herein. This implication should be beneficial for policymakers and university administrators to effectively improve relevant features to foster startup readiness.

Designing, implementing, evaluating and interpreting the framework and the assessment model in the aforementioned manner have led us to a better understanding of how to address the proposed research questions of this thesis, which can be revisited as follows.

Primary RQ: What are the implications of this empirical research using the logistic regression model to assess academic researchers' startup readiness based on the variables derived and constructed from the relevant digital data sources, related to the growing topics of interest in the biopharmaceutical domain?

Secondary RQ1: What are the potentially essential factors/features that can be derived from relevant digital data sources, to assess startup readiness of academic researchers who have intense scientific linkage such as those in the biopharmaceutical domain?

Secondary RQ2: What are the appropriate methodologies to be deployed, in order to construct a logistic regression model to assess academic researchers' startup readiness, with respect to preprocessing data, selecting and constructing variables, and, building and implementing the model?

This assessment method can be implemented even by stakeholders with little or no biopharmaceutical domain expertise, since this method is structured in a fashion that does not need such expertise, using digital data available on a real-time basis for anyone. In this way, this model will allow a wide range of stakeholders to benefit from its capabilities of startup readiness assessment and important variable identification, at an earlier stage, in a timelier manner, on a larger scale, and in greater detail, than conventional means. It will enable business professionals like venture capitalists and managerial entrepreneurs to retrieve and evaluate potential scientific founders to work with, from publicly available data sources as discussed before. It will also allow policymakers and university administrators to come up with effective policies while avoiding counterproductive ones, as this method can identify the effects of variables. For

academic researchers themselves, this method makes it possible to detect the variables they could work on to improve their startup readiness.

7.2. Limitations and Future Work

Regarding future research directions of my research, strengthening the legitimacy of my methods requires further research as explained below. Firstly, selection of industry segments and research topics demand further study because they can influence the results. Practicality of my proposed methodologies should be tried out and explored in a more real-world setting further. Secondly, additional study is necessary to address the prospect of evaluating research institutions that are not a university, which do not provide Academic Organization-related Features. Thirdly, additional development of data sources and explanatory variables, as well as their analyses/coordination could also be beneficial to build better variables, which could assess startup readiness more precisely. For example, as experts suggested, Patent-related Features can be developed further to consider the time lag between invention and publication, and the characteristics of important patents such as patent examiner citations, countries in which patents are applied for, and swiftness of securing patents can be surveyed. Hot Topic Factors can also be explored to reflect timings regarding the degree of excitement and penetration, covering more paper-related aspects. In parallel, target variables other than Participant and Exit, such as fundability or Time-to-Exit, could be explored. Fourthly, the development of an efficient method to consolidate researcher identities such that we can verify them across different data sources (e.g., databases regarding papers, patents, and startups) is anticipated, which can substantially reduce or eliminate the need for human discretion. Lastly, other classifier models, not limited to logistic regression model, can be explored, to make construction of explanatory variables easier, albeit at the expense of intuitive interpretability.

7.3. Concluding Remarks

Since the beginning of the 21st century until 2019, the number of Japanese Nobel Prize laureates in the field of natural sciences (18) is second only to the U.S (75), followed by the U.K. (14) according to the Nobel Foundation and Ministry of Education, Culture, Sports, Science and Technology of Japan. In fact, no other non-Western country has ever matched Japan in the number of Nobel Prize winners. Moreover, the Japanese government has strived to promote academic startups to utilize scientific outcomes, as any other governments do as discussed in the beginning of Chapter 1. In 2018, 2278

university-based startups were identified across Japan, a record high as recognized by the Ministry of Economy, Trade and Industry (METI) of Japan, as seen in Figure 7-1 [105].

Despite historic success of scientific advancement and recent endeavor of enhancing academic entrepreneurship, we can argue that now the foothold of Japanese academic startups, especially those based on intense science linkage, is being threatened.

Nature, an international weekly journal of science, recently reported the country's malaise in its scientific policy as follows [106, 107, 108]. Following reforms in 2004, Japanese national universities' budgets have declined by 1% every year. The move was meant to make universities themselves more responsible for their budgets, by aligning their research with industrial needs. But it has triggered a few negative changes such as decreasing research funding, ballooning administrative burdens for grants, curbing new hires of permanent faculty members, and forcing young researchers into unstable limited term employment. University researchers say they now can put little more than one-third of their work time into research, compared to just under half in 2002. This reportedly forced young researchers to aim for results that can be accomplished in the short term, and possibly to aim for less of academic research than commercial experimentation, in which scientific originality and creativity are difficult to realize. As a result of such austerity government policy, Japan's number of publications in all scientific fields has stagnated, whereas the U.S., China, the U.K., and South Korea are rising, according to publisher Elsevier's Scopus database. Between 2005 and 2015, Japan's global share declined by more than a third, while China experienced extraordinary growth. Given such situations, we can assume that an indispensable part of the recent driving force in Japan of increasing academic startups is arguably such austerity, at the expense of intrinsic scientific strength of researchers' future generations, whereas, in other leading countries above, continuing prosperity of academic startups have been propelled by their growing scientific presence and prominence.

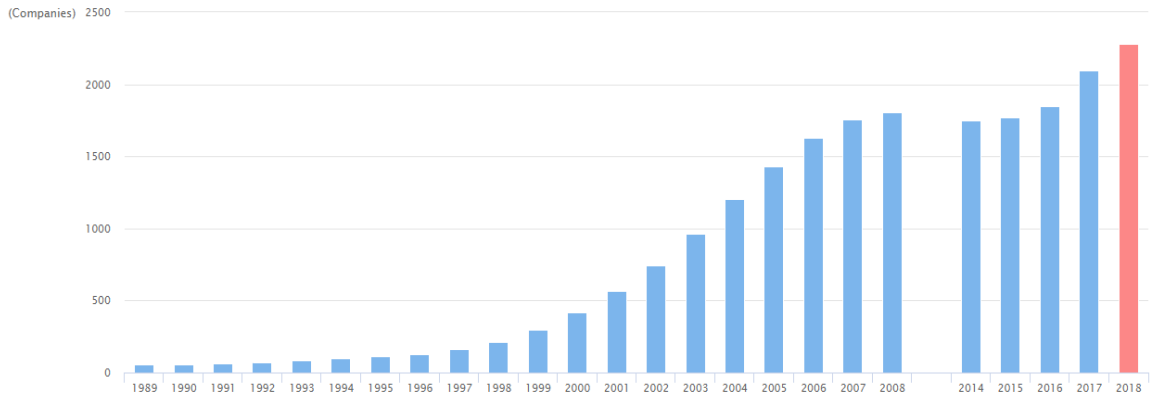


Figure 7-1 Trend in the Number of Japanese University Startups (METI survey [105])

As clarified in this thesis, as far as academic startups intensively linked to science in the emerging biopharmaceutical domain are concerned, the most performing determinants regarding startup readiness are Paper-related Features, by a wide margin than other groups of features such as Patent-related Features, Hot Topic Factors and Ecosystem Factors. Paper-related Features represent researchers' prominence as scientists in terms of academic capability, profile, initiative and responsibility, and they are particularly important for assessing Exit within a certain range of biopharmaceutical research topics.

The government and university administrators of Japan are anticipated to move on to take measures to support and encourage such determinants, in order to positively impact promise of academic startups with intense scientific linkage in the country, in a way that mere austerity cannot. It is also expected that Japanese academic researchers with interest in such startups, to know this finding, and develop their academic presence and prominence prior to their participation in startups, in order to achieve a successful outcome such as Exit in their foreseeable future.

References

[1] J. C. Mankins, "Technology readiness assessments: A retrospective," *Acta Astronautica,* vol. 65, no. 9–10, pp. 1216-1223, 2009.

[2] H. Nakamura, Y. Kajikawa and S. Suzuki, "Multi-level perspectives with technology readiness measures for aviation innovation," *Sustainability science,* vol. 8, no. 1, pp. 87-101, 2013.

[3] T. Goji, T. Matsuda and I. Sakata, "Measuring" Start-Up Readiness" of Scientific Research-Based Start-Ups Using Analysis of Citation Networks: Case Study of CRISPR-Cas9," *Portland International Conference on Management of Engineering and Technology (PICMET),* pp. 1-14, 2017.

[4] T. Goji, Y. Hayashi, H. Yamano, T. Matsuda and I. and Sakata, "Assessing" Start-up Readiness" for Research Topics and Researchers: Case Studies of Research-Based Start-Ups in the Biopharmaceutical Domain," *Portland International Conference on Management of Engineering and Technology (PICMET),* pp. 1-17, 2018.

[5] R. A. Rader, "(Re) defining biopharmaceutical," *Nature biotechnology,* p. 743, 2008.

[6] P. Stephan, S. Gurmu, A. J. Sumell and G. Black, "Who's patenting in the university? Evidence from the survey of doctorate recipients," *Econ. Innov. New Techn.,* vol. 16, no. 2, pp. 71-99, 2007.

[7] D. E. Stokes, Pasteur's quadrant: Basic science and technological innovation, Brookings Institution Press, 2011.

[8] R. Henderson, A. B. Jaffe and M. Trajtenberg, "Universities as a source of commercial technology: a detailed analysis of university patenting, 1965-1988," *Review of Economics and statistics,* vol. 80, no. 1, pp. 119-127, 1998.

[9] A. N. Link, D. S. Siegel and B. Bozeman, "An empirical analysis of the propensity of academics to engage in informal university technology transfer," *Industrial and Corporate Change,* vol. 16, no. 4, pp. 641-655, 2007.

[10] W. W. Powell and J. Owen-Smith, "Universities and the market for intellectual property in the life sciences," *Journal of Policy Analysis and Management: The Journal of the Association for Public Policy Analysis and Management,* vol. 17, no. 2, pp. 253-277, 1998.

[11] EvaluatePharma, EvaluatePharma World Preview 2018, Outlook to 2024, London, UK: EvaluatePharma, 2018.

[12] Ministry of Health, Labor and Welfare (MHLW) of Japan, "Report from "Council

for Promoting Ventures Innovating Healthcare"," Ministry of Health, Labor and Welfare (MHLW) of Japan, July 2016.

[13] W. Dolfsma and D. Seo, "Government policy and technological innovation — a suggested typology," *Technovation,* vol. 33, no. 6-7, pp. 173-179, 2013.

[14] F. T. Rothaermel, S. D. Agung and L. Jiang, "University entrepreneurship: a taxonomy of the literature," *Industrial and corporate change,* vol. 16, no. 4, pp. 691-791, 2007.

[15] J. Bercovitz and M. Feldman, "Academic entrepreneurs: Organizational change at the individual level," *Organization science,* vol. 19, no. 1, pp. 69-89, 2008.

[16] S. Jain, G. George and M. Maltarich, "Academics or entrepreneurs? Investigating role identity modification of university scientists involved in commercialization activity," *Research policy,* vol. 38, no. 6, pp. 922-935, 2009.

[17] B. Clarysse, V. Tartari and A. Salter, "The impact of entrepreneurial capacity, experience and organizational support on academic entrepreneurship," *Research policy,* vol. 40, no. 8, pp. 1084-1093, 2011.

[18] M. Abreu and V. Grinevich, "The nature of academic entrepreneurship in the UK: Widening the specifically examine entrepreneurial activities," *Research Policy,* vol. 42, no. 2, pp. 408-422, 2013.

[19] T. Aldridge, D. Audretsch, S. Desai and V. Nadella, "Scientist entrepreneurship across scientific fields," *The Journal of Technology Transfer,* vol. 39, no. 6, pp. 819-835, 2014.

[20] T. Goji, Y. Hayashi, H. Yamano, T. Matsuda and I. Sakata, "Researchers' "Startup Readiness" in the Biopharmaceutical Domain Assessed using Logistic Regression for Features of Their Papers, Patents, Institutes, and Nations," *2019 Portland International Conference on Management of Engineering and Technology (PICMET),* 2019.

[21] P. E. Auerswald and L. Dani, "The adaptive life cycle of entrepreneurial ecosystems: the biotechnology cluster," *Small Business Economics,* vol. 49, no. 1, pp. 97-117, 2017.

[22] C. Declan, C. van Egeraat and C. O'Gorman, "Inherited competence and spin-off performance," *European Planning Studies,* vol. 24, no. 3, pp. 443-462, 2016.

[23] T. G. P. C. A. Allen, S. Woerner and O. Raz, "The power of reciprocal knowledge sharing relationships for startup success," *Journal of Small Business and Enterprise Development,* vol. 23, no. 3, pp. 636-651, 2016.

[24] R. Landry, N. Amara and I. Rherrad, "Why are some university researchers more

likely to create spin-offs than others? Evidence from Canadian universities.," *Research Policy,* vol. 35, no. 10, pp. 1599-1615, 2006.

[25] S. Krabel and P. Mueller, "What drives scientists to start their own company?: An empirical investigation of Max Planck Society scientists," *Research Policy,* vol. 38, no. 6, pp. 947-956., 2009.

[26] G. Criaco, T. Minola, P. Migliorini and C. Serarols-Tarrés, "“To have and have not”: founders’ human capital and university start-up survival," *The Journal of Technology Transfer,* vol. 39, no. 4, pp. 567-593, 2014.

[27] T. Huynh, D. Patton, D. Arias-Aranda and L. Molina-Fernández, "University spin-off's performance: Capabilities and networks of founding teams at creation phase," *Journal of Business Research,* vol. 78, pp. 10-22, 2017.

[28] J. Barney, "Firm resources and sustained competitive advantage," *Journal of management,* vol. 17, no. 1, pp. 99-120, 1991.

[29] K. R. Conner and C. K. Prahalad, "A resource-based theory of the firm: Knowledge versus opportunism," *Organization science,* vol. 7, no. 5, pp. 477-501, 1996.

[30] R. M. Grant, "Toward a knowledge-based theory of the firm," *Strategic management journal,* vol. 17, no. S2, pp. 109-122, 1996.

[31] E. Rasmussen and O. J. Borch, "University capabilities in facilitating entrepreneurship: A longitudinal study of spin-off ventures at mid-range universities," *Research policy,* vol. 39, no. 5, pp. 602-612, 2010.

[32] M. Knockaert, A. Spithoven and B. Clarysse, "The knowledge paradox explored: what is impeding the creation of ICT spin-offs?," *Technology Analysis & Strategic Management,* vol. 22, no. 4, pp. 479-493, 2010.

[33] C. Corsi, A. Prencipe, M. Rodríguez-Gulías, D. Rodeiro-Pazos and S. Fernández-López, "Growth of KIBS and non-KIBS firms: evidences from university spin-offs," *The Service Industries Journal,* vol. 39, no. 1, pp. 43-64, 2019.

[34] B. Kogut and U. Zander, "Knowledge of the firm, combinative capabilities, and the replication of technology," *Organization science,* vol. 3, no. 3, pp. 383-397, 1992.

[35] OECD, "Entrepreneurship at a Glance," OECD, 2018.

[36] P. Azoulay, W. Ding and T. Stuart, "The impact of academic patenting on the rate, quality and direction of (public) research output," *The Journal of Industrial Economics,* vol. 57, no. 4, pp. 637-676, 2009.

[37] D. C. Fehder, F. Murray and S. Stern, "Intellectual property rights and the evolution of scientific journals as knowledge platforms," *International Journal of Industrial Organization,* vol. 36, pp. 83-94, 2014.

[38] N. Shibata, Y. Kajikawa, Y. Takeda and K. Matsushima, "Detecting emerging research fronts based on topological measures in citation networks of scientific publications," *Technovation,* vol. 28, no. 11, pp. 758-775, 2008.

[39] N. Shibata, Y. Kajikawa, Y. Takeda, I. Sakata and K. Matsushima, "Detecting emerging research fronts in regenerative medicine by the citation," *Technological Forecasting & Social Change,* vol. 78, pp. 274-282, 2011.

[40] H. Sasaki, T. Hara and I. Sakata, "Identifying Emerging Research Related to Solar Cells Field using a Machine Learning Approach," *Journal of Sustainable Development of Energy, Water and Environment Systems,* vol. 4, no. 4, pp. 418-429, 2016.

[41] L. C. Freeman, "Centrality in social networks conceptual clarification," *Social networks,* vol. 1, no. 3, pp. 215-239, 1978.

[42] L. C. Freeman, "A set of measures of centrality based on betweenness," *Sociometry,* pp. 35-41, 1977.

[43] P. Bonacich, "Technique for analyzing overlapping memberships," *Sociological methodology,* vol. 4, pp. 176-185, 1972.

[44] R. S. Burt, "Structural holes and good ideas," *American journal of sociology,* vol. 110, no. 2, pp. 349-399, 2004.

[45] D. J. Watts and S. H. Strogatz, "Collective dynamics of ‘small-world’networks," *nature,* vol. 393, no. 6684, p. 440, 1998.

[46] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine," *Computer networks and ISDN systems,* vol. 30, no. 1-7, pp. 107-117, 1998.

[47] E. Van der Pol, A. Böing, P. Harrison, A. Sturk and R. Nieuwland, "Classification, functions, and clinical relevance of extracellular vesicles," *Pharmacological reviews,* vol. 64, no. 3, pp. 676-705, 2012.

[48] S. Keller, M. Sanderson, A. Stoeck and P. Altevogt, "Exosomes: from biogenesis and secretion to biological function," *Immunology letters,* vol. 107, no. 2, pp. 102-108, 2006.

[49] A. M. Booth, Y. Fang, J. K. Fallon, J.-M. Yang, J. E. Hildreth and S. J. Gould, "Exosomes and HIV Gag bud from endosome-like domains of the T cell plasma membrane," *J Cell biol,* vol. 172, no. 6, pp. 923-935, 2006.

[50] J. Lederberg and A. T. McCray, "Ome SweetOmics - A Genealogical Treasury of Words," *The Scientist,* vol. 15, no. 7, p. 8, 2001.

[51] J. Peterson, S. Garges, M. Giovanni, P. McInnes, L. Wang, J. Schloss, V. Bonazzi, J. McEwen, K. Wetterstrand, C. Deal and C. Baker, "The NIH human microbiome project," *Genome research,* vol. 19, no. 12, pp. 2317-2323, 2009.

[52] F. Bäckhed, R. E. Ley, J. L. Sonnenburg, D. A. Peterson and J. I. Gordon, "Host-bacterial mutualism in the human intestine," *Science,* vol. 307, no. 5717, pp. 1915-1920, 2005.

[53] P. Turnbaugh, R. Ley, M. Hamady, C. Fraser-Liggett, R. Knight and J. Gordon, "The human microbiome project," *Nature,* vol. 449, no. 7164, p. 804, 2007.

[54] R. E. Ley, D. A. Peterson and J. I. Gordon, "Ecological and evolutionary forces shaping microbial diversity in the human intestine," *Cell,* vol. 124, no. 4, pp. 837-848, 2006.

[55] T. Nakatsuji, T. Chen, S. Narala, K. Chun, A. Two, T. Yun, F. Shafiq, P. Kotol, A. Bouslimani, A. Melnik and H. Latif, "Antimicrobials from human skin commensal bacteria protect against Staphylococcus aureus and are deficient in atopic dermatitis," *Science translational medicine,* vol. 9, no. 378, p. eaah4680, 2017.

[56] K. Egelie, G. Graff, S. Strand and B. Johansen, "The emerging patent landscape of CRISPR–Cas gene editing technology," *Nature biotechnology,* vol. 34, no. 10, p. 1025, 2016.

[57] E. Deltcheva, K. Chylinski, C. Sharma, K. Gonzales, Y. Chao, Z. Pirzada, M. Eckert, J. Vogel and E. Charpentier, "CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III," *Nature,* vol. 471, no. 7340, p. 602, 2011.

[58] L. Cong, F. Ran, D. Cox, S. Lin, R. Barretto, N. Habib, P. Hsu, X. Wu, W. Jiang, L. Marraffini and F. Zhang, "Multiplex genome engineering using CRISPR/Cas systems," *Science,* vol. 339, no. 6121, pp. 819-823, 2013.

[59] M. Sadelain, R. Brentjens and I. Rivière, "The basic principles of chimeric antigen receptor design," *Cancer discovery,* vol. 3, no. 4, pp. 388-398, 2013.

[60] S. Srivastava and S. R. Riddell, "Engineering CAR-T cells: design concepts," *Trends in immunology* , vol. 36, no. 8, pp. 494-502, 2015.

[61] J. Hartmann, M. Schüßler-Lenz, A. Bondanza and C. Buchholz, "Clinical development of CAR T cells—challenges and opportunities in translating innovative treatment concepts," *EMBO molecular medicine,* vol. 9, no. 9, pp. 1183-1197, 2017.

[62] World Health Organization, "WHO statement on the first meeting of the

International Health Regulations Emergency Committee on Zika virus," 1 February 2016.

[63] L. H. Chen and D. H. Hamer, "Zika virus: rapid spread in the western hemisphere," *Annals of internal medicine,* vol. 164, no. 9, pp. 613-615, 2016.

[64] D. Musso, E. J. Nilles and V. Cao-Lormeau, "Rapid spread of emerging Z ika virus in the P acific area," *Clinical Microbiology and Infection,* vol. 20, no. 10, pp. 595-596., 2014.

[65] European Centre for Disease Prevention and Control, "Factsheet about Zika virus disease," 23 Jun 2016,.

[66] P. Brasil, J. Pereira Jr, M. Moreira, R. Ribeiro Nogueira, L. Damasceno, M. Wakimoto, R. Rabello, S. Valderramos, U. Halai, T. Salles and A. Zin, "Zika virus infection in pregnant women in Rio de Janeiro," *New England Journal of Medicine,* vol. 375, no. 24, pp. 2321-2334, 2016.

[67] World Health Organization, "Zika situation report," 5 February 2016.

[68] V. Sikka, V. Chattu, R. Popli, S. Galwankar, D. Kelkar, S. Sawicki, S. Stawicki and T. Papadimos, "The emergence of Zika virus as a global health security threat: a review and a consensus statement of the INDUSEM Joint Working Group (JWG)," *Journal of global infectious diseases,* vol. 8, no. 1, p. 3, 2016.

[69] R. A. Fisher, "On the interpretation of χ 2 from contingency tables, and the calculation of P," *Journal of the Royal Statistical Society,* vol. 85, no. 1, pp. 87-94, 1922.

[70] R. A. Fisher, Statistical methods for research workers, Genesis Publishing Pvt. Ltd., 1925.

[71] A. Agresti, "A survey of exact inference for contingency tables," *Statistical science,* vol. 7, no. 1, pp. 131-153, 1992.

[72] F. Mosteller, "Association and estimation in contingency tables," *Journal of the American Statistical Association,* vol. 63, no. 321, pp. 1-28, 1968.

[73] A. W. Edwards, "The measure of association in a 2× 2 table," *Journal of the Royal Statistical Society: Series A (General),* vol. 126, no. 1, pp. 109-114, 1963.

[74] S. Kuhle, B. Maguire, H. Zhang, D. Hamilton, A. C. Allen, K. S. Joseph and V. M. Allen, "Comparison of logistic regression with machine learning methods for the prediction of fetal growth abnormalities: a retrospective cohort study," *BMC Pregnancy Childbirth,* vol. 18, no. 333, 2018.

[75] A. Grandi and R. Grimaldi, "Exploring the networking characteristics of new

venture founding teams: A stdy of italian academic spin-off," *Small Business Economics,* vol. 21, no. 4, pp. 329-341, 2003.

[76] M. Feldman, I. Feller, J. Bercovitz and R. Burton, "Equity and the technology transfer strategies of American research universities," *Management Science,* vol. 48, no. 1, pp. 105-121, 2002.

[77] D. Di Gregorio and S. Shane, "Why do some universities generate more start-ups than others?," *Research policy,* vol. 32, no. 2, pp. 209-227, 2003.

[78] J. Bercovitz and M. Feldman, "Academic entrepreneurs: Social learning and participation in university technology transfer," Hubert H. Humphrey Institute of Public Affairs, University of Minnesota, 2004.

[79] L. G. Zucker, M. R. Darby and J. S. Armstrong, "Commercializing knowledge: University science, knowledge capture, and firm performance in biotechnology," *Management science,* vol. 48, no. 1, pp. 138-153, 2002.

[80] E. Autio, M. Kenney, P. Mustar, D. Siegel and M. Wright, "Entrepreneurial innovation: The importance of context," *Research Policy,* vol. 43, no. 7, pp. 1097-1108, 2014.

[81] Z. Zhang, "Multivariable fractional polynomial method for regression model," *Annals of translational medicine,* vol. 4, no. 9, 2016.

[82] G. Shmueli, "To explain or to predict?," *Statistical science,* vol. 25, no. 3, pp. 289-310, 2010.

[83] B. G. Tabachnick and L. S. Fidell, Using Multivariate Statistics, Boston: Pearson Education, 2013.

[84] P. Allison, "Logistic regression for rare events," *Statistical Horizons,* vol. 13, 2012.

[85] G. James, D. Witten, T. Hastie and R. Tibshirani, An introduction to statistical learning. Vol. 112., New York: springer, 2013.

[86] P. Bruce and A. Bruce, Practical statistics for data scientists: 50 essential concepts, O'Reilly Media, Inc., 2017.

[87] H. Akaike, "A new look at the statistical model identification," *.AKAIKE, H. "A new look at the statistical model identification." IEEE Transactions on Automatic Control 19 (1974): 716-723." Selected Papers of Hirotugu Akaike. Springer, New York, NY, 1974. 215-222.,* vol. 19, pp. 716-723, 1974.

[88] T. Yamashita, K. Yamashita and R. Kamimura, "A stepwise AIC method for variable selection in linear regression," *Communications in Statistics—Theory and Methods,* vol. 36, no. 13, pp. 2395-2403, 2007.

[89] G. Ambler and A. Benner, "The mfp Package," CRAN, 2006.

[90] R. M. O'brien, "A caution regarding rules of thumb for variance inflation factors," *Quality & quantity,* vol. 41, no. 5, pp. 673-690, 2007.

[91] D. W. Hosmer, S. Lemeshow and R. X. Sturdivant, Applied Logistic Regression, Third Edition, John Wiley & Sons, Inc., 2013.

[92] G. Osius and D. Rojek, "Normal goodness-of-fit tests for multinomial models with large degrees of freedom," *Journal of the American Statistical Association,* vol. 87, no. 420, pp. 1145-52, 1992.

[93] Z. Zhang, "Residuals and regression diagnostics: focusing on logistic regression," *Annals of translational medicine,* vol. 4, no. 10, p. 195, 2016.

[94] S. K. Sarkar, H. Midi and S. Rana, "Detection of outliers and influential observations in binary logistic regression: An empirical study," *Journal of Applied Sciences* , vol. 11, no. 1 , pp. 26-35, 2011.

[95] R. D. Cook, "Detection of influential observation in linear regression," *Technometrics,* vol. 19, no. 1, pp. 15-18, 1977.

[96] N. Martín and L. Pardo, "On the asymptotic distribution of Cook's distance in logistic regression models," *Journal of Applied Statistics,* vol. 36, no. 10, pp. 1119-1146., 2009.

[97] The Pennsylvania State University, "9.2 - Using Leverages to Help Identify Extreme X Values, STAT 462: Applied Regression Analysis," 2018. [Online]. Available: https://newonlinecourses.science.psu.edu/stat462/node/171/. [Accessed 30 9 2019].

[98] The Pennsylvania State University, "9.4 - Studentized Residuals, STAT 462: Applied Regression Analysis," 2018. [Online]. Available: https://newonlinecourses.science.psu.edu/stat462/node/247/. [Accessed 30 9 2019].

[99] The Pennsylvania State University, "9.5 - Identifying Influential Data Points, STAT 462: Applied Regression Analysis," 2018. [Online]. Available: https://newonlinecourses.science.psu.edu/stat462/node/173/. [Accessed 30 9 2019].

[100] J. Fox, Regression Diagnostics: An Introduction, Sage Publications., 1991.

[101] J. Fox, A Mathematical Primer for Social Statistics, SAGE Publications, Inc, 2009.

[102] S. A. Shane, . Academic entrepreneurship: University spinoffs and wealth creation, Edward Elgar Publishing, 2004.

[103] M. Jinek, K. Chylinski, I. Fonfara, M. Hauer, J. A. Doudna and E. Charpentier, "A programmable dual-RNA–guided DNA endonuclease in adaptive bacterial

immunity," *Science,* vol. 337, no. 6096, pp. 816-821, 2012.

[104] E. S. Lander, "The heroes of CRISPR," *Cell,* vol. 164, no. 1-2, pp. 18-28, 2016.

[105] Ministry of Economy, Trade and Industry (METI) of Japan, "FY2018 University-Oriented Venture Business Survey," Ministry of Economy, Trade and Industry (METI) of Japan, May 2019.

[106] Editorial, "Budget cuts fuel frustration among Japan's academics," *Nature,* vol. 548, p. 259, 2017.

[107] N. Phillips, "Japan loses share of research articles to China," *Nature,* vol. 550, pp. 310-311, 24 3 2017.

[108] N. Phillips, "Japanese research leaders warn about national science decline," *Nature,* vol. 550, pp. 310-311, 2017.

Appendices

APPENDIX A-1 HEAT MAP OF TOP 100 AUTHORS HIGHLIGHTING FOUNDERS, RANKED BY FIVE CENTRALITIES (2012～2016, LAST NAME ONLY)

(… CONTINUED ON NEXT PAGE)

	2012					2013					2014					2015					2016				
Ranking	Betweenness Centrality	Closeness Centrality	Degree Centrality	Eigenvector Centrality	Pagerank	Betweenness Centrality	Closeness Centrality	Degree Centrality	Eigenvector Centrality	Pagerank	Betweenness Centrality	Closeness Centrality	Degree Centrality	Eigenvector Centrality	Pagerank	Betweenness Centrality	Closeness Centrality	Degree Centrality	Eigenvector Centrality	Pagerank	Betweenness Centrality	Closeness Centrality	Degree Centrality	Eigenvector Centrality	Pagerank
1	Barrangou	Barrangou	Barrangou	Barrangou	Barrangou	Church	Church	Church	Church	Church	Zhang	Zhang	Zhang	Zhang	Zhang	Zhang	Zhang	Zhang	Zhang	Zhang	Zhang	Zhang	Zhang	Zhang	Zhang
2	Gasiunas	Gasiunas	Gasiunas	Gasiunas	Gasiunas	Esvelt	Esvelt	Esvelt	Esvelt	Esvelt	Church	Church	Church	Church	Church	Hsu	Hsu	Hsu	Hsu	Hsu	Hsu	Hsu	Hsu	Hsu	Hsu
3	Horvath	Horvath	Horvath	Horvath	Horvath	Jinek	Aach	Jinek	Aach	Aach	Doudna	Hsu	Hsu	Hsu	Esvelt	Ran	Ran	Ran	Ran	Ran	Ran	Ran	Ran	Ran	Ran
4	Siksnys	Siksnys	Siksnys	Siksnys	Siksnys	Aach	Norville	Aach	Norville	Norville	Esvelt	Esvelt	Esvelt	Esvelt	Hsu	Church	Marraffini	Marraffini	Marraffini	Marraffini	Scott	Marraffini	Marraffini	Wu	Scott
5	Fremaux	Fremaux	Fremaux	Fremaux	Fremaux	Norville	Mali	Norville	Mali	Mali	Hsu	Ran	Ran	Ran	Ran	Esvelt	Wu	Wu	Wu	Church	Church	Wu	Wu	Marraffini	Marraffini
6	Sapranauskas	Sapranauskas	Sapranauskas	Sapranauskas	Sapranauskas	Zhang	Yang	Zhang	Yang	Yang	Ran	Mali	Mali	Mali	Mali	Marraffini	Church	Church	Church	Wu	Doudna	Church	Church	Church	Wu
7	Charpentier	Charpentier	Charpentier	Charpentier	Charpentier	Mali	Zhang	Mali	Zhang	Zhang	Mali	Aach	Aach	Marraffini	Aach	Doudna	Jiang	Jiang	Jiang	Esvelt	Esvelt	Scott	Cong	Cong	Church
8	Chylinski	Chylinski	Chylinski	Chylinski	Chylinski	Yang	DiCarlo	Yang	DiCarlo	DiCarlo	Aach	Marraffini	Marraffini	Aach	Marraffini	Wu	Esvelt	Esvelt	Esvelt	Jiang	Marraffini	Cong	Scott	Jiang	Jiang
9	Doudna	Doudna	Doudna	Doudna	Doudna	DiCarlo	Hsu	DiCarlo	Hsu	Hsu	Norville	Wu	Wu	Wu	Norville	Scott	Cox	Cox	Cong	Mali	Wu	Jiang	Jiang	Habib	Cong
10	Fonfara	Fonfara	Fonfara	Fonfara	Fonfara	Guell	Jiang	Guell	Jiang	Jiang	Marraffini	Norville	Norville	Norville	Wu	Mali	Lin	Lin	Mali	Cox	Mali	Doudna	Doudna	Mali	Esvelt
11	Hauer	Hauer	Hauer	Hauer	Hauer	Doudna	Marraffini	Doudna	Marraffini	Marraffini	Wu	Jiang	Jiang	Yang	DiCarlo	Lin	Mali	Habib	Habib	Lin	Aach	Cox	Cox	Cox	Cox
12	Jinek	Jinek	Jinek	Jinek	Jinek	Hsu	Ran	Hsu	Ran	Ran	DiCarlo	Cox	Cox	Jiang	Yang	Jiang	Habib	Mali	Cox	Habib	Jiang	Habib	Habib	Doudna	Mali
13	n/a	n/a	n/a	n/a	n/a	Jiang	Wu	Jiang	Wu	Wu	Charpentier	Lin	Lin	Lin	Jiang	Aach	Cong	Cong	Lin	Cong	Cong	Esvelt	Esvelt	Esvelt	Habib
14	n/a	n/a	n/a	n/a	n/a	Marraffini	Guell	Marraffini	Guell	Guell	Lin	Yang	Yang	Cox	Cox	Cox	Aach	Aach	Aach	Aach	Lin	Mali	Mali	Aach	Lin
15	n/a	n/a	n/a	n/a	n/a	Ran	Cox	Ran	Cox	Cox	Yang	DiCarlo	DiCarlo	DiCarlo	Lin	Habib	Barretto	Barretto	Barretto	Barretto	Charpentier	Lin	Lin	Lin	Doudna
16	n/a	n/a	n/a	n/a	n/a	Wu	Lin	Wu	Lin	Lin	Jiang	Cong	Cong	Cong	Guell	Cong	Norville	Norville	Yang	Scott	Cox	Aach	Aach	Barretto	Barretto
17	n/a	n/a	n/a	n/a	n/a	Cox	Barretto	Cox	Barretto	Barretto	Cox	Barretto	Barretto	Barretto	Cong	Norville	Doudna	Doudna	Doudna	Norville	Habib	Barretto	Barretto	Scott	Aach
18	n/a	n/a	n/a	n/a	n/a	Lin	Cong	Lin	Cong	Cong	Guell	Habib	Habib	Habib	Barretto	Barretto	Yang	Yang	Norville	Yang	Norville	Norville	Norville	Yang	Norville
19	n/a	n/a	n/a	n/a	n/a	Barretto	Habib	Barretto	Habib	Habib	Jinek	Guell	Guell	Doudna	Habib	Yang	DiCarlo	DiCarlo	DiCarlo	Doudna	Yang	Yang	Yang	Norville	Yang
20	n/a	n/a	n/a	n/a	n/a	Cong	Jinek	Cong	Jinek	Jinek	Cong	Doudna	Doudna	Guell	Doudna	Charpentier	Scott	Scott	Guell	DiCarlo	Joung	DiCarlo	DiCarlo	DiCarlo	DiCarlo
21	n/a	n/a	n/a	n/a	n/a	Habib	Doudna	Habib	Doudna	Doudna	Barretto	Charpentier	Charpentier	Charpentier	Charpentier	DiCarlo	Guell	Guell	Scott	Guell	Barretto	Charpentier	Charpentier	Charpentier	Guell
22	n/a	n/a	n/a	n/a	n/a	Charpentier	Charpentier	Charpentier	Charpentier	Charpentier	Habib	Jinek	Jinek	Jinek	Jinek	Guell	Charpentier	Charpentier	Charpentier	Charpentier	Agarwala	Guell	Guell	Guell	Agarwala
23	n/a	n/a	n/a	n/a	n/a	Chylinski	Chylinski	Chylinski	Chylinski	Chylinski	Chylinski	Chylinski	Chylinski	Hauer	Chylinski	Jinek	Jinek	Jinek	Jinek	Jinek	Jinek	Jinek	Jinek	Jinek	Charpentier
24	n/a	n/a	n/a	n/a	n/a	Fonfara	Fonfara	Fonfara	Fonfara	Fonfara	Hauer	Hauer	Hauer	Chylinski	Hauer	Agarwala	Hauer	Hauer	Konermann	Agarwala	DiCarlo	Hauer	Hauer	Konermann	Jinek
25	n/a	n/a	n/a	n/a	n/a	Hauer	Hauer	Hauer	Hauer	Hauer	Fonfara	Fonfara	Fonfara	Fonfara	Fonfara	Chylinski	Chylinski	Chylinski	Hauer	Hauer	Shalem	Chylinski	Chylinski	Hauer	Shalem
26	n/a	n/a	n/a	n/a	n/a	Fu	Fu	Fu	Fu	Fu	Scott	Scott	Scott	Scott	Scott	Hauer	Konermann	Konermann	Chylinski	Chylinski	Guell	Konermann	Konermann	Chylinski	Hauer
27	n/a	n/a	n/a	n/a	n/a	Hwang	Hwang	Hwang	Hwang	Hwang	Cheng	Cheng	Cheng	Cheng	Cheng	Fonfara	Fonfara	Fonfara	Fonfara	Fonfara	Lander	Shalem	Shalem	Fonfara	Joung
28	n/a	n/a	n/a	n/a	n/a	Joung	Joung	Joung	Joung	Joung	Jaenisch	Jaenisch	Jaenisch	Jaenisch	Jaenisch	Lander	Agarwala	Agarwala	Shalem	Konermann	Sander	Fonfara	Fonfara	Shalem	Chylinski
29	n/a	n/a	n/a	n/a	n/a	Maeder	Maeder	Maeder	Maeder	Maeder	Konermann	Konermann	Konermann	Konermann	Konermann	Konermann	Shalem	Shalem	Cheng	Shalem	Kim	Joung	Joung	Joung	Fonfara
30	n/a	n/a	n/a	n/a	n/a	Peterson	Peterson	Peterson	Peterson	Peterson	Shivalila	Shivalila	Shivalila	Shivalila	Shivalila	Shalem	Cheng	Cheng	Jaenisch	Cheng	Chylinski	Agarwala	Agarwala	Jaenisch	Konermann
31	n/a	n/a	n/a	n/a	n/a	Reyon	Reyon	Reyon	Reyon	Reyon	Wang	Wang	Wang	Wang	Wang	Joung	Jaenisch	Jaenisch	Bao	Jaenisch	Hauer	Jaenisch	Jaenisch	Agarwala	Sander
32	n/a	n/a	n/a	n/a	n/a	Sander	Sander	Sander	Sander	Sander	Yang	Yang	Yang	Yang	Yang	Cheng	Wang	Wang	Wang	Wang	Fonfara	Sander	Sander	Cheng	Lander
33	n/a	n/a	n/a	n/a	n/a	Tsai	Tsai	Tsai	Tsai	Tsai	Joung	Joung	Joung	Joung	Joung	Jaenisch	Bao	Shivalila	Cradick	Joung	Konermann	Cheng	Cheng	Bao	Jaenisch
34	n/a	n/a	n/a	n/a	n/a	Yeh	Yeh	Yeh	Yeh	Yeh	Shalem	Shalem	Shalem	Sander	Shalem	Wang	Shivalila	Yang	Agarwala	Shivalila	Jaenisch	Bao	Bao	Sander	Cheng
35	n/a	n/a	n/a	n/a	n/a	Cheng	Cheng	Cheng	Cheng	Cheng	Reyon	Sander	Sander	Shalem	Sander	Reyon	Yang	Bao	Joung	Yang	Musunuru	Kim	Kim	Cradick	Kim
36	n/a	n/a	n/a	n/a	n/a	Jaenisch	Jaenisch	Jaenisch	Jaenisch	Jaenisch	Sander	Reyon	Reyon	Reyon	Reyon	Sander	Joung	Joung	Shivalila	Bao	Tsai	Wang	Wang	Kim	Wang
37	n/a	n/a	n/a	n/a	n/a	Shivalila	Shivalila	Shivalila	Shivalila	Shivalila	Dawlaty	Dawlaty	Agarwala	Fu	Dawlaty	Shivalila	Cradick	Cradick	Yang	Cradick	Cheng	Cradick	Lander	Wang	Bao
38	n/a	n/a	n/a	n/a	n/a	Wang	Wang	Wang	Wang	Wang	Agarwala	Agarwala	Dawlaty	Lin	Agarwala	Yang	Fine	Fine	Fine	Fine	Wright	Lander	Cradick	Fine	Yang
39	n/a	n/a	n/a	n/a	n/a	Yang	Yang	Yang	Yang	Yang	Fu	Fu	Fu	Bao	Fu	Fu	Li	Li	Li	Lander	Qi	Yang	Yang	Gootenberg	Cradick
40	n/a	n/a	n/a	n/a	n/a	Dawlaty	Lin	Dawlaty	Lin	Lin	Tsai	Bao	Bao	Cradick	Bao	Kim	Reyon	Reyon	Sander	Reyon	Reyon	Gootenberg	Gootenberg	Li	Wright
41	n/a	n/a	n/a	n/a	n/a	Lin	Dawlaty	Lin	Dawlaty	Dawlaty	Bao	Cradick	Cradick	Kim	Cradick	Bao	Sander	Sander	Weinstein	Sander	Wang	Fine	Shivalila	Yang	Shivalila
42	n/a	n/a	n/a	n/a	n/a	Ma	Ma	Ma	Ma	Ma	Cradick	Fine	Fine	Fine	Fine	Cradick	Fu	Fu	Reyon	Li	Bao	Shivalila	Fine	Shivalila	Gootenberg
43	n/a	n/a	n/a	n/a	n/a	Cho	Cho	Cho	Cho	Cho	Lin	Lin	Lin	Agarwala	Tsai	Maeder	Weinstein	Weinstein	Lin	Fu	Yang	Li	Li	Trevino	Tsai
44	n/a	n/a	n/a	n/a	n/a	Kim	Kim	Kim	Kim	Kim	Hwang	Kim	Kim	Dawlaty	Lin	Fine	Kim	Kim	Kim	Dawlaty	Maeder	Tsai	Tsai	Reyon	Fine
45	n/a	n/a	n/a	n/a	n/a	Kim	Kim	Kim	Kim	Kim	Maeder	Tsai	Tsai	Ma	Li	Chen	Dawlaty	Dawlaty	Fu	Weinstein	Cradick	Reyon	Reyon	Lin	Reyon
46	n/a	n/a	n/a	n/a	n/a	Kim	Kim	Kim	Kim	Kim	Peterson	Li	Li	Tsai	Kim	Tsai	Lin	Lin	Trevino	Kim	Fu	Qi	Qi	Tsai	Li
47	n/a	n/a	n/a	n/a	n/a	Cheng	Cheng	Cheng	Cheng	Cheng	Fine	Weinstein	Weinstein	Li	Weinstein	Dawlaty	Lander	Lander	Ma	Lin	Cho	Trevino	Trevino	Qi	Qi
48	n/a	n/a	n/a	n/a	n/a	Doudna	Doudna	Doudna	Doudna	Doudna	Yeh	Hwang	Hwang	Cho	Hwang	Kondo	Ma	Trevino	Dawlaty	Maeder	Shivalila	Fu	Lin	Weinstein	Fu
49	n/a	n/a	n/a	n/a	n/a	East	East	East	East	East	Li	Maeder	Maeder	Kim	Maeder	Li	Maeder	Maeder	Gootenberg	Tsai	Gootenberg	Lin	Fu	Lander	Trevino
50	n/a	n/a	n/a	n/a	n/a	Chang	Xiong	Chang	Chang	Xiong	Kim	Peterson	Peterson	Weinstein	Peterson	Lin	Trevino	Ma	Maeder	Trevino	Chen	Weinstein	Weinstein	Fu	Lin

Author is a founder

Author is not a founder

(CONTINUED FROM PREVIOUS PAGE)

	2012					2013					2014					2015					2016				
Ranking	Betweenness Centrality	Closeness Centrality	Degree Centrality	Eigenvector Centrality	Pagerank	Betweenness Centrality	Closeness Centrality	Degree Centrality	Eigenvector Centrality	Pagerank	Betweenness Centrality	Closeness Centrality	Degree Centrality	Eigenvector Centrality	Pagerank	Betweenness Centrality	Closeness Centrality	Degree Centrality	Eigenvector Centrality	Pagerank	Betweenness Centrality	Closeness Centrality	Degree Centrality	Eigenvector Centrality	Pagerank
51	n/a	n/a	n/a	n/a	n/a	Xiong	Zhu	Xiong	Xiong	Zhu	Weinstein	Ma	Ma	Hwang	Ma	Peterson	Tsai	Tsai	Kim	Gootenberg	Kim	Sharp	Maeder	Ma	Weinstein
52	n/a	n/a	n/a	n/a	n/a	Zhu	Chang	Zhu	Zhu	Chang	Trevino	Yeh	Yeh	Maeder	Yeh	Zhang	Gootenberg	Gootenberg	Tsai	Ma	Fine	Maeder	Cho	Sharp	Maeder
53	n/a	n/a	n/a	n/a	n/a	Gao	Bao	Gao	Bao	Bao	Ma	Kim	Cho	Peterson	Trevino	Hwang	Kim	Kim	Cho	Peterson	Huang	Dawlaty	Kim	Kim	Dawlaty
54	n/a	n/a	n/a	n/a	n/a	Sun	Cradick	Sun	Cradick	Cradick	Cho	Cho	Kim	O'Connor–Giles	Kim	Ueda	Cho	Cho	Huang	Kim	Lin	Cho	Dawlaty	Cho	Cho
55	n/a	n/a	n/a	n/a	n/a	Xi	Fine	Xi	Fine	Fine	Kim	Trevino	Trevino	Trevino	Cho	Kim	Peterson	Peterson	Lander	Cho	Peterson	Ma	Sharp	Maeder	Kim
56	n/a	n/a	n/a	n/a	n/a	Xu	Gao	Xu	Agarwala	Gao	Concordet	O'Connor–Giles	O'Connor–Giles	Harrison	O'Connor–Giles	Weinstein	Huang	Huang	Peterson	Hwang	Li	Kim	Ma	Huang	Root
57	n/a	n/a	n/a	n/a	n/a	Zhu	Sun	Zhu	Scott	Sun	O'Connor–Giles	Harrison	Harrison	Wildonger	Harrison	Wright	Hwang	Hwang	Zhang	Yeh	Root	Huang	Huang	Dawlaty	Sharp
58	n/a	n/a	n/a	n/a	n/a	Gilbert	Xi	Gilbert	Gratz	Xi	Harrison	Wildonger	Wildonger	Yeh	Wildonger	Cho	Yeh	Zhang	Guilinger	Zhang	Zhang	Root	Root	Liu	Peterson
59	n/a	n/a	n/a	n/a	n/a	Larson	Xu	Larson	Harrison	Xu	Wildonger	Kim	Kim	Huang	Kim	Jao	Zhang	Yeh	Liu	Huang	Dawlaty	Peterson	Peterson	Zhang	Ma
60	n/a	n/a	n/a	n/a	n/a	Lim	Zhu	Lim	O'Connor–Giles	Zhu	Auer	Huang	Huang	Kim	Huang	Wente	Shen	Inoue	Inoue	Inoue	Li	Zhang	Zhang	Peterson	Huang
61	n/a	n/a	n/a	n/a	n/a	Qi	Agarwala	Qi	Wildonger	Agarwala	Del Bene	Zhang	Zhang	Shen	Zhang	Yeh	Inoue	Lin	Lin	Lin	Trevino	Doench	Doench	Chen	Doench
62	n/a	n/a	n/a	n/a	n/a	Weissman	Scott	Weissman	Gao	Scott	Kim	Gootenberg	Gootenberg	Zhang	Gootenberg	Huang	Lin	Matoba	Matoba	Matoba	Doench	Shen	Wright	Inoue	Ebert
63	n/a	n/a	n/a	n/a	n/a	Gasiunas	Gratz	Gasiunas	Sun	Gratz	Gilbert	Inoue	Inoue	Gootenberg	Inoue	Trevino	Matoba	Shen	Shen	Wright	Ebert	Inoue	Ebert	Lin	Zhang
64	n/a	n/a	n/a	n/a	n/a	Siksnys	Harrison	Siksnys	Xi	Harrison	Qi	Lin	Lin	Inoue	Lin	Gootenberg	Kim	Kim	Kim	Kim	Sharp	Lin	Shen	Matoba	Hartenian
65	n/a	n/a	n/a	n/a	n/a	Bao	O'Connor–Giles	Bao	Xu	O'Connor–Giles	Weissman	Matoba	Matoba	Lin	Matoba	Ma	Guilinger	Sharp	Hwang	Shen	Yamamoto	Matoba	Inoue	Root	Sanjana
66	n/a	n/a	n/a	n/a	n/a	Cradick	Wildonger	Cradick	Zhu	Wildonger	Huang	Moosburner	Moosburner	Matoba	Moosburner	Basak	Liu	Guilinger	Sharp	Root	Sanjana	Liu	Lin	Guilinger	Shen
67	n/a	n/a	n/a	n/a	n/a	Fine	Konermann	Fine	Konermann	Konermann	Moosburner	Shen	Shen	Guilinger	Shen	Nithin	Sharp	Liu	Yeh	Sharp	Shen	Ebert	Matoba	Weissman	Weissman
68	n/a	n/a	n/a	n/a	n/a	Agarwala	Li	Agarwala	Bikard	Bikard	Lander	Guilinger	Guilinger	Liu	Shi	Shen	Root	Root	Moosburner	Qi	Cowan	Chen	Weissman	Gilbert	Inoue
69	n/a	n/a	n/a	n/a	n/a	Scott	Shalem	Scott	Li	Li	Duroure	Liu	Liu	Moosburner	Guilinger	Root	Qi	Qi	Qi	Guilinger	Weinstein	Hartenian	Gilbert	Shen	Lin
70	n/a	n/a	n/a	n/a	n/a	Bikard	Weinstein	Bikard	Shalem	Shalem	Shi	Shi	Shi	Pattanayak	Liu	Kim	Moosburner	Moosburner	Chen	Liu	Sakuma	Weissman	Hartenian	Gersbach	Matoba
71	n/a	n/a	n/a	n/a	n/a	Sampson	Bikard	Sampson	Weinstein	Weinstein	Zhang	Pattanayak	Pattanayak	Wang	Lander	Li	Gilbert	Chen	Pattanayak	Moosburner	Hwang	Gilbert	Chen	Lim	Gilbert
72	n/a	n/a	n/a	n/a	n/a	Weiss	Zhang	Weiss	Zhang	Gilbert	Gootenberg	Lander	Lander	Zhang	Kosuri	Qi	Weissman	Wang	O'Connor–Giles	Gilbert	Jao	Wright	Liu	Moosburner	Hwang
73	n/a	n/a	n/a	n/a	n/a	Konermann	Cummings	Konermann	Cummings	Larson	Inoue	Kosuri	Kosuri	Jenkins	Stranges	Sharp	Chen	Gilbert	Wang	Weissman	Wente	Sanjana	Sanjana	Doench	Chen
74	n/a	n/a	n/a	n/a	n/a	Llewellyn	Donohue	Llewellyn	Donohue	Lim	Lin	Stranges	Stranges	Shi	Gilbert	Gilbert	Wang	Weissman	Root	Chen	Kim	Hwang	Hwang	Pattanayak	Liu
75	n/a	n/a	n/a	n/a	n/a	Saroj	Hamm	Saroj	Hamm	Qi	Matoba	Wang	Wang	Lander	Qi	Weissman	Pattanayak	Pattanayak	Harrison	Wang	Hartenian	Guilinger	Lim	Ebert	Yeh
76	n/a	n/a	n/a	n/a	n/a	Tzeng	Nguyen	Tzeng	Nguyen	Weissman	De Cian	Zhang	Jenkins	Bikard	Weissman	Sakuma	O'Connor–Giles	Shi	Wildonger	Ebert	Kondo	Lim	Gersbach	Anderson	Kim
77	n/a	n/a	n/a	n/a	n/a	Gratz	Gilbert	Gratz	Gilbert	Zhang	Larson	Jenkins	Gilbert	Kosuri	Pattanayak	Yamamoto	Ebert	Ebert	Gilbert	Doench	Yeh	Yeh	Guilinger	Hartenian	Lim
78	n/a	n/a	n/a	n/a	n/a	Harrison	Larson	Harrison	Larson	Cummings	Lim	Gilbert	Qi	Stranges	Wang	Inoue	Doench	O'Connor–Giles	Weissman	Hartenian	Zon	Gersbach	Kim	Kim	Shi
79	n/a	n/a	n/a	n/a	n/a	O'Connor–Giles	Lim	O'Connor–Giles	Lim	Donohue	Barrangou	Qi	Weissman	Cheng	Jenkins	Lin	Hartenian	Doench	Shi	Shi	Gersbach	Pattanayak	Yeh	Hwang	Gersbach
80	n/a	n/a	n/a	n/a	n/a	Wildonger	Qi	Wildonger	Qi	Hamm	Arkin	Weissman	Zhang	Doudna	Bikard	Matoba	Shi	Hartenian	Anderson	Pattanayak	Ueda	Kim	Wang	Bae	Guilinger
81	n/a	n/a	n/a	n/a	n/a	Li	Weissman	Li	Weissman	Nguyen	Shen	Cheng	Cheng	East	Cheng	Wang	Harrison	Anderson	Kosuri	O'Connor–Giles	Wang	Li	Anderson	Li	Heckl
82	n/a	n/a	n/a	n/a	n/a	Shalem	Gasiunas	Shalem	Gasiunas	Gasiunas	Kosuri	Doudna	Doudna	Zhang	Doudna	Musunuru	Wildonger	Harrison	Stranges	Anderson	Sabatini	Anderson	Li	Sternberg	Mikkelsen
83	n/a	n/a	n/a	n/a	n/a	Weinstein	Siksnys	Weinstein	Siksnys	Siksnys	Stranges	East	East	Gratz	East	O'Connor–Giles	Anderson	Wildonger	Sternberg	Harrison	Wei	Moosburner	Moosburner	Barrangou	Wang
84	n/a	n/a	n/a	n/a	n/a	Zhang	Calarco	Zhang	Calarco	Calarco	Calarco	Bikard	Bikard	Gilbert	Zhang	Ebert	Kosuri	Kosuri	Cheng	Wildonger	Weissman	Wang	Pattanayak	Graham	Wang
85	n/a	n/a	n/a	n/a	n/a	Arkin	Colaiacovo	Arkin	Colaiacovo	Colaiacovo	Colaiacovo	Calarco	Calarco	Qi	Larson	Harrison	Stranges	Stranges	Doudna	Sanjana	Gilbert	Graham	Zhou	Sanjana	Sabatini
86	n/a	n/a	n/a	n/a	n/a	Calarco	Friedland	Calarco	Friedland	Friedland	Friedland	Colaiacovo	Colaiacovo	Weissman	Lim	Wildonger	Zhang	Sanjana	East	Kosuri	Ma	Bae	Graham	Yeh	Wei
87	n/a	n/a	n/a	n/a	n/a	Colaiacovo	Tzur	Colaiacovo	Tzur	Tzur	Tzur	Friedland	Friedland	Chang	Calarco	Chang	Sanjana	Sternberg	Doench	Stranges	Barrangou	Zhou	Bae	Kriz	Li
88	n/a	n/a	n/a	n/a	n/a	Friedland	Wang	Friedland	Gao	Wang	Sampson	Tzur	Tzur	Xiong	Colaiacovo	Barrangou	Zhou	Zhou	Hartenian	Sabatini	Chen	Wang	Sternberg	Wang	Anderson
89	n/a	n/a	n/a	n/a	n/a	Tzur	Gao	Tzur	Jiao	Gao	Weiss	Larson	Larson	Cummings	Friedland	Anderson	Graham	Graham	Ebert	Wang	Chen	Sabatini	Barrangou	Zhou	Moosburner
90	n/a	n/a	n/a	n/a	n/a	Cummings	Jiao	Cummings	Yu	Jiao	Jenkins	Lim	Lim	Donohue	Tzur	Xiong	Sternberg	Bikard	Sanjana	Wei	Bae	Wei	Shi	Kosuri	Pattanayak
91	n/a	n/a	n/a	n/a	n/a	Donohue	Yu	Donohue	Wang	Yu	Guilinger	Chang	Arkin	Zhu	Arkin	Chen	Cheng	Cheng	Zhou	Sternberg	Auer	Shi	Shi	Stranges	Bae
92	n/a	n/a	n/a	n/a	n/a	Hamm	Arkin	Hamm	Arkin	Arkin	Liu	Xiong	Chang	Calarco	Chang	Doench	Doudna	Doudna	Bikard	Arkin	Del Bene	Heckl	Wang	Shi	Zhou
93	n/a	n/a	n/a	n/a	n/a	Nguyen	Rios	Nguyen	Rios	Rios	Bikard	Arkin	Xiong	Colaiacovo	Xiong	Hartenian	East	East	Xue	Bikard	Lim	Kriz	Kriz	Xue	Barrangou
94	n/a	n/a	n/a	n/a	n/a	Wang	Cencic	Wang	Cencic	Cencic	Zhang	Zhu	Gratz	Friedland	Gratz	Moosburner	Li	Sabatini	Gersbach	Zhou	Shi	Mikkelsen	Sabatini	Yin	Graham
95	n/a	n/a	n/a	n/a	n/a	Rios	Dostie	Rios	Dostie	Dostie	Wang	Gratz	Zhang	Tzur	Zhu	Guilinger	Sabatini	Wang	Yin	Graham	Inoue	Sternberg	Wei	Zhang	Sternberg
96	n/a	n/a	n/a	n/a	n/a	Gao	Fraser	Gao	Fraser	Fraser	Pattanayak	Zhang	Zhu	Bassett	Zhang	Liu	Wang	Wei	Zhang	Cheng	Lin	Zetsche	Heckl	Doudna	Shi
97	n/a	n/a	n/a	n/a	n/a	Jiao	Malina	Jiao	Malina	Malina	Cheng	Wang	Cummings	Liu	Cummings	Zhu	Wei	Arkin	Li	Doudna	Matoba	Shi	Mikkelsen	Zetsche	Yamamoto
98	n/a	n/a	n/a	n/a	n/a	Yu	Mills	Yu	Mills	Mills	Doudna	Gao	Donohue	Ma	Donohue	Sabatini	Bikard	Zhang	Zhang	East	Chen	Barrangou	Zetsche	Koonin	Kriz
99	n/a	n/a	n/a	n/a	n/a	Moosburner	Paquet	Moosburner	Paquet	Paquet	East	Sun	Wang	Zhang	Wang	Wang	Arkin	Li	Graham	Zhang	Heckl	Yamamoto	Yamamoto	Makarova	Zetsche
100	n/a	n/a	n/a	n/a	n/a	Huang	Pelletier	Huang	Pelletier	Pelletier	Sternberg	Xi	Gao	Gao	Gao	Wei	Larson	Xue	Bae	Li	Mikkelsen	Zhang	Zhang	Cheng	Sakuma

Author is a founder

Author is not a founder

APPENDIX A-2 HEAT MAP OF TOP 100 AUTHORS HIGHLIGHTING FOUNDERS, RANKED BY NUMBER OF CITATIONS (2012-2016) (LAST NAME ONLY)

	2012	2013	2014	2015	2016
Ranking	Author	Author	Author	Author	Author
1	Gasiunas	Church	Zhang	Zhang	Doudna
2	Barrangou	Zhang	Charpentier	Doudna	Zhang
3	Horvath	Esvelt	Huang	Gersbach	Qi
4	Siksnys	Mali	Doudna	Kim	Gersbach
5	Jinek	Hsu	Church	Church	Kim
6	Chylinski	Ran	O'Connor-Giles	Charpentier	Barrangou
7	Fonfara	Scott	Wildonger	Huang	Joung
8	Hauer	Siksnys	Harrison	Joung	Hsu
9	Doudna	Gasiunas	Kim	Hsu	Yamamoto
10	Charpentier	Marraffini	Joung	Sakuma	Lim
11	n/a	Harrison	Hsu	Yamamoto	Huang
12	n/a	Gratz	Shen	Anderson	Charpentier
13	n/a	Agarwala	Jenkins	Barrangou	Bao
14	n/a	Wildonger	Jinek	Xue	Khalili
15	n/a	Doudna	Gersbach	O'Connor-Giles	Church
16	n/a	O'Connor-Giles	Esvelt	Yin	Concordet
17	n/a	Yan	Kabadi	Hilton	Sakuma
18	n/a	Fraser	Liu	Qi	Liu
19	n/a	Cencic	Sander	Shen	Hu
20	n/a	Paquet	Ma	Sternberg	Tsai
21	n/a	Dostie	Zhang	Bao	Shui
22	n/a	Pelletier	Mali	Wildonger	Wolfe
23	n/a	Mills	Scott	Harrison	Li
24	n/a	Aach	Cencic	Hu	Bortesi
25	n/a	Malina	Reyon	Khalili	Hilton
26	n/a	Schippers	Malina	Kabadi	Fischer
27	n/a	Wang	Pelletier	Marraffini	Sternberg
28	n/a	Jaenisch	Anders	Mou	Anderson
29	n/a	Shivalila	Musunuru	Scott	Xue
30	n/a	Yang	Bao	Shalem	Cradick
31	n/a	Cheng	Konermann	Graham	Musunuru
32	n/a	Qi	Chen	Hotta	Yin
33	n/a	Gilbert	Sharp	Musunuru	Qi
34	n/a	Weissman	Siksnys	Kennedy	Liu
35	n/a	Lin	Wieland	Cullen	Jinek
36	n/a	Zhang	Gaul	Cheng	Weissman
37	n/a	Konermann	Foerstemann	Xu	Gupta
38	n/a	Bao	Boettcher	Sun	Root
39	n/a	Fine	Hollmann	Ren	Schillberg
40	n/a	Cradick	Merk	Liu	La Russa
41	n/a	Friedland	Nitschko	Ni	Gilbert
42	n/a	Colaiacovo	Obermaier	Root	Smith
43	n/a	Tzur	Philippou-Massier	Reddy	Mahfouz
44	n/a	Calarco	Cradick	Jacks	Fine
45	n/a	Moosburner	Bassett	Konermann	Wang
46	n/a	Karvelis	Gasiunas	Pelletier	Hotta
47	n/a	Shi	Ran	Ran	Zhu
48	n/a	Norville	Li	Kildegaard	Farre
49	n/a	Jiang	Guo	Mali	Christou
50	n/a	Frokjaer-Jensen	Peng	Makarova	Kim

	2012	2013	2014	2015	2016
Ranking	Author	Author	Author	Author	Author
51	n/a	Wright	Xu	Liu	Xu
52	n/a	Morita	Ni	Koonin	Kaminski
53	n/a	Larson	Sun	Jenkins	Li
54	n/a	Lim	Yang	Sharp	Schaeffer
55	n/a	Horii	Liu	Esvelt	Nakata
56	n/a	Sasaki	Ren	Qiao	Maeder
57	n/a	Beekman	Cho	Wang	Toki
58	n/a	Nieuwenhuis	Mao	Kaminski	Bae
59	n/a	van der Ent	Ji	Jinek	Gaj
60	n/a	Cuppen	Kim	Li	O'Connor-Giles
61	n/a	Boymans	Marraffini	Dow	Lee
62	n/a	Demircan	Zhou	Kim	Bolukbasi
63	n/a	Heo	Zhang	Ramakrishna	Perrimon
64	n/a	Dekkers	Yeh	Wang	Ventura
65	n/a	Sasselli	Fu	Perrimon	Thakore
66	n/a	Koo	Chylinski	Birmingham	Kim
67	n/a	Schwank	Wang	Smith	Kim
68	n/a	Clevers	Dostie	Gootenberg	Beisel
69	n/a	Hatada	Liu	Kornepati	Shen
70	n/a	Tamura	Cheng	Siksnys	Peng
71	n/a	Kimura	Guilinger	Burgess	Guo
72	n/a	Grosshans	Jaenisch	Varshney	Villorbina
73	n/a	Jiao	Makarova	Lin	Zischewski
74	n/a	Gao	Koonin	Malina	Housden
75	n/a	Wu	Sampson	Sanjana	Perez
76	n/a	Nekrasov	Cowan	Kim	Capell
77	n/a	Kamoun	Weiss	Lee	Twyman
78	n/a	Lin	Lin	Gasiunas	Munoz
79	n/a	Chaparro-Garcia	Trevino	Crawford	Soto
80	n/a	Belhaj	Barrangou	Dostie	Medina
81	n/a	Katic	Yamamoto	Housden	Bassie
82	n/a	Yu	Kim	Zhang	Nadi
83	n/a	Tang	Sakuma	Ma	Forni
84	n/a	Wu	Gratz	Tsai	Jin
85	n/a	Blitz	Nemudryi	Cho	Lade
86	n/a	Biesinger	Valetdinova	Bassett	Hatada
87	n/a	Xie	Medvedev	Sander	Horii
88	n/a	Liang	Zakian	Riordan	Xiong
89	n/a	Wang	Auer	Ye	Zhao
90	n/a	Cho	Hu	Zhang	Berman
91	n/a	Yan	Del Bene	Heruth	Burgess
92	n/a	Bao	Perrimon	Bae	Varshney
93	n/a	Wang	Campbell	Jiang	Rajan
94	n/a	Bai	Wang	Chen	Prykhozhij
95	n/a	Li	Qiao	Maeder	Jaenisch
96	n/a	Li	Panetta	Ward	Scott
97	n/a	Reiner	Boxem	Kim	Mali
98	n/a	Ward	Waaijers	Reyon	Carroll
99	n/a	Dickinson	Ekker	Yeh	Koonin
100	n/a	Meyer	Clark	Fine	Makarova

Author is a founder
Author is not a founder

APPENDIX B STARTUP PARTICIPANT AUTHORS RANKED BY ORDERS OF DEGREE CENTRALITY IN BOTH AUTHOR CITATION NETWORKS & CO-AUTHORSHIP NETWORKS, RELATIVE TO EMERGING RESEARCH TOPICS IN ACTIVELY FINANCED BIOPHARMACEUTICAL INDUSTRY FIELDS IN 2014–2017

(…CONTINUED ON NEXT PAGE)

[Exosome]

Rank	Startup Participant Author	Order in Author Citation Network	Order in Co-authorship Network
1	Zhang, Bin	204	15
2	Zhou, Yu	207	187
3	Chen, Wei	301	165
4	Zhang, Jing	351	23
5	Xu, Bin	316	184
6	Johansson, Henrik J.	319	258
7	Huang, Wei	359	216
8	Iordanskiy, Sergey	343	244
9	Chen, Yun	462	7
10	Wang, Jing	407	229
11	Miller, Donald	436	286
12	Wang, Jin	414	371
13	Zhang, Yu	403	385
14	Qiu, Yu	248	510
15	Weissleder, Ralph	119	573
16	Chen, Xi	605	172
17	Lin, Pengnian	109	629
18	Yen, Yun	103	634
19	Wu, Jun	93	636
20	Zhang, Xu	544	383
21	Chen, Li	465	485
22	Filant, Justyna	393	575
23	Yan, Jun	641	329
24	Zhang, Xin	648	320
25	Steinmetz, Lars M.	780	127
26	Wang, Mei	552	610
27	Wang, Xiao	381	746
28	Wang, Hai	382	759
29	Liu, Yuan	829	251
30	Wang, Wei	930	202
31	Lotvall, Jan	1067	5
32	Allan, David S.	831	691
33	Liu, Yi	1036	344
34	Liu, Xia	600	1025
35	Niu, Xin	1068	569
36	Liu, Yuchen	1087	533
37	Wang, Jian	996	1028
39	Chen, Hua	527	1027
40	Chen, Yu	770	866
41	Lewis, James D.	596	1017
42	Li, Ning	1091	701
43	Littman, Dan R.	997	1091
44	Clark, Andrew G.	59	1560
45	Zengler, Karsten	1004	1215
46	Wu, Jing	707	1561
47	Lander, Eric S.	556	1623
48	Huang, Yong	1271	1245
49	Johnson, Timothy J.	767	1649
50	Bicalho, Rodrigo C.	1458	1126
51	Li, Xin	825	1693
52	de Vos, Willem	449	1837
53	Inouye, Michael	431	1937
54	Harris, R. Alan	1600	1187
55	Leff, Jonathan W.	727	1874
56	Chen, Jing	1883	881
57	Peterson, Scott N.	1733	1182
58	Quake, Stephen R.	1607	1369
59	DeSantis, Todd Z.	1264	1726
60	Liu, Li	2068	594
61	Kassam, Zain	1320	1727
62	Stein, Richard R.	608	2120
63	Huang, Yan	1942	1294
64	Redinbo, Matthew R.	1400	1879
65	Hou, Yong	1962	1321
66	Zhang, Xu	1786	1677
67	Tian, Gang	1354	2059
68	Chen, Li	1333	2076
69	Hao, Qin	2324	1015
70	Zhang, Hua	2340	1033
71	Chen, Peng	2522	531
72	Wu, Hao	885	2437
73	Hyoty, Heikki	2204	1570
74	Xu, Xin	2136	1666
75	Gill, Steven R.	2537	1188
76	Courtney, Michael	2451	1380
77	Morrow, Casey	2526	1486
78	Wu, Michael C.	2097	2156

[Microbiome]

Rank	Startup Participant Author	Order in Author Citation Network	Order in Co-authorship Network
1	Xavier, Ramnik J.	4	7
2	de Vos, Willem M.	62	14
3	Wang, Jian	66	20
4	Mazmanian, Sarkis K.	108	239
5	Mills, David A.	244	224
6	Walker, Alan W.	85	422
7	Li, Jun	352	265
8	Frank, Daniel N.	389	233
9	Yu, Jun	477	216
10	Li, Xiaoping	369	387
11	Leonard, Pierre	150	516
12	Yeoman, Carl J.	347	450
13	Bode, Lars	371	534
14	Oresic, Matej	311	575
15	Li, Yin	129	652
16	Xu, Wei	209	665
17	Wang, Yan	558	426
18	Knip, Mikael	149	725
19	Braun, Jonathan	641	390
20	Li, Jing	651	538
21	Finlay, B. Brett	857	61
22	Shi, Wenyuan	603	623
23	Morrow, Casey D.	719	511
24	Cox, Michael J.	160	873
35	Zhao, Hui	971	580
36	Honda, Kenya	1022	495
37	Liu, Xin	1065	416
38	Xu, Jian	883	747
79	Targan, Stephan R.	2899	867
80	Lam, Thomas Yuen Tung	2672	1503
81	Lin, John C.	2687	1650
82	Weiss, Paul S.	2404	2063
83	Martin, John	1119	2956
84	Nagler, Cathryn R.	636	3193
85	Liang, Chao	2953	1793
86	Zhang, Yan	3194	1363
87	Flavell, Richard A.	2645	2413
88	Quintana, Francisco J.	1828	3166
89	Tanzi, Rudolph E.	2654	2542
90	Andersen, Gary L.	3243	1831
91	Nielsen, Jens	780	3642
92	Li, Dong	2235	3007
93	Hu, Min	2305	3091
94	Miller, Andrew D.	2095	3266
95	Rodriguez, Juan M.	3489	1820
96	Kahn, C. Ronald	3564	1774
97	Chen, Xu	1633	3621
98	Huang, Hu	1583	3672
99	Sun, Yan	1859	3652
100	Kim, Sun A.	2266	3604
101	Pardoll, Drew M.	3461	2687
102	Vogelstein, Bert	3465	2685
103	Sidransky, David	2890	3312
104	Brown, Richard	2732	3579
105	Li, Peng	3014	3380
106	Kelly, William J.	3347	3078
107	Kane, Anne V.	3385	3071

(CONTINUED FROM PREVIOUS PAGE & CONTINUED ON NEXT PAGE)

[CRISPR]

Rank	Startup Participant Author	Order in Author Citation Network	Order in Co-authorship Network
1	Zhang, Feng	1	1
2	Doudna, Jennifer A.	2	9
3	Joung, J. Keith	5	30
4	Barrangou, Rodolphe	23	24
5	Li, Wei	43	7
6	Church, George M.	61	38
7	Jacks, Tyler	85	17
8	Zhou, Qi	92	32
9	Lander, Eric S.	3	158
10	Anderson, Daniel G.	48	162
11	Langer, Robert	110	168
12	Charpentier,	10	258
13	Sharp, Phillip A.	13	300
14	Zhang, Yu	306	29
15	Xu, Han	140	284
16	Wang, Hong	147	304
17	Wurst, Wolfgang	115	418
18	Wang, Yan	449	10
19	Wu, Jun	462	33
20	Yang, Hui	276	425
21	Collins, James J.	390	506
22	Wang, Jing	610	264
23	Cowan, Chad A.	113	659
24	Scott, David A.	4	670
25	Liu, Wei	708	16
26	Wang, Ying	264	655
27	Jiang, Yu	531	514
28	Porteus, Matthew H.	304	678
29	Nureki, Osamu	41	742
30	Zhang, Yan	632	428
31	Li, Hao	621	454
32	Lowe, Scott W.	229	749
33	Xavier, Ramnik J.	60	783
34	Wang, Fang	212	766
35	Wang, Xiao	855	169
36	Lim, Wendell A.	131	857
37	Bradley, Allan	700	536
38	Zhang, Jun	208	863
39	Wang, Zhen	272	853
40	Wang, Gang	497	781
41	Davidson, Alan R.	511	775
42	Hacohen, Nir	123	936
43	Keasling, Jay D.	580	758
44	Brown, Kevin R.	624	732
45	Durocher, Daniel	648	737
46	Ebert, Benjamin L.	14	988
47	Zhang, Yi	370	976
48	Wang, Xia	1056	44
49	Wang, Feng	988	388
50	Wang, Yong	1028	297
51	Church, George	324	1031
52	Li, Zhe	368	1021
53	Wang, Jianying	78	1091
54	Zhang, Lu	939	619
55	Yang, Yi	1064	377
56	Wang, Lu	246	1097
57	Zhang, Bin	1018	560
58	Zhang, Bo	1036	527
59	Yang, Luhan	605	997
60	Lu, Timothy K.	244	1145
61	Cox, David B. T.	450	1113
62	Zhang, Wei	1236	4
63	Yan, Wei	360	1184
64	Li, Mo	736	1006
65	Zhang, Jie	1202	395
66	Li, Ming	1215	389
67	Zhao, Hui	1004	854
68	Harrington, William F.	884	980
69	Zhang, Ying	1380	370
70	Kuehn, Ralf	108	1442
71	Li, Yilong	155	1444
72	Li, Kai	1384	501
73	Liu, Yong	419	1419
74	Huang, Yong	714	1338
75	Hornung, Veit	715	1356
76	Zhang, Li	1191	1028
77	Olson, Eric N.	209	1557
78	Li, Ning	1580	128
79	Wang, Xin	1260	962
80	Conklin, Bruce R.	900	1304
81	Liu, Bo	1019	1301
82	Clevers, Hans	1607	420
83	Liu, David R.	57	1687
84	Chen, Yu	1268	1131
85	Lin, Lin	1532	778
86	Li, Jun	1107	1312
87	Wei, Wei	1644	516
88	Gao, Guangping	588	1626
89	Marraffini, Luciano	109	1752
90	Sabatini, David M.	42	1768
91	Liu, Jin	591	1671
92	Jaenisch, Rudolf	84	1770
93	Asokan, Aravind	410	1735
94	Schlabach, Michael R.	1622	780
95	Stegmeier, Frank	1625	776
96	Li, Nan	1411	1129
97	Wang, Bin	876	1602
98	Chen, Yan	895	1597
99	Zhang, Xu	517	1806
100	Wang, Yi	1071	1568
101	May, Andrew P.	1108	1563
102	Ma, Jing	696	1803
103	Tang, Li	1558	1194
104	Wang, Ping	1139	1609
105	Li, Song	1724	991
106	Young, Richard A.	2037	495
107	Zhang, Rui	2079	383
108	Li, Xin	1983	787
109	Edwards, David	1822	1110
110	Liu, Yang	2135	225
111	Murray, Stephen A.	2120	414
112	Wang, Wen	1886	1149
113	Li, Hui	451	2166
114	Li, Amy	680	2153
115	Liu, Qi	1941	1185
116	Liu, Xin	2262	394
117	Sun, Jin	688	2178
118	Fu, Xin	2100	939
119	Bailey, Scott	1095	2026
120	Sidhu, Sachdev	874	2143
121	Chen, Li	139	2373
122	Zhang, Zhiying	792	2244
123	Roberts, Charles W.	1081	2138
124	Fraichard, Alexandre	1480	1903
125	Ha, Gavin	1085	2175
126	Naldini, Luigi	2280	918
127	Li, Liang	2325	949
128	Zhang, Min	832	2374
129	Ren, Bing	2127	1368
130	Zhang, Na	1996	1841
131	Li, Bin	1581	2208
132	Wang, Dan	1570	2227
133	Yu, Qian	2257	1566
134	Li, Xiaoping	2227	1616
135	Liang, Chao	2342	1474
136	Xu, Yong	1460	2369
137	Yan, Sen	1677	2223
138	Xu, Jian	2338	1532
139	Zhou, Hai	2146	1882
140	Doudna, Jennifer	2487	1673
141	Li, Ling	2122	2274
142	Gabriel, Richard	2530	2359

(CONTINUED FROM PREVIOUS PAGE)

[Cas9]

Rank	Start-up Participant Author	Order in Author Citation Network	Order in Co-authorship Network
1	Zhang, Feng	1	1
2	Doudna, Jennifer A.	2	15
3	Joung, J. Keith	5	33
4	Li, Wei	43	20
5	Church, George M.	65	18
6	Jacks, Tyler	76	23
7	Anderson, Daniel G.	39	150
8	Langer, Robert	93	151
9	Sharp, Phillip A.	9	266
10	Xu, Han	144	255
11	Zhang, Yu	293	35
12	Zhou, Qi	97	350
13	Wu, Jun	366	25
14	Wang, Yan	369	7
15	Charpentier,	15	390
16	Wang, Hong	139	375
17	Cowan, Chad A.	90	466
18	Scott, David A.	3	485
19	Wang, Ying	241	449
20	Nureki, Osamu	42	532
21	Porteus, Matthew H.	260	472
22	Lander, Eric S.	6	545
23	Wang, Jianying	63	564
24	Lowe, Scott W.	190	541
25	Liu, Wei	600	10
26	Wurst, Wolfgang	100	597
27	Jiang, Yu	494	388
28	Wang, Jing	531	340
29	Zhang, Jun	156	619
30	Hacohen, Nir	113	694
31	Wang, Gang	420	566
32	Yang, Hui	561	483
33	Bradley, Allan	653	353
34	Wang, Yong	642	377
35	Ebert, Benjamin L.	12	752
36	Keasling, Jay D.	609	558
37	Church, George	300	771
38	Wang, Fang	184	809
39	Wang, Lu	194	807
40	Li, Zhe	323	774
41	Durocher, Daniel	660	528
42	Lim, Wendell A.	185	825
43	Collins, James J.	433	738
44	Wang, Feng	845	346
45	Yan, Wei	314	865
46	Wang, Xia	931	36
47	Zhang, Lu	812	457
48	Kuehn, Ralf	87	937
49	Li, Hao	664	712
50	Liu, David R.	34	1001
51	Li, Mo	654	773
52	Clevers, Hans	868	537
53	Liu, Yong	349	960
54	Zhang, Bin	841	596
55	Zhang, Bo	981	395
56	Huang, Yong	576	893
57	Yang, Yi	1043	344
58	Yang, Luhan	864	765
59	Brown, Kevin R.	877	790
60	Liu, Jin	448	1098
61	Gao, Guangping	490	1081
62	Liu, Bo	872	832
63	Zhao, Hui	893	819
64	Wang, Xin	992	728
65	Olson, Eric N.	171	1223
66	Harrington, William F.	989	749
67	Wang, Xiao	1182	530
68	Zhang, Li	1036	779
69	Zhang, Xu	424	1241
70	Chen, Yan	797	1066
71	Wang, Bin	771	1085
72	Conklin, Bruce R.	682	1156
73	Zhang, Jie	1161	733
74	Ma, Jing	554	1265
75	Zhang, Ying	1364	330
76	Hornung, Veit	796	1186
77	May, Andrew P.	1012	1017
78	Wang, Ping	983	1063
79	Li, Nan	1220	824
80	Zhang, Yan	1072	1037
81	Wang, Yi	1077	1036
82	Zhang, Wei	1509	6
83	Barrangou, Rodolphe	82	1514
84	Jaenisch, Rudolf	96	1526
85	Wei, Wei	1496	391
86	Zhang, Zhiying	605	1519
87	Lin, Lin	1444	777
88	Chen, Li	123	1646
89	Sun, Jin	579	1572
90	Xavier, Ramnik J.	53	1678
91	Schlabach, Michael R.	1592	568
92	Stegmeier, Frank	1593	571
93	Li, Amy	727	1546
94	Li, Jian	829	1593
95	Zhang, Min	697	1668
96	Lu, Timothy K.	376	1776
97	Liu, Yang	1806	261
98	Li, Ning	1580	1008
99	Zhang, Rui	1858	342
100	Wang, Wen	1686	863
101	Fraichard, Alexandre	1302	1398
102	Li, Song	1626	1052
103	Sidhu, Sachdev	899	1712
104	Ha, Gavin	1147	1566
105	Fu, Xin	1807	724
106	Jin, Xin	434	1894
107	Li, Jun	913	1717
108	Roberts, Charles W.	1151	1579
109	Liu, Yu	1682	1059
110	Murray, Stephen A.	1960	362
111	Wang, Dan	1330	1598
112	Liang, Chao	1854	959
113	Li, Xiaoping	1839	1053
114	Yan, Sen	1480	1537
115	Zhang, Na	1739	1358
116	Li, Bin	1558	1561
117	Zhou, Hai	1895	1392
118	Li, Ling	1953	1635

[CAR-T]

Rank	Startup Participant	Order in Author	Order in Co-
1	June, Carl H.	1	1
2	Sadelain, Michel	3	7
3	Riviere, Isabelle	9	40
4	Levine, Bruce L.	57	6
5	Jensen, Michael C.	59	5
6	Brentjens, Renier J.	62	18
7	Zhao, Yangbing	70	23
8	Lee, Daniel W.	43	85
9	Liu, Hao	97	20
10	Li, Daniel	110	36
11	Turtle, Cameron J.	98	80
12	Chen, Xueyan	111	110
13	Barrett, David M.	161	42
14	Wang, Yao	174	41
15	Gill, Saar	200	52
16	Pule, Martin	207	81
17	Brentjens, Renier	14	226
18	Liu, Yang	202	116
19	Heimfeld, Shelly	115	229
20	Chew, Anne	88	256
21	Loew, Andreas	230	199
22	Zhou, Li	231	198
23	Zhang, Qing	269	308

[Zika]

Rank	Startup Participant	Order in Author	Order in Co-
1	Osorio, Jorge E.	52	74
2	Brooks, John T.	138	390
3	Schinazi, Raymond F.	199	564
4	Shi, Yi	525	411
5	Busch, Michael P.	675	171
6	Whitehead, Stephen S.	676	172
7	Zhang, Bo	705	199
8	Zhao, Hui	540	491
9	Yao, Bing	215	701
10	Wu, Hao	583	472
11	Li, Zhenfeng	562	682
12	Mercado, Noe B.	563	685
13	Busch, Michael	934	306
14	Zheng, Wei	724	818
15	Yazdy, Mahsa M.	851	833
16	Shi, Jian	792	905
17	Gregory, Christopher J.	761	933
18	Lu, Lu	836	1202
19	Thomas, Dana L.	1092	1059
20	Lima de Mendonca,	1102	1255

Note: Rank here is based on closeness to the point of origin.

APPENDIX C-1 DESCRIPTIVE STATISTICS OF VARIABLES IN CAS9 DATASET

Variables (n=19611)								
Target Variable	**Type of Variable**		**YES (=1)**	**NO (=0)**				
Participant	Categorical:		669	18942				
	Yes = 1, No = 0		3.41%	96.59%				
Exit	Categorical:		345	19266				
	Yes = 1, No = 0		1.76%	98.24%				
Explanatory Variable	**Type of Variable**	**Minimum**	**Maximum**	**Mean**	**Median**	**S.D.**	**Skewness**	**Kurtosis**
<<Original Variables>>								
Individual Factors								
(Paper Features)								
PUB	Continuous: integral	0	70	1.54	1.00	1.63	11.43	288.24
PAPER_CITED	Continuous: integral	0	9258	74.79	8.00	257.28	10.78	195.40
PAPER_CITING	Continuous: integral	0	1833	74.59	37.00	108.08	3.81	23.56
CORRESP_AUTH	Continuous: integral	0	42	1.03	0.00	2.77	4.56	27.73
FIRST_AUTH	Continuous: integral	0	7	0.20	0.00	0.50	3.53	19.24
CITATION_DEG_CENT	Continuous: numerical	0	0.47	0.01	0.00	0.01	8.46	131.37
CITATION_INDEG_CENT	Continuous: numerical	0	0.47	0.00	0.00	0.01	10.74	194.13
CITATION_OUTDEG_CENT	Continuous: numerical	0	0.09	0.00	0.00	0.01	3.76	22.78
COAUTH_DEG_CENT	Continuous: numerical	0	0.02	0.00	0.00	0.00	4.37	31.79
(Patent Features)			**YES (=1)**	**NO (=0)**				
IP_BINARY	Categorical:		1090	18521				
	Yes = 1, No = 0		5.56%	94.44%				
		Minimum	**Maximum**	**Mean**	**Median**	**S.D.**	**Skewness**	**Kurtosis**
IP_NUM	Continuous: integral	0	69	0.11	0	0.87	33.96	2153.43
PAPER_CITED_NUM_IN_IP	Continuous: integral	0	6455	4.04	0	72.03	48.31	3552.45
			YES (=1)	**NO (=0)**				
PAPER_CITED_BINARY_IN_IP	Categorical:		769	18842				
	Yes = 1, No = 0		3.92%	96.08%				
		Minimum	**Maximum**	**Mean**	**Median**	**S.D.**	**Skewness**	**Kurtosis**
IP_CITED	Continuous: integral	0	3659	1.36	0	33.01	80.71	8164.35
IP_CITING	Continuous: integral	0	3555	2.05	0	42.54	51.46	3494.21
Ecosystem Factors								
(Academic Organization Features)								
UNIV_SIZE	Continuous: integral	0	135606	18387.69	15658.00	22085.38	2.09	7.36
UNIV_RESEARCH	Continuous: numerical	0	99.10	34.66	26.30	36.05	0.51	-1.26
UNIV_INNOV	Continuous: numerical	0	99.00	12.45	0.00	22.58	2.04	3.39
(Nation Features)								
NATION_VC	Continuous: numerical	0	0.3773	0.17	0.07	0.16	0.22	-1.87
NATION_STARTUP	Continuous: numerical	0	99.96	88.12	91.23	8.28	-6.47	65.49
NATION_TURNOVER	Continuous: numerical	0	20.7	10.18	12.00	3.33	-1.38	1.17
<<Selected & Applied Variables in the Models>>[a,b]								
IP_NUM.c	Continuous: numerical	-0.11	68.89	0.00	-0.11	0.87	33.96	2153.43
PUB_MFP.c	Continuous: numerical	-0.85	3.41	0.00	-0.16	0.33	2.65	8.83
IP_CITED.c	Continuous: numerical	-1.36	3657.64	0.00	-1.36	33.01	80.71	8164.35
FIRST_AUTH_MFP.c	Continuous: numerical	-0.86	0.13	0.00	0.13	0.29	-1.83	1.40
CORRESP_AUTH.c	Continuous: numerical	-1.03	40.97	0.00	-1.03	2.77	4.56	27.73
CITATION_OUTDEG_CENT.c	Continuous: numerical	0.00	0.09	0.00	0.00	0.01	3.76	22.78
PAPER_CITED_NUM_IN_IP.c	Continuous: numerical	-4.04	6450.96	0.00	-4.04	72.03	48.31	3552.45
NATION_VC.c	Continuous: numerical	-0.17	0.21	0.00	-0.10	0.16	0.22	-1.87
NATION_STARTUP.c	Continuous: numerical	-0.17	99.79	87.95	91.06	8.28	-6.47	65.49
IP_BINARY	Categorical, Same as above IP_BINARY							
PAPER_CITED_BINARY_IN_IP	Categorical, Same as above PAPER_CITED_BINARY_IN_IP							
COAUTH_DEG_CENT.c	Continuous: numerical	0.00	0.02	0.00	0.00	0.00	4.37	31.79
NATION_TURNOVER.c	Continuous: numerical	-10.18	10.52	0.00	1.82	3.33	-1.38	1.17
UNIV_RESEARCH.c	Continuous: numerical	-34.66	64.44	0.00	-8.36	36.05	0.51	-1.26
FIRST_AUTH_MFP.c * NATION_STARTUP.c	Continuous: numerical	-80.81	12.73	-0.01	11.32	25.41	-1.84	1.47
IP_CITED.c * CORRESP_AUTH.c	Continuous: numerical	-1225.03	87690.18	6.68	1.40	632.99	135.72	18774.06
IP_NUM.c * IP_CITED.c	Continuous: numerical	-5.31	251966.34	21.80	0.15	1838.49	132.05	17999.17
CORRESP_AUTH.c * IP_BINARY	Continuous: numerical	-1.03	40.97	0.07	0.00	1.08	16.85	388.08
FIRST_AUTH_MFP.c * IP_BINARY	Continuous: numerical	-0.84	0.13	-0.01	0.00	0.10	-6.35	42.89
FIRST_AUTH_MFP.c * NATION_VC.c	Continuous: numerical	-0.16	0.14	0.00	-0.01	0.05	-0.46	1.97
FIRST_AUTH_MFP.c * CORRESP_AUTH.c	Continuous: numerical	-31.20	4.46	-0.04	-0.13	0.93	-10.41	200.78
NATION_STARTUP.c * FIRST_AUTH_MFP.c	Continuous: numerical	-80.81	12.73	-0.01	11.32	25.41	-1.84	1.47
NATION_VC.c * FIRST_AUTH_MFP.c	Continuous: numerical	-0.16	0.14	0.00	-0.01	0.05	-0.46	1.97
CORRESP_AUTH.c * CITATION_OUTDEG_CENT.c	Continuous: numerical	-0.10	3.51	0.00	0.00	0.04	47.77	3520.05
PUB_MFP.c * CITATION_OUTDEG_CENT.c	Continuous: numerical	-0.01	0.24	0.00	0.00	0.01	17.37	490.39
FIRST_AUTH_MFP.c * CITATION_OUTDEG_CENT.c	Continuous: numerical	-0.07	0.01	0.00	0.00	0.00	-8.36	114.97
COAUTH_DEG_CENT.c * IP_BINARY	Continuous: numerical	0.00	0.02	0.00	0.00	0.00	17.71	498.25
FIRST_AUTH_MFP.c * IP_BINARY	Continuous: numerical	-0.84	0.13	-0.01	0.00	0.10	-6.35	42.89
CORRESP_AUTH.c * IP_BINARY	Continuous: numerical	-1.03	40.97	0.07	0.00	1.08	16.85	388.08
CITATION_OUTDEG_CENT.c * PAPER_CITED_BINARY_IN_IP	Continuous: numerical	0.00	0.09	0.00	0.00	0.00	15.33	318.90
CORRESP_AUTH.c * PAPER_CITED_BINARY_IN_IP	Continuous: numerical	-1.03	40.97	0.05	0.00	0.96	21.43	601.58
CITATION_OUTDEG_CENT.c * IP_BINARY	Continuous: numerical	0.00	0.09	0.00	0.00	0.00	13.24	243.91
COAUTH_DEG_CENT.c * NATION_TURNOVER.c	Continuous: numerical	-0.03	0.02	0.00	0.00	0.00	-1.99	27.81
IP_BINARY * UNIV_RESEARCH.c	Continuous: numerical	-34.66	64.44	0.50	0.00	9.20	3.28	27.40
PUB_MFP.c * UNIV_RESEARCH.c	Continuous: numerical	-72.51	186.16	1.34	1.88	12.07	2.87	22.31
NATION_VC.c * IP_CITED.c	Continuous: numerical	-36.02	556.60	0.17	0.17	5.33	75.13	6938.90

a) "_MFP" indicates that these variables were turned into their multivariable fractional polynomial forms. Regarding Cas9, the same MFPs were applied both to Participant and Exit.

b) "_c" indicates that these variables were centered from their originals, i.e. adjusted so that their means became zero. This indication is omitted from the paper.

APPENDIX C-2 DESCRIPTIVE STATISTICS OF VARIABLES IN MICROBIOME DATASET

Variables (n=35932)								
Target Variable	**Type of Variable**		**YES (=1)**	**NO (=0)**				
Participant	Categorical:		1164	34768				
	Yes = 1, No = 0		3.24%	96.76%				
Exit	Categorical:		558	35374				
	Yes = 1, No = 0		1.55%	98.45%				
Explanatory Variable	**Type of Variable**	**Minimum**	**Maximum**	**Mean**	**Median**	**S.D.**	**Skewness**	**Kurtosis**
<<Original Variables>>								
Individual Factors								
(Paper Features)								
PUB	Continuous: integral	0	154	1.23	1.00	2.43	15.30	568.61
PAPER_CITED	Continuous: integral	0	6962	48.70	7.00	176.49	12.05	224.47
PAPER_CITING	Continuous: integral	0	3080	48.54	22.00	83.94	6.51	95.66
CORRESP_AUTH	Continuous: integral	0	94	1.01	0.00	4.20	13.34	243.12
FIRST_AUTH	Continuous: integral	0	13	0.24	0.00	0.56	3.87	30.17
CITATION_DEG_CENT	Continuous: numerical	0	0.24	0.00	0.00	0.01	10.11	182.66
CITATION_INDEG_CENT	Continuous: numerical	0	0.19	0.00	0.00	0.00	12.05	225.48
CITATION_OUTDEG_CENT	Continuous: numerical	0	0.08	0.00	0.00	0.00	6.38	92.96
COAUTH_DEG_CENT	Continuous: numerical	0	0.02	0.00	0.00	0.00	7.89	142.16
(Patent Features)			**YES (=1)**	**NO (=0)**				
IP_BINARY	Categorical:		198	35734				
	Yes = 1, No = 0		0.55%	99.45%				
		Minimum	**Maximum**	**Mean**	**Median**	**S.D.**	**Skewness**	**Kurtosis**
IP_NUM	Continuous: integral	0	81	0.01	0	0.74	106.39	11536.49
PAPER_CITED_NUM_IN_IP	Continuous: integral	0	323	0.19	0	5.26	39.45	1797.43
			YES (=1)	**NO (=0)**				
PAPER_CITED_BINARY_IN_IP	Categorical:		168	35764				
	Yes = 1, No = 0		0.47%	99.53%				
		Minimum	**Maximum**	**Mean**	**Median**	**S.D.**	**Skewness**	**Kurtosis**
IP_CITED	Continuous: integral	0	156	0.04	0	1.50	62.79	5199.82
IP_CITING	Continuous: integral	0	305	0.07	0	3.01	88.66	8676.76
Ecosystem Factors								
(Academic Organization Features)								
UNIV_SIZE	Continuous: integral	0	256470	15542.37	1987.00	20788.74	2.12	9.48
UNIV_RESEARCH	Continuous: numerical	0	99.10	25.87	8.10	31.38	0.86	-0.63
UNIV_INNOV	Continuous: numerical	0	100.00	9.45	0.00	21.09	2.52	5.62
(Nation Features)								
NATION_VC	Continuous: numerical	0	0.3773	0.17	0.05	0.16	0.26	-1.83
NATION_STARTUP	Continuous: numerical	0	99.96	88.92	91.23	10.47	-6.19	48.58
NATION_TURNOVER	Continuous: numerical	0	23.2	9.20	12.00	4.13	-0.81	-0.16
<<Selected & Applied Variables in the Models>>[a,b]								
CORRESP_AUTH_MFPp.c	Continuous: numerical	-0.15	7.70	0.00	-0.15	0.48	7.68	87.11
CITATION_OUTDEG_CENT.c	Continuous: numerical	0.00	0.08	0.00	0.00	0.00	6.38	92.96
PAPER_CITED_BINARY_IN_IP	Categorical, Same as above PAPER_CITED_BINARY_IN_IP							
FIRST_AUTH.c	Continuous: numerical	-0.24	12.76	0.00	-0.24	0.56	3.87	30.17
NATION_VC.c	Continuous: numerical	-0.17	0.21	0.00	-0.12	0.16	0.26	-1.83
PAPER_CITED_NUM_IN_IP.c	Continuous: numerical	-0.19	322.81	0.00	-0.19	5.26	39.45	1797.43
UNIV_INNOV.c	Continuous: numerical	-9.45	90.55	0.00	-9.45	21.09	2.52	5.62
UNIV_RESEARCH.c	Continuous: numerical	-25.87	73.23	0.00	-17.77	31.38	0.86	-0.63
UNIV_SIZE.c	Continuous: numerical	-15542.37	240927.63	0.00	-13555.37	20788.74	2.12	9.48
CORRESP_AUTH_MFPe.c	Continuous: numerical	-0.59	12.71	0.00	-0.59	1.27	3.43	17.87
NATION_VC_MFPe.c	Continuous: numerical	-0.37	0.58	0.00	0.02	0.37	0.34	-1.43
CITATION_INDEG_CENT.c	Continuous: numerical	0.00	0.19	0.00	0.00	0.00	12.05	225.48
COAUTH_DEG_CENT.c	Continuous: numerical	0.00	0.02	0.00	0.00	0.00	7.89	142.16
CORRESP_AUTH_MFPp.c * CITATION_OUTDEG_CENT.c	Continuous: numerical	-0.01	0.45	0.00	0.00	0.00	68.23	7204.52
FIRST_AUTH.c * IP_NUM.c	Continuous: numerical	-19.15	58.78	0.00	0.00	0.36	115.79	20588.65
CORRESP_AUTH_MFPp.c * PAPER_CITED_BINARY_IN_IP	Continuous: numerical	-0.15	6.85	0.00	0.00	0.07	60.44	4732.17
IP_NUM.c * IP_CITED_MFPp.c	Continuous: numerical	-739429.65	2937282.37	372.34	-473.04	28079.99	87.28	9070.42
CORRESP_AUTH_MFPp.c * UNIV_INNOV.c	Continuous: numerical	-72.77	288.46	0.50	1.38	10.50	10.60	199.54
FIRST_AUTH.c * UNIV_RESEARCH.c	Continuous: numerical	-252.56	753.01	1.16	3.42	18.93	5.67	119.11
NATION_VC.c * UNIV_RESEARCH.c	Continuous: numerical	-10.62	14.37	1.38	2.08	5.09	0.29	-0.23
CITATION_OUTDEG_CENT.c * UNIV_SIZE.c	Continuous: numerical	-392.52	2161.83	5.13	4.00	44.80	9.91	273.81
CORRESP_AUTH_MFPp.c * UNIV_SIZE.c	Continuous: numerical	-119727.85	280020.87	377.32	384.71	9104.52	4.87	141.53
UNIV_INNOV.c * UNIV_RESEARCH.c	Continuous: numerical	-1307.70	3751.84	265.39	244.36	573.79	2.35	7.17
PAPER_CITED_NUM_IN_IP.c * UNIV_SIZE.c	Continuous: numerical	-3571741.19	4193098.31	-164.38	3004.06	64980.36	-2.51	1670.66
CITATION_OUTDEG_CENT.c * FIRST_AUTH.c	Continuous: numerical	-0.01	0.20	0.00	0.00	0.00	24.85	893.02
CORRESP_AUTH_MFPe * CITATION_OUTDEG_CENT.c	Continuous: numerical	-0.02	0.96	0.00	0.00	0.01	35.93	2620.36
CORRESP_AUTH_MFPe * FIRST_AUTH.c	Continuous: numerical	-3.38	85.45	0.13	-0.24	1.85	14.33	384.06
FIRST_AUTH.c * NATION_VC_MFPe.c	Continuous: numerical	-4.71	2.78	-0.01	-0.03	0.20	-1.61	33.97
CITATION_OUTDEG_CENT.c * CITATION_INDEG_CENT.c	Continuous: numerical	0.00	0.02	0.00	0.00	0.00	103.78	14684.15
FIRST_AUTH.c * COAUTH_DEG_CENT.c	Continuous: numerical	0.00	0.06	0.00	0.00	0.00	46.15	3925.20
CITATION_INDEG_CENT.c * COAUTH_DEG_CENT.c	Continuous: numerical	0.00	0.00	0.00	0.00	0.00	89.33	11191.77
UNIV_INNOV.c * CITATION_INDEG_CENT.c	Continuous: numerical	-0.73	5.65	0.01	0.01	0.15	19.93	523.06
CORRESP_AUTH_MFPe * CITATION_INDEG_CENT.c	Continuous: numerical	-0.02	2.17	0.00	0.00	0.02	45.52	3396.24

a) "_MFP" indicates that these variables were turned into their multivariable fractional polynomial forms, with "p" for Participant and "e" for Exit.
b) ".c" indicates that these variables were centered from their originals, i.e. adjusted so that their means became zero. This indication is omitted from the paper.

APPENDIX C-3 DESCRIPTIVE STATISTICS OF VARIABLES IN 5-BIOPHARMA-TOPICS DATASET

(…CONTINUED ON NEXT PAGE)

Variables (n=94669)								
Target Variable	**Type of Variable**		**YES (=1)**	**NO (=0)**				
Participant	Categorical:		3156	91513				
	Yes = 1, No = 0		3.33%	96.67%				
Exit	Categorical:		1556	93113				
	Yes = 1, No = 0		1.64%	98.36%				
Explanatory Variable	**Type of Variable**	**Minimum**	**Maximum**	**Mean**	**Median**	**S.D.**	**Skewness**	**Kurtosis**
<<Original Variables>>								
Individual Factors								
(Paper Features)								
PUB	Continuous: integral	0	154	1.38	1	2.05	14.05	518.20
PAPER_CITED	Continuous: integral	0	11219	60.23	8	217.71	12.55	279.18
PAPER_CITING	Continuous: integral	0	3080	60.07	28	95.95	4.90	48.14
CORRESP_AUTH	Continuous: integral	0	244	1.23	0	5.22	16.56	447.20
FIRST_AUTH	Continuous: integral	0	13	0.21	0	0.53	3.93	30.46
CITATION_DEG_CENT	Continuous: numerical	0	0.64	0.01	0.00	0.02	12.68	238.99
CITATION_INDEG_CENT	Continuous: numerical	0	0.55	0.00	0.00	0.02	16.64	372.92
CITATION_OUTDEG_CENT	Continuous: numerical	0	0.23	0.00	0.00	0.01	7.73	93.75
COAUTH_DEG_CENT	Continuous: numerical	0	0.08	0.00	0.00	0.00	11.06	278.55
(Patent Features)			**YES (=1)**	**NO (=0)**				
IP_BINARY	Categorical:		2884	91785				
	Yes = 1, No = 0		3.05%	96.95%				
		Minimum	**Maximum**	**Mean**	**Median**	**S.D.**	**Skewness**	**Kurtosis**
IP_NUM	Continuous: integral	0	88	0.06	0	0.75	68.74	6913.10
PAPER_CITED_NUM_IN_IP	Continuous: integral	0	9631	2.03	0	51.85	99.57	15448.03
			YES (=1)	**NO (=0)**				
PAPER_CITED_BINARY_IN_IP	Categorical:		2133	92536				
	Yes = 1, No = 0		2.25%	97.75%				
		Minimum	**Maximum**	**Mean**	**Median**	**S.D.**	**Skewness**	**Kurtosis**
IP_CITED	Continuous: integral	0	4105	0.62	0	22.20	130.12	21158.53
IP_CITING	Continuous: integral	0	4460	0.99	0	27.44	89.93	11691.64
Hot Topic Factors								
FINANCED_AMOUNT	Continuous: numerical	23.94	32.17	26.70	27.51	2.59	0.65	-0.19
FINANCED_FREQ	Continuous: numerical	67.00	140.00	100.97	67.00	34.74	0.09	-1.92
KW_GROWTH	Continuous: numerical	2.20	5.32	3.68	4.52	1.24	0.11	-1.78
IP_GROWTH	Continuous: numerical	2.81	18.38	5.95	6.03	2.55	3.68	16.27
Ecosystem Factors								
(Academic Organization Features)								
UNIV_SIZE	Continuous: integral	0	256470	16834.81	11946	22122.01	2.32	9.64
UNIV_RESEARCH	Continuous: numerical	0	99.10	29.32	15.90	33.73	0.74	-0.89
UNIV_INNOV	Continuous: numerical	0	100.00	10.68	0.00	21.90	2.32	4.61
(Nation Features)								
NATION_VC	Continuous: numerical	0	0.38	0.16	0.04	0.16	0.28	-1.83
NATION_STARTUP	Continuous: numerical	0	99.96	88.41	91.23	9.66	-6.31	54.14
NATION_TURNOVER	Continuous: numerical	0	23.20	9.66	12.00	3.80	-1.04	0.32
<<Selected & Applied Variables in the Models>>[a,b]								
CORRESP_AUTH_MFPp.c	Continuous: numerical	-0.39	13.59	0.00	-0.06	0.53	8.20	106.84
IP_BINARY	Categorical, Same as above IP_BINARY							
IP_NUM.c	Continuous: numerical	-0.06	87.94	0.00	-0.06	0.75	68.74	6913.10
NATION_VC_MFPp.c	Continuous: numerical	-0.41	0.48	0.00	0.18	0.37	0.03	-1.74
IP_GROWTH.c	Continuous: numerical	-3.14	12.43	0.00	0.08	2.55	3.68	16.27
PUB_MFPe.c	Continuous: numerical	-0.46	0.54	0.00	-0.04	0.26	1.09	0.47
FINANCED_AMOUNT.c	Continuous: numerical	-2.76	5.47	0.00	0.82	2.59	0.65	-0.19
FIRST_AUTH_MFPp.c	Continuous: numerical	-0.84	0.09	0.00	0.09	0.20	-1.80	1.50
COAUTH_DEG_CENT_MFPp.c	Continuous: numerical	-1.70	4.18	0.00	-0.16	0.95	1.62	3.97
IP_CITED.c	Continuous: numerical	-0.62	4104.38	0.00	-0.62	22.20	130.12	21158.53
NATION_TURNOVER_MFPp.c	Continuous: numerical	-567.31	9532.08	0.00	-566.41	2321.09	3.86	12.92
NATION_STARTUP_MFPp.c	Continuous: numerical	-0.64	0.36	0.00	0.06	0.16	-0.82	1.28
UNIV_RESEARCH.c	Continuous: numerical	-29.32	69.78	0.00	-13.42	33.73	0.74	-0.89
IP_CITING.c	Continuous: numerical	-0.99	4459.01	0.00	-0.99	27.44	89.93	11691.64
UNIV_INNOV.c	Continuous: numerical	-10.68	89.32	0.00	-10.68	21.90	2.32	4.61
CORRESP_AUTH_MFPe.c	Continuous: numerical	-0.17	13.54	0.00	-0.17	0.56	7.29	85.78
NATION_VC.c	Continuous: numerical	-0.16	0.21	0.00	-0.13	0.16	0.28	-1.83
FIRST_AUTH_MFPe.c	Continuous: numerical	-0.86	0.14	0.00	0.14	0.30	-1.72	1.03
PUB_MFPp.c	Continuous: numerical	-0.46	0.54	0.00	-0.04	0.26	1.09	0.47
UNIV_SIZE.c	Continuous: numerical	-16834.81	239635.19	0.00	-4888.81	22122.01	2.32	9.64
CORRESP_AUTH_MFPp.c * IP_GROWTH.c	Continuous: numerical	-15.73	61.43	0.00	0.02	0.71	21.15	1359.75
CORRESP_AUTH_MFPp.c * FINANCED_AMOUNT.c	Continuous: numerical	-21.19	27.42	0.00	-0.05	1.01	-2.81	162.24
IP_BINARY * COAUTH_DEG_CENT_MFPp.c	Continuous: numerical	-1.70	4.18	-0.01	0.00	0.15	-0.91	102.90
IP_NUM * FIRST_AUTH_MFPp.c	Continuous: numerical	-50.43	7.55	-0.01	-0.01	0.28	-113.57	17835.49
COAUTH_DEG_CENT_MFPp.c * PAPER_CITED_BINARY_IN	Continuous: numerical	-1.70	4.18	-0.01	0.00	0.13	-1.51	135.16
IP_GROWTH.c * PAPER_CITED_NUM_IN_IP.c	Continuous: numerical	-2817.00	9370.66	1.15	1.03	58.74	76.42	9814.39
CORRESP_AUTH_MFPp.c * COAUTH_DEG_CENT_MFPp.c	Continuous: numerical	-19.84	18.11	-0.03	0.01	0.46	4.50	287.43
NATION_VC_MFPp.c * PUB_MFPp.c	Continuous: numerical	-0.22	0.26	0.01	0.00	0.09	0.13	1.42
PUB_MFPp.c * KW_GROWTH.c	Continuous: numerical	-0.80	0.88	-0.12	-0.03	0.29	-0.65	0.00
FIRST_AUTH_MFPp.c * IP_CITED.c	Continuous: numerical	-2353.62	162.06	-0.14	-0.06	10.93	-182.37	37133.10
CORRESP_AUTH_MFPp.c * PAPER_CITED_BINARY_IN_IP.c	Continuous: numerical	-0.39	13.59	0.00	0.00	0.14	33.13	1728.43
FIRST_AUTH_MFPp.c * PAPER_CITED_BINARY_IN_IP.c	Continuous: numerical	-0.84	0.09	0.00	0.00	0.05	-10.63	128.15
CORRESP_AUTH_MFPp.c * PAPER_CITED_NUM_IN_IP.c	Continuous: numerical	-666.01	63020.61	1.43	0.12	219.03	260.14	73099.17

(CONTINUED FROM PREVIOUS PAGE)

CORRESP_AUTH_MFPp.c * PUB_MFPp.c	Continuous: numerical	-6.26	4.14	-0.01	0.00	0.14	-3.15	254.51
CORRESP_AUTH_MFPp.c * IP_CITED.c	Continuous: numerical	-683.88	26862.74	0.58	0.04	95.45	252.52	68215.66
IP_GROWTH.c * PUB_MFPp.c	Continuous: numerical	-5.72	1.44	-0.12	0.00	0.53	-4.87	33.42
FINANCED_AMOUNT.c * NATION_TURNOVER_MFPp.c	Continuous: numerical	-26269.32	52136.26	-335.13	-461.12	6228.16	1.30	34.33
FIRST_AUTH_MFPp.c * NATION_STARTUP_MFPp.c	Continuous: numerical	-0.24	0.45	0.00	0.01	0.03	1.05	13.00
IP_CITED.c * NATION_STARTUP_MFPp.c	Continuous: numerical	-31.28	147.53	0.03	-0.03	1.11	70.71	6885.12
COAUTH_DEG_CENT_MFPp.c * NATION_STARTUP_MFPp.c	Continuous: numerical	-2.67	1.53	0.01	0.00	0.15	-1.79	42.76
NATION_TURNOVER_MFPp.c * NATION_STARTUP_MFPp.	Continuous: numerical	-6077.00	2832.83	2.53	-31.25	788.54	-3.19	32.67
NATION_VC_MFPp.c * NATION_STARTUP_MFPp.c	Continuous: numerical	-0.30	0.18	-0.03	-0.02	0.06	-1.30	4.79
IP_GROWTH.c * NATION_TURNOVER_MFPp.c	Continuous: numerical	-29902.86	118468.51	10.53	-44.60	6341.36	14.37	280.25
KW_GROWTH.c * NATION_TURNOVER_MFPp.c	Continuous: numerical	-14132.57	15621.14	-176.57	-472.04	2811.55	-0.52	15.62
IP_NUM.c * NATION_STARTUP_MFPp.c	Continuous: numerical	-3.22	4.47	0.00	0.00	0.06	6.85	1287.31
IP_BINARY * NATION_STARTUP_MFPp.c	Continuous: numerical	-0.64	0.30	0.00	0.00	0.02	-6.17	137.90
FIRST_AUTH_MFPp.c * KW_GROWTH.c	Continuous: numerical	-1.28	1.13	0.00	0.08	0.25	-0.19	2.37
IP_NUM.c * UNIV_RESEARCH.c	Continuous: numerical	-2373.20	5348.02	0.90	0.83	32.99	61.62	10907.89
IP_BINARY * UNIV_RESEARCH.c	Continuous: numerical	-29.32	69.78	0.39	0.00	6.91	5.98	60.57
IP_CITED.c * UNIV_RESEARCH.c	Continuous: numerical	-11211.53	249604.88	19.42	8.57	1234.86	155.63	28637.66
NATION_VC_MFPp.c * UNIV_RESEARCH.c	Continuous: numerical	-27.90	28.83	-3.93	-5.27	11.97	0.00	-0.62
COAUTH_DEG_CENT_MFPp.c * UNIV_RESEARCH.c	Continuous: numerical	-122.59	291.75	-3.33	-0.49	31.98	1.00	10.95
PUB_MFPp.c * UNIV_RESEARCH.c	Continuous: numerical	-31.57	37.55	-1.03	0.29	8.27	0.35	3.91
FINANCED_FREQ.c * UNIV_RESEARCH.c	Continuous: numerical	-2370.65	2723.26	-137.55	-192.96	1164.69	0.07	-0.79
FIRST_AUTH_MFPp.c * UNIV_RESEARCH.c	Continuous: numerical	-52.00	23.92	-0.19	-1.12	6.87	-1.50	6.89
NATION_VC_MFPp.c * FIRST_AUTH_MFPp.c	Continuous: numerical	-0.37	0.34	0.00	0.02	0.07	0.15	2.50
PAPER_CITED_NUM_IN_IP.c * UNIV_RESEARCH.c	Continuous: numerical	-59079.76	585578.91	64.73	27.43	2939.02	118.30	20180.64
PAPER_CITED_NUM_IN_IP.c * IP_CITING.c	Continuous: numerical	-86.32	42935655.13	1237.32	2.01	153402.23	247.30	66459.84
CORRESP_AUTH_MFPp.c * IP_CITING.c	Continuous: numerical	-348.95	29183.74	0.73	0.06	108.54	228.62	57770.27
IP_BINARY * FINANCED_AMOUNT.c	Continuous: numerical	-2.76	5.47	0.03	0.00	0.36	10.58	175.79
IP_GROWTH.c * UNIV_INNOV.c	Continuous: numerical	-280.21	1097.70	-0.32	1.21	51.43	7.34	147.48
PUB_MFPp.c * UNIV_INNOV.c	Continuous: numerical	-38.31	45.92	-0.35	0.41	5.88	2.62	24.56
IP_GROWTH.c * UNIV_RESEARCH.c	Continuous: numerical	-364.41	867.25	1.13	1.53	83.19	1.85	36.78
IP_CITING.c * UNIV_INNOV.c	Continuous: numerical	-22338.77	38695.64	-0.01	10.60	296.76	56.71	7619.50
PAPER_CITED_NUM_IN_IP.c * UNIV_INNOV.c	Continuous: numerical	-21516.61	129210.32	2.86	21.66	764.81	109.73	17595.00
CORRESP_AUTH_MFPp.c * IP_NUM.c	Continuous: numerical	-17.36	575.56	0.03	0.00	2.15	218.46	55562.13
NATION_STARTUP_MFPp.c * UNIV_INNOV.c	Continuous: numerical	-30.92	17.12	0.00	-0.26	3.45	-1.44	10.33
NATION_VC_MFPp.c * UNIV_INNOV.c	Continuous: numerical	-34.29	39.85	-0.79	-2.14	8.15	1.17	8.53
NATION_STARTUP_MFPp.c * UNIV_RESEARCH.c	Continuous: numerical	-23.53	18.69	1.18	0.86	4.62	0.54	1.86
COAUTH_DEG_CENT_MFPp.c * IP_CITING.c	Continuous: numerical	-6924.32	930.93	-0.84	0.15	35.02	-122.48	20438.60
IP_NUM.c * COAUTH_DEG_CENT_MFPp.c	Continuous: numerical	-136.56	24.50	-0.03	0.01	0.75	-100.19	16435.34
CORRESP_AUTH_MFPe.c * FINANCED_AMOUNT.c	Continuous: numerical	-21.15	27.56	0.00	-0.14	1.09	-2.02	123.54
CORRESP_AUTH_MFPe.c * KW_GROWTH.c	Continuous: numerical	-8.75	11.36	0.03	0.06	0.61	0.95	54.18
COAUTH_DEG_CENT.c * FINANCED_AMOUNT.c	Continuous: numerical	-0.05	0.06	0.00	0.00	0.00	4.19	76.00
IP_BINARY * FIRST_AUTH_MFPe.c	Continuous: numerical	-0.86	0.14	0.00	0.00	0.07	-8.65	83.27
CORRESP_AUTH_MFPe.c * IP_CITING.c	Continuous: numerical	-255.22	29181.35	0.78	0.17	109.01	226.78	56921.72
CORRESP_AUTH_MFPe.c * UNIV_SIZE.c	Continuous: numerical	#######	1294927.76	577.40	890.45	13231.62	22.13	1542.92
PUB_MFPe.c * UNIV_SIZE.c	Continuous: numerical	-96263.16	69956.58	-389.03	161.33	5648.99	1.88	32.51
FINANCED_AMOUNT.c * PUB_MFPe .c	Continuous: numerical	-2.51	2.94	-0.08	-0.03	0.84	1.14	4.67
IP_CITING.c * IP_NUM.c	Continuous: numerical	-8.87	392126.35	12.27	0.06	1441.25	235.49	60598.87
CORRESP_AUTH_MFPe.c * IP_NUM.c	Continuous: numerical	-14.10	575.51	0.03	0.01	2.16	216.00	54556.06
CORRESP_AUTH_MFPe.c * PAPER_CITED_BINARY_IN_IP.c	Continuous: numerical	-0.17	13.54	0.01	0.00	0.14	30.74	1492.67
FIRST_AUTH_MFPe.c * NATION_STARTUP_MFPe.c	Continuous: numerical	-0.29	0.53	0.00	0.01	0.05	1.11	12.60
FINANCED_AMOUNT.c * NATION_STARTUP_MFPe.c	Continuous: numerical	-3.49	2.00	-0.03	-0.01	0.41	0.00	6.71
COAUTH_DEG_CENT.c * NATION_STARTUP_MFPe.c	Continuous: numerical	-0.01	0.00	0.00	0.00	0.00	-7.61	196.87
NATION_VC.c * FIRST_AUTH_MFPe.c	Continuous: numerical	-0.18	0.14	0.00	-0.02	0.05	-0.56	1.71

a) "_MFP" indicates that these variables were turned into their multivariable fractional polynomial forms, with "p" for Participant and "e" for Exit.
b) ".c" indicates that these variables were centered from their originals, i.e. adjusted so that their means became zero. This indication is omitted from the paper.

(…CONTINUED ON NEXT PAGE)

Paper-related Features

PUB

Research Topic	Target Variable		n	Minimum	Maximum	Mean	Median	S.D.	Skewness	Kurtosis
Cas9	Participant	0	18942	0	49	1.50	1	1.45	9.09	177.98
		1	669	1	70	2.56	1	4.15	8.67	115.57
	Exit	0	19266	0	49	1.51	1	1.48	8.79	165.49
		1	345	1	70	3	1	5.25	7.86	84.50
	Participant & No Exit		324	1	20	2	1	2.47	3.37	14.26
Microbiome	Participant	0	34768	0	154	1.21	1	2.41	15.83	607.82
		1	1164	0	43	1.70	1	3.10	7.14	73.06
	Exit	0	35374	0	154	1.22	1	2.42	15.49	584.87
		1	558	0	43	1.63	1	3.08	8.69	102.86
	Participant & No Exit		606	0	38	2	1	3.12	5.77	47.25
Top5 Biopharmaceutical Topics	Participant	0	91513	0	154	1.36	1	1.94	14.22	582.70
		1	3156	0	85	2.17	1	4.03	8.98	124.92
	Exit	0	93113	0	154	1.37	1	1.97	13.86	547.41
		1	1556	0	85	2.32	1	4.75	9.07	116.13
	Participant & No Exit		1600	0	52	2.02	1	3.19	6.62	70.32

PAPER_CITED

Research Topic	Target Variable		n	Minimum	Maximum	Mean	Median	S.D.	Skewness	Kurtosis
Cas9	Participant	0	18942	0	5515	70.57	7.00	226.46	8.62	109.58
		1	669	0	9258	194.14	15.00	688.64	7.33	67.98
	Exit	0	19266	0	5515	72.03	7.00	233.87	8.81	114.66
		1	345	0	9258	228.49	17.00	828.35	6.78	55.11
	Participant & No Exit		324	0	5394	157.56	13.50	497.37	6.45	52.08
Microbiome	Participant	0	34768	0	6962	48.24	7.00	174.80	12.11	228.33
		1	1164	0	4049	62.54	9.00	220.79	10.49	148.27
	Exit	0	35374	0	6962	48.55	7.00	176.30	12.15	227.80
		1	558	0	2067	58.09	9.00	187.94	7.10	58.70
	Participant & No Exit		606	0	4049	66.64	10.00	247.30	11.45	162.12
Top5 Biopharmaceutical Topics	Participant	0	91513	0	6962	57.94	8.00	198.56	10.39	166.18
		1	3156	0	11219	126.74	12.00	523.58	10.31	143.54
	Exit	0	93113	0	6962	58.88	8.00	203.91	10.57	171.98
		1	1556	0	11219	140.97	11.00	623.92	10.02	126.98
	Participant & No Exit		1600	0	6547	112.90	12.00	402.44	8.33	90.88

PAPER_CITING

Research Topic	Target Variable		n	Minimum	Maximum	Mean	Median	S.D.	Skewness	Kurtosis
Cas9	Participant	0	18942	0	1476	73.57	37.00	105.23	3.62	20.31
		1	669	0	1833	103.36	46.00	167.56	4.24	26.59
	Exit	0	19266	0	1476	73.91	37.00	105.77	3.60	20.00
		1	345	0	1833	112.43	53.00	194.71	4.34	25.30
	Participant & No Exit		324	0	861	93.71	44.50	132.32	2.80	9.53
Microbiome	Participant	0	34768	0	3080	48.42	22.00	83.54	6.58	98.80
		1	1164	0	1095	52.15	23.00	94.84	4.92	34.40
	Exit	0	35374	0	3080	48.59	22.00	84.05	6.53	96.30
		1	558	0	809	45.74	22.50	76.38	4.67	31.67
	Participant & No Exit		606	0	1095	58.04	23.50	108.83	4.68	29.98
Top5 Biopharmaceutical Topics	Participant	0	91513	0	3080	59.39	28.00	93.74	4.69	43.49
		1	3156	0	2536	79.99	33.00	144.68	5.61	53.65
	Exit	0	93113	0	3080	59.70	28.00	94.49	4.70	43.41
		1	1556	0	2536	82.15	34.00	159.08	6.15	60.07
	Participant & No Exit		1600	0	1793	77.90	32.00	129.16	4.36	31.32

CORRESP_AUTH

Research Topic	Target Variable		n	Minimum	Maximum	Mean	Median	S.D.	Skewness	Kurtosis
Cas9	Participant	0	18942	0	36	0.99	0.00	2.68	4.42	24.79
		1	669	0	42	1.92	0.00	4.65	4.06	21.59
	Exit	0	19266	0	38	1.01	0.00	2.71	4.47	25.73
		1	345	0	42	2.15	0.00	5.09	3.83	18.93
	Participant & No Exit		324	0	38	1.68	0.00	4.14	4.25	24.17
Microbiome	Participant	0	34768	0	94	0.79	0.00	2.62	11.52	260.84
		1	1164	0	94	7.63	2.00	17.15	3.62	13.53
	Exit	0	35374	0	94	0.86	0.00	2.99	12.66	283.13
		1	558	0	94	10.93	2.00	21.62	2.71	6.63
	Participant & No Exit		606	0	94	4.58	1.00	10.75	5.37	36.04
Top5 Biopharmaceutical Topics	Participant	0	91513	0	189	1.04	0.00	3.67	12.45	350.27
		1	3156	0	244	6.63	1.00	19.91	5.88	42.22
	Exit	0	93113	0	189	1.09	0.00	4.02	14.29	416.34
		1	1556	0	244	9.38	1.00	24.94	4.66	25.93
	Participant & No Exit		1600	0	189	3.96	0.00	12.77	8.47	91.93

FIRST_AUTH

Research Topic	Target Variable		n	Minimum	Maximum	Mean	Median	S.D.	Skewness	Kurtosis
Cas9	Participant	0	18942	0	7	0.20	0.00	0.50	3.43	18.08
		1	669	0	6	0.21	0.00	0.65	4.55	26.38
	Exit	0	19266	0	7	0.20	0.00	0.50	3.45	18.29
		1	345	0	6	0.26	0.00	0.75	4.22	21.54
	Participant & No Exit		324	0	5	0.17	0.00	0.52	4.57	28.79
Microbiome	Participant	0	34768	0	13	0.24	0.00	0.56	3.79	28.59
		1	1164	0	10	0.22	0.00	0.61	5.69	63.31
	Exit	0	35374	0	13	0.24	0.00	0.56	3.88	30.42
		1	558	0	4	0.22	0.00	0.56	3.31	13.89
	Participant & No Exit		606	0	10	0.23	0.00	0.65	6.96	84.80
Top5 Biopharmaceutical Topics	Participant	0	91513	0	13	0.21	0.00	0.53	3.74	26.04
		1	3156	0	13	0.22	0.00	0.66	6.56	76.82
	Exit	0	93113	0	13	0.21	0.00	0.53	3.91	30.39
		1	1556	0	7	0.23	0.00	0.66	4.47	27.20
	Participant & No Exit		1600	0	13	0.20	0.00	0.66	8.64	127.10

(CONTINUED FROM PREVIOUS PAGE & CONTINUED ON NEXT PAGE)

COAUTH_DEG

Research Topic	Target Variable		n	Minimum	Maximum	Mean	Median	S.D.	Skewness	Kurtosis
Cas9	Participant	0	18942	0.0001	0.0127	0.0007	0.0005	0.0008	4.05	24.58
		1	669	0.0001	0.0162	0.0010	0.0006	0.0013	4.92	37.46
	Exit	0	35374	0.0001	0.0127	0.0007	0.0005	0.0008	4.02	24.21
		1	558	0.0001	0.0162	0.0011	0.0006	0.0016	4.60	29.35
	Participant & No Exit		324	0.0001	0.0068	0.0008	0.0006	0.0009	2.79	11.08
Microbiome	Participant	0	34768	0.0000	0.0205	0.0004	0.0003	0.0005	7.87	144.27
		1	1164	0.0000	0.0099	0.0004	0.0003	0.0006	7.91	97.97
	Exit	0	35374	0.0000	0.0205	0.0004	0.0003	0.0005	7.90	142.08
		1	558	0.0000	0.0033	0.0003	0.0002	0.0003	3.47	16.80
	Participant & No Exit		606	0.0000	0.0099	0.0004	0.0003	0.0007	7.23	73.21
Top5 Biopharmaceutical Topics	Participant	0	91513	0.0000	0.0517	0.0007	0.0004	0.0012	9.28	170.56
		1	3156	0.0000	0.0756	0.0010	0.0004	0.0026	12.66	263.81
	Exit	0	35374	0.0000	0.0756	0.0007	0.0004	0.0013	10.97	284.18
		1	558	0.0000	0.0410	0.0011	0.0004	0.0025	8.62	106.05
	Participant & No Exit		1600	0.0000	0.0756	0.0010	0.0004	0.0028	15.21	349.35

Patent-related Features

IP_NUM

Research Topic	Target Variable		n	Minimum	Maximum	Mean	Median	S.D.	Skewness	Kurtosis
Cas9	Participant	0	18942	0	26	0.09	0.00	0.60	15.79	402.70
		1	669	0	69	0.66	0.00	3.42	13.57	243.95
	Exit	0	35374	0	26	0.10	0.00	0.66	15.58	368.18
		1	558	0	69	0.75	0.00	4.29	12.58	186.91
	Participant & No Exit		324	0	21	0.56	0.00	2.14	6.14	43.96
Microbiome	Participant	0	34768	0	77	0.01	0.00	0.42	173.07	31425.49
		1	1164	0	81	0.17	0.00	3.36	23.84	569.89
	Exit	0	35374	0	81	0.01	0.00	0.74	105.70	11378.14
		1	558	0	2	0.03	0.00	0.18	7.84	67.43
	Participant & No Exit		606	0	81	0.31	0.00	4.66	17.18	294.64
Top5 Biopharmaceutical Topics	Participant	0	91513	0	77	0.05	0.00	0.49	56.16	7136.14
		1	3156	0	88	0.37	0.00	3.13	21.72	544.28
	Exit	0	35374	0	81	0.05	0.00	0.64	70.22	8031.71
		1	558	0	88	0.36	0.00	3.13	21.73	547.95
	Participant & No Exit		1600	0	81	0.38	0.00	3.14	21.70	540.11

PAPER_CITED_NUM_IN_IP

Research Topic	Target Variable		n	Minimum	Maximum	Mean	Median	S.D.	Skewness	Kurtosis
Cas9	Participant	0	18942	0	2048	2.93	0.00	47.39	29.35	1034.93
		1	669	0	6455	35.56	0.00	295.95	16.69	336.22
	Exit	0	35374	0	2048	3.30	0.00	50.55	26.93	867.65
		1	558	0	6455	45.71	0.00	388.40	13.88	216.08
	Participant & No Exit		324	0	1578	24.76	0.00	142.29	8.00	70.78
Microbiome	Participant	0	34768	0	323	0.18	0.00	5.13	41.95	2003.30
		1	1164	0	130	0.73	0.00	8.17	14.04	206.17
	Exit	0	35374	0	323	0.19	0.00	5.30	39.23	1775.19
		1	558	0	35	0.15	0.00	1.71	16.95	322.75
	Participant & No Exit		606	0	130	1.25	0.00	11.19	10.32	109.36
Top5 Biopharmaceutical Topics	Participant	0	91513	0	2154	1.53	0.00	31.49	39.78	1981.24
		1	3156	0	9631	16.41	0.00	227.36	32.17	1221.24
	Exit	0	35374	0	2154	1.70	0.00	33.27	36.80	1692.54
		1	558	0	9631	21.60	0.00	311.41	25.09	699.88
	Participant & No Exit		1600	0	1578	11.36	0.00	87.38	11.52	153.85

IP_CITED

Research Topic	Target Variable		n	Minimum	Maximum	Mean	Median	S.D.	Skewness	Kurtosis
Cas9	Participant	0	18942	0	1715	1.02	0.00	19.61	54.74	3988.86
		1	669	0	3659	11.24	0.00	144.86	23.95	598.25
	Exit	0	19266	0	1715	1.08	0.00	19.72	53.17	3835.77
		1	345	0	3659	17.18	0.00	200.20	17.50	314.48
	Participant & No Exit		324	0	264	4.90	0.00	25.17	7.30	59.72
Microbiome	Participant	0	34768	0	156	0.03	0.00	1.42	73.58	6717.25
		1	1164	0	44	0.32	0.00	3.13	10.93	125.10
	Exit	0	35374	0	156	0.04	0.00	1.50	64.01	5329.94
		1	558	0	19	0.18	0.00	1.71	9.83	97.17
	Participant & No Exit		606	0	44	0.46	0.00	4.01	9.18	83.99
Top5 Biopharmaceutical Topics	Participant	0	91513	0	1739	0.46	0.00	12.79	84.99	9589.77
		1	3156	0	4105	5.20	0.00	100.09	37.30	1451.90
	Exit	0	93113	0	1739	0.50	0.00	12.95	81.35	9003.81
		1	1556	0	4105	7.75	0.00	141.12	26.91	741.75
	Participant & No Exit		1600	0	377	2.71	0.00	19.69	12.00	174.53

IP_CITING

Research Topic	Target Variable		n	Minimum	Maximum	Mean	Median	S.D.	Skewness	Kurtosis
Cas9	Participant	0	18942	0	1466	1.36	0.00	24.78	36.88	1752.42
		1	669	0	3555	21.51	0.00	187.93	14.80	244.99
	Exit	0	19266	0	1466	1.60	0.00	28.09	33.89	1447.62
		1	345	0	3555	27.03	0.00	241.56	12.70	168.11
	Participant & No Exit		324	0	1444	15.62	0.00	104.05	10.11	119.25
Microbiome	Participant	0	34768	0	298	0.05	0.00	1.98	111.04	15426.40
		1	1164	0	305	0.71	0.00	12.73	23.49	558.17
	Exit	0	35374	0	305	0.07	0.00	3.03	88.15	8566.80
		1	558	0	11	0.11	0.00	0.92	9.22	91.13
	Participant & No Exit		606	0	305	1.26	0.00	17.61	16.95	289.22
Top5 Biopharmaceutical Topics	Participant	0	91513	0	1466	0.71	0.00	16.28	47.16	3034.11
		1	3156	0	4460	9.04	0.00	121.83	27.38	865.08
	Exit	0	93113	0	1466	0.81	0.00	17.84	44.17	2635.07
		1	1556	0	4460	11.66	0.00	163.33	22.32	535.86
	Participant & No Exit		1600	0	1444	6.50	0.00	57.69	15.11	293.20

(CONTINUED FROM PREVIOUS PAGE & CONTINUED ON NEXT PAGE)

Hot Topic Factors/Features

FINANCED_AMOUNT

Research Topic	Target Variable		n	Minimum	Maximum	Mean	Median	S.D.	Skewness	Kurtosis
Cas9	Participant	0	18942	27.51	27.51	27.51	27.51	0.00	NA	NA
		1	669	27.51	27.51	27.51	27.51	0.00	NA	NA
	Exit	0	19266	27.51	27.51	27.51	27.51	0.00	NA	NA
		1	345	27.51	27.51	27.51	27.51	0.00	NA	NA
	Participant & No Exit		324	27.51	27.51	27.51	27.51	0.00	NA	NA
Microbiome	Participant	0	34768	23.94	23.94	23.94	23.94	0.00	NA	NA
		1	1164	23.94	23.94	23.94	23.94	0.00	NA	NA
	Exit	0	35374	23.94	23.94	23.94	23.94	0.00	NA	NA
		1	558	23.94	23.94	23.94	23.94	0.00	NA	NA
	Participant & No Exit		606	23.94	23.94	23.94	23.94	0.00	NA	NA
Top5 Biopharmaceutical Topics	Participant	0	91513	23.94	32.17	26.70	27.51	2.59	0.65	-0.19
		1	3156	23.94	32.17	26.70	27.51	2.54	0.63	-0.12
	Exit	0	93113	23.94	32.17	26.69	27.51	2.59	0.65	-0.19
		1	1556	23.94	32.17	26.77	27.51	2.56	0.62	-0.17
	Participant & No Exit		1600	23.94	32.17	26.64	27.51	2.51	0.65	-0.07

FINANCED_FREQ

Research Topic	Target Variable		n	Minimum	Maximum	Mean	Median	S.D.	Skewness	Kurtosis
Cas9	Participant	0	18942	67	67	67.00	67.00	0.00	NA	NA
		1	669	67	67	67.00	67.00	0.00	NA	NA
	Exit	0	19266	67	67	67.00	67.00	0.00	NA	NA
		1	345	67	67	67.00	67.00	0.00	NA	NA
	Participant & No Exit		324	67	67	67.00	67.00	0.00	NA	NA
Microbiome	Participant	0	34768	140	140	140.00	140.00	0.00	NA	NA
		1	1164	140	140	140.00	140.00	0.00	NA	NA
	Exit	0	35374	140	140	140.00	140.00	0.00	NA	NA
		1	558	140	140	140.00	140.00	0.00	NA	NA
	Participant & No Exit		606	140	140	140.00	140.00	0.00	NA	NA
Top5 Biopharmaceutical Topics	Participant	0	91513	67	140	101.01	67.00	34.74	0.09	-1.92
		1	3156	67	140	99.84	67.00	34.74	0.15	-1.91
	Exit	0	93113	67	140	101.00	67.00	34.74	0.09	-1.92
		1	1556	67	140	99.46	67.00	34.60	0.17	-1.90
	Participant & No Exit		1600	67	140	100.22	67.00	34.88	0.14	-1.92

KW_GROWTH

Research Topic	Target Variable		n	Minimum	Maximum	Mean	Median	S.D.	Skewness	Kurtosis
Cas9	Participant	0	18942	5.32	5.32	5.32	5.32	0.00	NA	NA
		1	669	5.32	5.32	5.32	5.32	0.00	NA	NA
	Exit	0	19266	5.32	5.32	5.32	5.32	0.00	NA	NA
		1	345	5.32	5.32	5.32	5.32	0.00	NA	NA
	Participant & No Exit		324	5.32	5.32	5.32	5.32	0.00	NA	NA
Microbiome	Participant	0	34768	2.54	2.54	2.54	2.54	0.00	NA	NA
		1	1164	2.54	2.54	2.54	2.54	0.00	NA	NA
	Exit	0	35374	2.54	2.54	2.54	2.54	0.00	NA	NA
		1	558	2.54	2.54	2.54	2.54	0.00	NA	NA
	Participant & No Exit		606	2.54	2.54	2.54	2.54	0.00	NA	NA
Top5 Biopharmaceutical Topics	Participant	0	91513	2.20	5.32	3.68	4.52	1.24	0.11	-1.78
		1	3156	2.20	5.32	3.74	4.52	1.25	0.03	-1.80
	Exit	0	93113	2.20	5.32	3.68	4.52	1.24	0.11	-1.78
		1	1556	2.20	5.32	3.75	4.52	1.26	0.02	-1.80
	Participant & No Exit		1600	2.20	5.32	3.72	4.52	1.24	0.05	-1.80

IP_GROWTH

Research Topic	Target Variable		n	Minimum	Maximum	Mean	Median	S.D.	Skewness	Kurtosis
Cas9	Participant	0	18942	6.52	6.52	6.52	6.52	0.00	NA	NA
		1	669	6.52	6.52	6.52	6.52	0.00	NA	NA
	Exit	0	19266	6.52	6.52	6.52	6.52	0.00	NA	NA
		1	345	6.52	6.52	6.52	6.52	0.00	NA	NA
	Participant & No Exit		324	6.52	6.52	6.52	6.52	0.00	NA	NA
Microbiome	Participant	0	34768	5.44	5.44	5.44	5.44	0.00	Inf	NaN
		1	1164	5.44	5.44	5.44	5.44	0.00	NA	NA
	Exit	0	35374	5.44	5.44	5.44	5.44	0.00	Inf	NA
		1	558	5.44	5.44	5.44	5.44	0.00	NA	NA
	Participant & No Exit		606	5.44	5.44	5.44	5.44	0.00	NA	NA
Top5 Biopharmaceutical Topics	Participant	0	91513	2.81	18.38	5.93	6.03	2.52	3.70	16.66
		1	3156	2.81	18.38	6.30	6.03	3.22	3.01	8.98
	Exit	0	93113	2.81	18.38	5.94	6.03	2.54	3.69	16.44
		1	1556	2.81	18.38	6.26	6.03	3.17	3.04	9.36
	Participant & No Exit		1600	2.81	18.38	6.34	6.03	3.27	2.98	8.62

Environment Factors

(Academic Organization Features)

UNIV_SIZE

Research Topic	Target Variable		n	Minimum	Maximum	Mean	Median	S.D.	Skewness	Kurtosis
Cas9	Participant	0	18942	0	135606	18326.74	15658.00	22048.07	2.08	7.33
		1	669	0	135606	20113.51	18945.00	23067.02	2.18	7.78
	Exit	0	19266	0	135606	18359.50	15658.00	22057.90	2.08	7.33
		1	345	0	135606	19962.03	18433.00	23549.31	2.26	8.03
	Participant & No Exit		324	0	135606	20274.80	19890.00	22577.42	2.08	7.38
Microbiome	Participant	0	34768	0	256470	15533.33	0.00	20845.73	2.13	9.55
		1	1164	0	135606	15812.28	12083.00	19014.06	1.69	5.83
	Exit	0	35374	0	256470	15555.44	1987.00	20820.02	2.12	9.48
		1	558	0	135606	14714.05	7613.50	18697.75	2.00	8.16
	Participant & No Exit		606	0	135606	16823.53	14091.33	19260.79	1.43	4.03
Top5 Biopharmaceutical Topics	Participant	0	91513	0	256470	16780.42	11554.00	22087.98	2.31	9.68
		1	3156	0	135606	18411.87	15557.00	23034.75	2.34	8.65
	Exit	0	93113	0	256470	16815.54	11946.00	22098.31	2.31	9.65
		1	1556	0	135606	17988.22	14091.33	23475.80	2.46	9.13
	Participant & No Exit		1600	0	135606	18823.86	16932.67	22597.30	2.21	8.12

(CONTINUED FROM PREVIOUS PAGE)

UNIV_RESEARCH

Research Topic	Target Variable		n	Minimum	Maximum	Mean	Median	S.D.	Skewness	Kurtosis
Cas9	Participant	0	18942	0	99	34.52	25.30	36.02	0.52	-1.25
		1	669	0	99	38.72	34.80	36.78	0.35	-1.39
	Exit	0	19266	0	99	34.62	26.15	36.05	0.51	-1.26
		1	345	0	99	37.01	33.05	36.13	0.42	-1.31
	Participant & No Exit		324	0	99	40.55	36.90	37.43	0.26	-1.47
Microbiome	Participant	0	34768	0	99	25.74	0.00	31.33	0.86	-0.61
		1	1164	0	99	29.55	21.25	32.61	0.64	-1.00
	Exit	0	35374	0	99	25.85	5.40	31.38	0.86	-0.62
		1	558	0	99	27.26	15.20	31.68	0.76	-0.79
	Participant & No Exit		606	0	99	31.66	24.87	33.34	0.53	-1.15
Top5 Biopharmaceutical Topics	Participant	0	91513	0	99	29.19	15.80	33.68	0.75	-0.88
		1	3156	0	99	33.24	25.19	34.81	0.56	-1.13
	Exit	0	93113	0	99	29.29	15.80	33.72	0.74	-0.89
		1	1556	0	99	31.23	22.90	33.97	0.66	-0.96
	Participant & No Exit		1600	0	99	35.20	29.45	35.52	0.46	-1.27

UNIV_INNOV

Research Topic	Target Variable		n	Minimum	Maximum	Mean	Median	S.D.	Skewness	Kurtosis
Cas9	Participant	0	18942	0	99	12.40	0.00	22.57	2.04	3.42
		1	669	0	96	13.97	0.00	22.85	1.83	2.67
	Exit	0	19266	0	99	12.42	0.00	22.56	2.04	3.40
		1	345	0	96	14.40	0.00	23.54	1.87	2.87
	Participant & No Exit		324	0	94	13.50	0.00	22.13	1.75	2.26
Microbiome	Participant	0	34768	0	100	9.44	0.00	21.11	2.52	5.61
		1	1164	0	95	9.63	0.00	20.41	2.53	5.93
	Exit	0	35374	0	100	9.45	0.00	21.11	2.52	5.61
		1	558	0	95	8.98	0.00	19.65	2.58	6.38
	Participant & No Exit		606	0	94	10.23	0.00	21.07	2.47	5.49
Top5 Biopharmaceutical Topics	Participant	0	91513	0	100	10.64	0.00	21.90	2.32	4.63
		1	3156	0	96	11.66	0.00	22.07	2.19	4.15
	Exit	0	93113	0	100	10.66	0.00	21.90	2.32	4.61
		1	1556	0	96	11.59	0.00	22.32	2.22	4.31
	Participant & No Exit		1600	0	95	11.73	0.00	21.82	2.15	3.97

(Nation Features)

NATION_VC

Research Topic	Target Variable		n	Minimum	Maximum	Mean	Median	S.D.	Skewness	Kurtosis
Cas9	Participant	0	18942	0	0	0.17	0.04	0.16	0.23	-1.86
		1	669	0	0	0.20	0.30	0.17	-0.18	-1.88
	Exit	0	19266	0	0	0.17	0.05	0.16	0.22	-1.86
		1	345	0	0	0.21	0.36	0.17	-0.29	-1.85
	Participant & No Exit		324	0	0	0.19	0.19	0.16	-0.06	-1.90
Microbiome	Participant	0	34768	0	0	0.17	0.04	0.16	0.27	-1.83
		1	1164	0	0	0.19	0.16	0.16	0.02	-1.87
	Exit	0	35374	0	0	0.17	0.05	0.16	0.27	-1.83
		1	558	0	0	0.19	0.16	0.16	0.02	-1.88
	Participant & No Exit		606	0	0	0.19	0.16	0.16	0.03	-1.87
Top5 Biopharmaceutical Topics	Participant	0	91513	0	0	0.16	0.04	0.16	0.30	-1.82
		1	3156	0	0	0.19	0.22	0.16	-0.08	-1.90
	Exit	0	93113	0	0	0.16	0.04	0.16	0.29	-1.83
		1	1556	0	0	0.20	0.27	0.17	-0.13	-1.90
	Participant & No Exit		1600	0	0	0.19	0.16	0.16	-0.04	-1.90

NATION_STARTUP

Research Topic	Target Variable		n	Minimum	Maximum	Mean	Median	S.D.	Skewness	Kurtosis
Cas9	Participant	0	18942	0	100	88.15	91.23	8.15	-6.44	66.21
		1	669	0	98	87.40	91.23	11.40	-6.05	43.58
	Exit	0	19266	0	100	88.14	91.23	8.18	-6.45	66.07
		1	345	0	98	86.85	91.23	12.63	-5.63	35.93
	Participant & No Exit		324	0	98	87.99	91.23	9.91	-6.48	54.76
Microbiome	Participant	0	34768	0	100	88.96	91.23	10.21	-6.23	50.17
		1	1164	0	100	87.87	91.23	16.49	-4.66	21.95
	Exit	0	35374	0	100	88.93	91.23	10.38	-6.20	49.09
		1	558	0	100	88.24	91.23	15.04	-5.05	26.79
	Participant & No Exit		606	0	100	87.53	91.23	17.73	-4.34	18.57
Top5 Biopharmaceutical Topics	Participant	0	91513	0	100	88.43	91.23	9.52	-6.32	54.96
		1	3156	0	100	87.84	91.23	12.91	-5.58	35.14
	Exit	0	93113	0	100	88.42	91.23	9.60	-6.31	54.49
		1	1556	0	100	87.69	91.23	12.57	-5.66	36.64
	Participant & No Exit		1600	0	100	87.99	91.23	13.23	-5.51	33.78

NATION_TURNOVER

Research Topic	Target Variable		n	Minimum	Maximum	Mean	Median	S.D.	Skewness	Kurtosis
Cas9	Participant	0	18942	0	21	10.16	12.00	3.34	-1.36	1.10
		1	669	0	14	10.69	12.00	3.02	-2.18	4.14
	Exit	0	19266	0	21	10.16	12.00	3.34	-1.37	1.12
		1	345	0	13	10.98	12.00	2.85	-2.56	6.02
	Participant & No Exit		324	0	14	10.38	12.00	3.17	-1.85	2.82
Microbiome	Participant	0	34768	0	23	9.18	12.00	4.14	-0.79	-0.20
		1	1164	0	16	10.06	12.00	3.82	-1.70	1.81
	Exit	0	35374	0	23	9.19	12.00	4.14	-0.80	-0.18
		1	558	0	16	10.25	12.00	3.61	-1.81	2.44
	Participant & No Exit		606	0	16	9.89	12.00	4.00	-1.59	1.30
Top5 Biopharmaceutical Topics	Participant	0	91513	0	23	9.63	12.00	3.81	-1.02	0.27
		1	3156	0	16	10.51	12.00	3.32	-2.01	3.36
	Exit	0	93113	0	23	9.64	12.00	3.81	-1.03	0.29
		1	1556	0	16	10.78	12.00	3.08	-2.27	4.71
	Participant & No Exit		1600	0	16	10.26	12.00	3.51	-1.80	2.39

APPENDIX E-1 COEFFICIENTS CHANGE BY REMOVING POTENTIAL INFLUENTIAL OBSERVATIONS PER EACH RESEARCHER GROUP IN CAS9 DATASET

	Coefficients of Variables for Cas9 Participant [a,b] Analyzing Potential Influential Observations' Influence	Original	Except 19610	Except 19600	Except 19598	Except 19103	Except All Influentials	Except All Influentials Left/Original	If > 200% or <50%
1	IP_NUM.c	0.127	0.135	0.130	0.133	0.120	0.129	1.021	-
2	PUB_MFP.c	1.176	1.176	1.172	1.174	1.174	1.169	0.994	-
3	IP_CITED.c	-0.002	-0.001	-0.003	-0.001	-0.004	-0.003	1.641	-
4	FIRST_AUTH_MFP.c	0.759	0.759	0.765	0.758	0.759	0.765	1.008	-
5	CORRESP_AUTH.c	0.020	0.020	0.020	0.020	0.020	0.019	0.958	-
6	CITATION_OUTDEG_CENT.c	-36.922	-36.874	-35.847	-36.310	-35.613	-34.213	0.927	-
7	PAPER_CITED_NUM_IN_IP.c	0.001	0.001	0.001	0.001	0.001	0.001	1.106	-
8	NATION_VC.c	1.224	1.224	1.230	1.217	1.215	1.216	0.994	-
9	NATION_STARTUP.c	-0.012	-0.012	-0.012	-0.012	-0.012	-0.012	0.996	-
10	IP_BINARY	0.869	0.853	0.857	0.855	0.876	0.849	0.977	-
11	PAPER_CITED_BINARY_IN_IP	-0.419	-0.428	-0.405	-0.418	-0.419	-0.409	0.975	-
12	FIRST_AUTH_MFP.c:NATION_STARTUP.c	0.049	0.049	0.049	0.049	0.049	0.049	0.998	-
13	IP_CITED.c:CORRESP_AUTH.c	0.000	0.000	0.000	0.001	0.000	0.000	0.569	-
14	IP_NUM.c:IP_CITED.c	0.000	0.000	0.000	0.000	0.000	0.000	0.887	-
15	CORRESP_AUTH.c:IP_BINARY	0.035	0.035	0.037	0.035	0.036	0.040	1.145	-
16	FIRST_AUTH_MFP.c:IP_BINARY	1.433	1.437	1.443	1.443	1.492	1.509	1.052	-
17	FIRST_AUTH_MFP.c:NATION_VC.c	-3.452	-3.449	-3.508	-3.392	-3.402	-3.410	0.988	-
18	FIRST_AUTH_MFP.c:CORRESP_AUTH.c	-0.073	-0.073	-0.073	-0.076	-0.073	-0.076	1.046	-

a) "_MFP" indicates that these variables were turned into their multivariable fractional polynomial forms. Regarding Cas9, the same MFPs were applied both to Participant and Exit.
b) ".c" indicates that these variables were centered from their originals, i.e. adjusted so that their means became zero. This indication is omitted from the paper.

	Coefficients of Variables for Cas9 Exit [a,b] Analyzing Potential Influential Observations' Influence	Original	Except 19537	Except 17473	Except 19599	Except 19570	Except 17712	Except 19564	Except All Influentials	Except All Influentials Left/Original	If > 200% or <50%
1	IP_NUM.c	0.077	0.090	0.079	0.068	0.070	0.080	0.074	0.086	1.081	-
2	PUB_MFP.c	1.252	1.245	1.260	1.238	1.257	1.294	1.268	1.328	1.054	-
3	CORRESP_AUTH.c	0.010	0.009	0.009	0.009	0.010	0.010	0.011	0.007	0.774	-
4	COAUTH_DEG_CENT.c	-8.072	-19.889	-10.265	2.103	-7.829	-27.329	-4.373	-36.555	3.561	FLAG
5	NATION_VC.c	1.388	1.421	1.374	1.383	1.387	1.421	1.381	1.433	1.042	-
6	NATION_STARTUP.c	-0.019	-0.018	-0.019	-0.019	-0.019	-0.019	-0.018	-0.018	0.971	-
7	IP_CITED.c	0.004	-0.007	0.005	0.004	0.004	0.004	0.004	-0.009	-1.901	FLAG
8	FIRST_AUTH_MFP.c	0.914	0.925	0.877	0.901	0.915	0.910	0.871	0.842	0.960	-
9	CITATION_OUTDEG_CENT.c	-39.584	-41.654	-38.894	-42.147	-39.818	-44.763	-45.937	-53.115	1.366	-
10	IP_BINARY	0.215	0.195	0.197	0.225	0.220	0.183	0.228	0.169	0.858	-
11	PAPER_CITED_BINARY_IN_IP	0.303	0.394	0.338	0.307	0.304	0.293	0.282	0.420	1.244	-
12	NATION_TURNOVER.c	0.070	0.071	0.069	0.068	0.069	0.068	0.069	0.069	0.988	-
13	UNIV_RESEARCH.c	-0.003	-0.003	-0.003	-0.003	-0.003	-0.003	-0.003	-0.003	0.972	-
14	NATION_STARTUP.c:FIRST_AUTH_MFP.c	0.054	0.054	0.053	0.053	0.054	0.054	0.052	0.053	0.992	-
15	NATION_VC.c:FIRST_AUTH_MFP.c	-4.446	-4.716	-4.394	-4.393	-4.434	-4.376	-4.379	-4.503	1.025	-
16	CORRESP_AUTH.c:CITATION_OUTDEG_CENT.c	3.559	4.047	4.203	4.193	3.603	3.164	3.235	4.103	0.976	-
17	PUB_MFP.c:CITATION_OUTDEG_CENT.c	-36.029	-32.228	-34.033	-31.478	-36.282	-42.104	-30.848	-30.436	0.894	-
18	FIRST_AUTH_MFP.c:CITATION_OUTDEG_CENT.c	-79.220	-75.234	-63.401	-76.120	-79.603	-93.546	-64.287	-54.341	0.857	-
19	COAUTH_DEG_CENT.c:IP_BINARY	219.021	255.144	273.858	262.129	221.181	264.580	200.338	336.678	1.229	-
20	FIRST_AUTH_MFP.c:IP_BINARY	1.605	1.649	1.540	1.611	1.627	1.754	1.494	1.614	1.048	-
21	CORRESP_AUTH.c:IP_BINARY	0.061	0.060	0.058	0.059	0.061	0.065	0.063	0.060	1.028	-
22	CITATION_OUTDEG_CENT.c:PAPER_CITED_BINARY_IN_IP	-65.519	-73.109	-87.161	-63.866	-67.373	-64.010	-63.498	-92.547	1.062	-
23	CORRESP_AUTH.c:PAPER_CITED_BINARY_IN_IP	-0.075	-0.077	-0.067	-0.078	-0.075	-0.072	-0.072	-0.064	0.953	-
24	CITATION_OUTDEG_CENT.c:IP_BINARY	53.511	48.724	45.614	47.096	53.710	64.634	56.363	54.053	1.185	-
25	COAUTH_DEG_CENT.c:NATION_TURNOVER.c	50.650	56.993	46.398	38.986	49.917	39.249	51.322	41.603	0.897	-
26	IP_BINARY:UNIV_RESEARCH.c	-0.008	-0.009	-0.007	-0.008	-0.008	-0.008	-0.008	-0.009	1.268	-
27	PUB_MFP.c:UNIV_RESEARCH.c	0.009	0.008	0.010	0.009	0.009	0.009	0.009	0.010	1.002	-
28	NATION_VC.c:IP_CITED.c	-0.029	0.033	-0.031	-0.030	-0.025	-0.028	-0.029	0.045	-1.460	FLAG

a) "_MFP" indicates that these variables were turned into their multivariable fractional polynomial forms. Regarding Cas9, the same MFPs were applied both to Participant and Exit.
b) ".c" indicates that these variables were centered from their originals, i.e. adjusted so that their means became zero. This indication is omitted from the paper.

APPENDIX E-2 COEFFICIENTS CHANGE BY REMOVING POTENTIAL INFLUENTIAL OBSERVATIONS PER EACH RESEARCHER GROUP IN MICROBIOME DATASET

	Coefficients of Variables for Microbiome Participant [a,b] Analyzing Potential Influential Observations' Influence	Original	Except 34966	Except 35621	Excpt 35583	Except All Influentials	Except All Influentials Left/Original	If > 200% or <50%
1	CORRESP_AUTH_MFPp.c	1.275	1.275	1.275	1.275	1.275	1.001	-
2	CITATION_OUTDEG_CENT.c	-14.828	-14.242	-10.039	-14.385	-8.699	0.587	-
3	PAPER_CITED_BINARY_IN_IP	1.978	1.850	1.967	1.980	1.806	0.913	-
4	FIRST_AUTH.c	-0.178	-0.184	-0.177	-0.180	-0.184	1.032	-
5	NATION_VC.c	0.777	0.783	0.781	0.782	0.795	1.023	-
6	PAPER_CITED_NUM_IN_IP.c	-0.009	-0.008	-0.009	-0.009	-0.008	0.918	-
7	UNIV_INNOV.c	0.001	0.001	0.001	0.001	0.001	0.890	-
8	UNIV_RESEARCH.c	0.007	0.007	0.007	0.008	0.007	1.002	-
9	UNIV_SIZE.c	0.000	0.000	0.000	0.000	0.000	0.998	-
10	CORRESP_AUTH_MFPp.c:CITATION_OUTDEG_CENT.c	-27.432	-27.204	-43.729	-27.469	-44.237	1.613	-
11	FIRST_AUTH.c:IP_NUM.c	-0.069	0.152	-0.069	-0.098	0.133	-1.933	FLAG
12	CORRESP_AUTH_MFPp.c:PAPER_CITED_BINARY_IN_IP	-0.994	-1.049	-0.913	-1.015	-0.986	0.992	-
13	IP_NUM.c:IP_CITED_MFPp.c	0.000	0.000	0.000	0.000	0.000	2.181	FLAG
14	CORRESP_AUTH_MFPp.c:UNIV_INNOV.c	-0.006	-0.006	-0.005	-0.006	-0.006	1.004	-
15	FIRST_AUTH.c:UNIV_RESEARCH.c	-0.005	-0.005	-0.005	-0.005	-0.005	1.107	-
16	NATION_VC.c:UNIV_RESEARCH.c	-0.026	-0.026	-0.026	-0.026	-0.026	0.996	-
17	CITATION_OUTDEG_CENT.c:UNIV_SIZE.c	-0.002	-0.002	-0.002	-0.002	-0.002	0.883	-
18	CORRESP_AUTH_MFPp.c:UNIV_SIZE.c	0.000	0.000	0.000	0.000	0.000	1.048	-
19	UNIV_INNOV.c:UNIV_RESEARCH.c	0.000	0.000	0.000	0.000	0.000	1.004	-
20	PAPER_CITED_NUM_IN_IP.c:UNIV_SIZE.c	0.000	0.000	0.000	0.000	0.000	0.455	FLAG
21	CITATION_OUTDEG_CENT.c:FIRST_AUTH.c	10.325	9.085	13.978	10.406	13.119	1.271	-

a) "_MFPp" indicates that these variables were turned into their multivariable fractional polynomial forms.
b) ".c" indicates that these variables were centered from their originals, i.e. adjusted so that their means became zero. This indication is omitted from the paper.

	Coefficients of Variables for Microbiome Exit [a,b] Analyzing Potential Influential Observations' Influence	Original	Except 35857	Except 22244	Excpt 31755	Excpt 35847	Except All Influentials	Except All Influentials Left/Original	If > 200% or <50%
1	CORRESP_AUTH_MFPe.c	0.538	0.537	0.538	0.538	0.538	0.539	1.002	-
2	CITATION_OUTDEG_CENT.c	-26.992	-24.818	-26.627	-25.085	-25.519	-22.008	0.815	-
3	FIRST_AUTH.c	-0.086	-0.084	-0.095	-0.083	-0.092	-0.097	1.122	-
4	PAPER_CITED_BINARY_IN_IP	2.798	2.795	3.131	2.803	2.804	3.136	1.121	-
5	NATION_VC_MFPe.c	-0.667	-0.667	-0.676	-0.667	-0.663	-0.671	1.005	-
6	PAPER_CITED_NUM_IN_IP.c	-0.078	-0.077	-0.154	-0.079	-0.078	-0.154	1.970	-
7	UNIV_INNOV.c	-0.004	-0.004	-0.005	-0.004	-0.005	-0.005	1.063	-
8	CITATION_INDEG_CENT.c	24.647	24.284	24.779	22.080	24.381	21.969	0.891	-
9	COAUTH_DEG_CENT.c	-300.995	-298.293	-293.974	-298.642	-304.651	-298.171	0.991	-
10	CORRESP_AUTH_MFPe.c:CITATION_OUTDEG_CENT.c	-29.651	-30.334	-29.480	-31.439	-29.039	-30.759	1.037	-
11	CORRESP_AUTH_MFPe.c:FIRST_AUTH.c	-0.052	-0.053	-0.051	-0.053	-0.054	-0.055	1.061	-
12	CITATION_OUTDEG_CENT.c:FIRST_AUTH.c	-106.218	-92.741	-108.818	-108.594	-81.026	-77.343	0.728	-
13	FIRST_AUTH.c:NATION_VC_MFPe.c	0.544	0.541	0.534	0.541	0.584	0.573	1.053	-
14	CITATION_OUTDEG_CENT.c:CITATION_INDEG_CENT.c	5099.086	3177.862	5146.791	5982.262	5108.073	4612.479	0.905	-
15	FIRST_AUTH.c:COAUTH_DEG_CENT.c	686.972	649.067	696.411	715.059	561.219	575.288	0.837	-
16	CITATION_INDEG_CENT.c:COAUTH_DEG_CENT.c	-26835.897	-19139.567	-27193.646	-31285.556	-26641.593	-24636.885	0.918	-
17	UNIV_INNOV.c:CITATION_INDEG_CENT.c	1.128	1.063	1.138	0.956	1.102	0.870	0.771	-
18	CORRESP_AUTH_MFPe.c:CITATION_INDEG_CENT.c	-10.133	-10.660	-10.019	-8.531	-10.309	-8.924	0.881	-

a) "_MFPe" indicates that these variables were turned into their multivariable fractional polynomial forms.
b) ".c" indicates that these variables were centered from their originals, i.e. adjusted so that their means became zero. This indication is omitted from the paper.

APPENDIX E-3 COEFFICIENTS CHANGE BY REMOVING POTENTIAL INFLUENTIAL OBSERVATIONS PER EACH RESEARCHER GROUP IN 5-BIOPHARMA-TOPICS DATASET

(…CONTINUED ON NEXT PAGE)

	Coefficients of Variables for 5-Bio-Topics Participant [a,b] Analyzing Potential Influential Observations' Influence	Original	Except 94576	Except 94487	Except 94422	Except 94333	Except All Influentials	Except All Influentials Left/Original	If > 200% or <50%
1	CORRESP_AUTH_MFPp.c	0.891	0.933	0.871	0.878	0.889	0.877	0.984	-
2	IP_BINARY	0.800	0.811	0.801	0.806	0.806	0.822	1.027	-
3	IP_NUM.c	0.079	0.077	0.078	0.078	0.077	0.073	0.931	-
4	NATION_VC_MFPp.c	-0.433	-0.435	-0.433	-0.433	-0.433	-0.432	0.998	-
5	IP_GROWTH.c	0.017	0.017	0.017	0.017	0.017	0.017	0.987	-
6	PUB_MFPp.c	-0.934	-0.935	-0.934	-0.935	-0.935	-0.935	1.001	-
7	FINANCED_AMOUNT.c	-0.002	-0.002	-0.002	-0.002	-0.002	-0.002	0.970	-
8	FIRST_AUTH_MFPp.c	-0.215	-0.282	-0.229	-0.189	-0.185	-0.231	1.072	-
9	COAUTH_DEG_CENT_MFPp.c	-0.525	-0.506	-0.528	-0.527	-0.517	-0.510	0.971	-
10	IP_CITED.c	0.005	0.004	0.005	0.005	0.004	0.003	0.587	-
11	NATION_TURNOVER_MFPp.c	0.000	0.000	0.000	0.000	0.000	0.000	0.981	-
12	NATION_STARTUP_MFPp.c	0.951	0.942	0.954	0.946	0.945	0.938	0.986	-
13	UNIV_RESEARCH.c	0.002	0.002	0.002	0.002	0.002	0.002	1.003	-
14	IP_CITING.c	0.003	0.003	0.003	0.002	0.003	0.003	1.023	-
15	UNIV_INNOV.c	-0.003	-0.003	-0.003	-0.003	-0.003	-0.003	1.011	-
16	CORRESP_AUTH_MFPp.c:IP_GROWTH.c	-0.022	-0.022	-0.023	-0.023	-0.022	-0.023	1.020	-
17	CORRESP_AUTH_MFPp.c:FINANCED_AMOUNT.c	-0.088	-0.088	-0.088	-0.088	-0.088	-0.088	1.002	-
18	IP_BINARY:COAUTH_DEG_CENT_MFPp.c	-0.589	-0.578	-0.593	-0.594	-0.591	-0.589	1.000	-
19	IP_NUM.c:FIRST_AUTH_MFPp.c	0.172	0.188	0.161	0.187	0.175	0.193	1.123	-
20	COAUTH_DEG_CENT_MFPp.c:PAPER_CITED_BINARY_IN_IP	0.543	0.524	0.547	0.546	0.536	0.528	0.973	-
21	IP_GROWTH.c:PAPER_CITED_NUM_IN_IP.c	0.000	-0.001	-0.001	0.000	0.000	-0.001	1.128	-
22	CORRESP_AUTH_MFPp.c:COAUTH_DEG_CENT_MFPp.c	0.006	0.007	0.005	0.006	0.006	0.005	0.841	-
23	NATION_VC_MFPp.c:PUB_MFPp.c	0.539	0.539	0.537	0.540	0.538	0.534	0.990	-
24	PUB_MFPp.c:KW_GROWTH.c	-0.411	-0.409	-0.411	-0.412	-0.411	-0.411	1.000	-
25	FIRST_AUTH_MFPp.c:IP_CITED.c	-0.003	-0.006	-0.002	-0.002	-0.001	-0.003	1.130	-
26	CORRESP_AUTH_MFPp.c:PAPER_CITED_BINARY_IN_IP	-0.210	-0.251	-0.189	-0.197	-0.208	-0.195	0.927	-
27	FIRST_AUTH_MFPp.c:PAPER_CITED_BINARY_IN_IP	0.870	0.937	0.883	0.844	0.840	0.884	1.016	-
28	CORRESP_AUTH_MFPp.c:PAPER_CITED_NUM_IN_IP.c	-0.001	-0.001	0.000	-0.001	-0.001	0.000	0.339	FLAG
29	CORRESP_AUTH_MFPp.c:PUB_MFPp.c	0.482	0.484	0.481	0.482	0.483	0.483	1.002	-
30	CORRESP_AUTH_MFPp.c:IP_CITED.c	0.005	0.006	0.004	0.005	0.005	0.004	0.769	-
31	IP_GROWTH.c:PUB_MFPp.c	-0.061	-0.062	-0.061	-0.061	-0.061	-0.062	1.009	-
32	FINANCED_AMOUNT.c:NATION_TURNOVER_MFPp.c	0.000	0.000	0.000	0.000	0.000	0.000	0.998	-
33	FIRST_AUTH_MFPp.c:NATION_STARTUP_MFPp.c	5.195	5.224	5.175	5.198	5.184	5.180	0.997	-
34	IP_CITED.c:NATION_STARTUP_MFPp.c	-0.108	-0.107	-0.103	-0.115	-0.116	-0.116	1.078	-
35	COAUTH_DEG_CENT_MFPp.c:NATION_STARTUP_MFPp.c	0.927	0.931	0.926	0.927	0.927	0.927	0.999	-
36	NATION_TURNOVER_MFPp.c:NATION_STARTUP_MFPp.c	0.000	0.000	0.000	0.000	0.000	0.000	0.999	-
37	NATION_VC_MFPp.c:NATION_STARTUP_MFPp.c	-2.974	-2.951	-2.973	-2.970	-2.963	-2.935	0.987	-
38	IP_GROWTH.c:NATION_TURNOVER_MFPp.c	0.000	0.000	0.000	0.000	0.000	0.000	1.000	-
39	NATION_TURNOVER_MFPp.c:KW_GROWTH.c	0.000	0.000	0.000	0.000	0.000	0.000	1.000	-
40	IP_NUM.c:NATION_STARTUP_MFPp.c	0.859	0.849	0.844	0.866	0.864	0.849	0.989	-
41	IP_BINARY:NATION_STARTUP_MFPp.c	-1.034	-1.056	-1.019	-1.001	-1.005	-0.984	0.952	-
42	FIRST_AUTH_MFPp.c:KW_GROWTH.c	0.131	0.134	0.131	0.129	0.131	0.131	0.999	-
43	IP_NUM.c:UNIV_RESEARCH.c	0.003	0.002	0.002	0.003	0.002	0.002	0.947	-
44	IP_BINARY:UNIV_RESEARCH.c	-0.006	-0.006	-0.006	-0.006	-0.006	-0.006	0.993	-
45	IP_CITED.c:UNIV_RESEARCH.c	0.000	0.000	0.000	0.000	0.000	0.000	0.612	-
46	NATION_VC_MFPp.c:UNIV_RESEARCH.c	0.011	0.011	0.011	0.011	0.011	0.011	1.002	-
47	COAUTH_DEG_CENT_MFPp.c:UNIV_RESEARCH.c	-0.002	-0.002	-0.002	-0.002	-0.002	-0.002	1.006	-
48	PUB_MFPp.c:UNIV_RESEARCH.c	-0.008	-0.008	-0.008	-0.008	-0.008	-0.008	0.984	-
49	UNIV_RESEARCH.c:FINANCED_FREQ.c	0.000	0.000	0.000	0.000	0.000	0.000	1.003	-
50	FIRST_AUTH_MFPp.c:UNIV_RESEARCH.c	0.005	0.006	0.006	0.005	0.005	0.005	0.989	-
51	NATION_VC_MFPp.c:FIRST_AUTH_MFPp.c	1.134	1.137	1.131	1.127	1.127	1.118	0.985	-
52	UNIV_RESEARCH.c:PAPER_CITED_NUM_IN_IP.c	0.000	0.000	0.000	0.000	0.000	0.000	0.811	-
53	IP_CITING.c:PAPER_CITED_NUM_IN_IP.c	0.000	0.000	0.000	0.000	0.000	0.000	0.635	-
54	CORRESP_AUTH_MFPp.c:IP_CITING.c	0.002	0.002	0.003	0.003	0.002	0.003	1.465	-
55	IP_BINARY:FINANCED_AMOUNT.c	-0.050	-0.052	-0.048	-0.051	-0.051	-0.052	1.036	-
56	IP_GROWTH.c:UNIV_INNOV.c	-0.002	-0.002	-0.002	-0.002	-0.002	-0.002	1.013	-
57	PUB_MFPp.c:UNIV_INNOV.c	-0.012	-0.012	-0.012	-0.012	-0.012	-0.012	1.028	-
58	IP_GROWTH.c:UNIV_RESEARCH.c	0.000	0.000	0.000	0.000	0.000	0.000	1.016	-
59	IP_CITING.c:UNIV_INNOV.c	0.000	0.000	0.000	0.000	0.000	0.000	1.876	-
60	UNIV_INNOV.c:PAPER_CITED_NUM_IN_IP.c	0.000	0.000	0.000	0.000	0.000	0.000	2.327	FLAG
61	CORRESP_AUTH_MFPp.c:IP_NUM.c	-0.050	-0.049	-0.056	-0.053	-0.048	-0.060	1.207	-
62	NATION_STARTUP_MFPp.c:UNIV_INNOV.c	-0.036	-0.036	-0.036	-0.036	-0.036	-0.036	0.997	-
63	NATION_VC_MFPp.c:UNIV_INNOV.c	-0.009	-0.009	-0.009	-0.009	-0.009	-0.009	0.991	-
64	NATION_STARTUP_MFPp.c:UNIV_RESEARCH.c	0.013	0.013	0.013	0.013	0.013	0.013	1.010	-
65	COAUTH_DEG_CENT_MFPp.c:IP_CITING.c	0.003	0.003	0.003	0.003	0.003	0.003	1.108	-
66	IP_NUM.c:COAUTH_DEG_CENT_MFPp.c	-0.096	-0.096	-0.096	-0.095	-0.096	-0.098	1.021	-

a) "_MFPp" indicates that these variables were turned into their multivariable fractional polynomial forms. Regarding Cas9, the same MFPs were applied both to Participant and Exit.
b) ".c" indicates that these variables were centered from their originals, i.e. adjusted so that their means became zero. This indication is omitted from the paper.

(CONTINUED FROM PREVIOUS PAGE)

	Coefficients of Variables for 5-Bio-Topics Exit [a,b] Analyzing Potential Influential Observations' Influence	Original	Except 94640 (Only Influential)	Except Influential /Original	If > 200% or <50%
1	CORRESP_AUTH_MFPe.c	1.356	1.354	0.998	-
2	IP_CITING.c	0.002	0.002	1.146	-
3	NATION_VC.c	1.348	1.348	1.000	-
4	IP_BINARY	0.923	0.899	0.973	-
5	FIRST_AUTH_MFPe.c	0.442	0.442	1.000	-
6	FINANCED_AMOUNT.c	0.014	0.014	0.999	-
7	PUB_MFPe.c	-0.653	-0.652	0.998	-
8	UNIV_SIZE.c	0.000	0.000	0.997	-
9	IP_NUM.c	0.040	0.055	1.362	-
10	CORRESP_AUTH_MFPe.c:FINANCED_AMOUNT.c	0.039	0.039	1.001	-
11	CORRESP_AUTH_MFPe.c:KW_GROWTH.c	-0.316	-0.316	1.000	-
12	FINANCED_AMOUNT.c:COAUTH_DEG_CENT.c	50.840	50.749	0.998	-
13	IP_BINARY:FIRST_AUTH_MFPe.c	0.652	0.656	1.006	-
14	CORRESP_AUTH_MFPe.c:IP_CITING.c	0.001	0.001	1.019	-
15	CORRESP_AUTH_MFPe.c:UNIV_SIZE.c	0.000	0.000	0.999	-
16	PUB_MFPe.c:UNIV_SIZE.c	0.000	0.000	1.001	-
17	FINANCED_AMOUNT.c:PUB_MFPe.c	-0.043	-0.043	0.999	-
18	IP_CITING.c:IP_NUM.c	0.000	0.000	1.488	-
19	CORRESP_AUTH_MFPe.c:IP_NUM.c	0.054	0.050	0.941	-
20	CORRESP_AUTH_MFPe.c:PAPER_CITED_BINARY_IN_IP	-0.583	-0.580	0.996	-
21	FIRST_AUTH_MFPe.c:NATION_STARTUP_MFPe.c	4.096	4.098	1.001	-
22	FINANCED_AMOUNT.c:NATION_STARTUP_MFPe.c	-0.244	-0.244	1.000	-
23	COAUTH_DEG_CENT.c:NATION_STARTUP_MFPe.c	-254.432	-253.959	0.998	-
24	NATION_VC.c:FIRST_AUTH_MFPe.c	-1.964	-1.963	1.000	-

a) "_MFPe" indicates that these variables were turned into their multivariable fractional polynomial forms.
b) ".c" indicates that these variables were centered from their originals, i.e. adjusted so that their means became zero.
This indication is omitted from the paper.