HEC MONTRĒAL

Innovation dynamics in the Chinese pharmaceutical industry
-
A Data Science perspective

par

Charlotte Marie Vorreuther

Sciences de la gestion
(Option Affaires Internationales)

Mémoire présenté en vue de l'obtention
du grade de maîtrise ès sciences en gestion
(M. Sc.)

Août 2018

Résumé

Depuis 2015, la Chine est devenue le premier pays en termes du nombre de brevets soumis chaque année, devançant les États-Unis. Le gouvernement chinois joue un rôle important dans le renforcement de l'attractivité de l'économie chinoise aux innovations mondiales et à la réduction de la dépendance du pays à la R&D internationale. Ce mémoire a pour but d'explorer les dynamiques d'innovations en Chine à travers le cas de l'industrie pharmaceutique. Plusieurs champs de la littérature sont mobilisés : l'innovation et la croissance économique, l'internationalisation de la R&D, l'innovation ouverte et, finalement, le concept de transfert des connaissances. D'un point de vue plus méthodologiques, une nouvelle technique issue de la Science des données est utilisée: l'analyse de texte sur l'ensemble des brevets. Ce qui nous permet de caractériser les dynamiques d'innovations au sein de l'industrie pharmaceutique en Chine. La base de données collectée est constituée de 238,870 brevets pharmaceutiques du Derwent World Patent Index entre 1990 et 2017 et constitue une source de données non-structurées. Cette base de données est utilisée pour analyser les thématiques d'innovation dans l'industrie pharmaceutique chinoise à travers le temps, en utilisant une analyse Latent Dirichlet Allocation (LDA). Finalement, nous proposons l'utilisation d'une variable préalablement inaperçue, la similitude entre brevets, pour analyser les dynamiques entre différents thématiques de brevets à travers le temps. Cette similitude est mésuré par l'index de similitude de Jaccard. L'analyse du mémoire s'achève sur une utilisation du concept d'utilisation de la variable similitude en étude de dynamiques d'innovation à travers une perspective économétrique. Nous voulons montrer que la similitude à travers de thématiques de brevets permet d'expliquer le comportement de brevetage dans les années successives et constitue ainsi un nouvel aspect du processus d'innovation. En final, ce mémoire est né de la motivation de montrer comment des nouvelles techniques de sciences de données peuvent être utilisé pour décrire les dynamiques d'innovation d'une manière sans précédente. Dans ce but un protocole a été développé que d'autres équipes de recherche peuvent exploiter pour des études futures.

Notre analyse a révélé que l'innovation en Chine focalise sur les traitements de cancer, les aminos et leur fonctionnement, les maladies (p.ex. le diabète), les anticorps/immunologie, et finalement polynucléotides, confortant les résultats d'autres chercheurs et experts comme Ni et al. (2017), Zhang et Zhou (2017) et Shen (2010). Le thématique des anticorps/immunologie est un thème présent dans les brevets plus anciens, des années 1990 à 2000, tandis que les thèmes aminos, maladies et polynucléotides sont des thèmes qui figurent dans les innovations plutôt récentes, des années 2001 à 2017. L'analyse de la similitude entre les classifications des brevets montre que les innovations sont de moins en moins liées aux innovations précédentes. Finalement, l'analyse économétrique indique l'utilité d'un indicateur de similitude dans l'analyse des dynamiques d'innovations. Plus précisément, la similitude entre des différents thématiques de brevets semble avoir un impact à courte-terme sur le brevetage. En fait, la similitude d'il y a une année ou d'une à cinq années expliquait un brevetage plus élevé. Cela était spéculé d'être originaire d'innovation incrémentale. Par rapport à un impact à long-terme, les résultats étaient peu concluants.

Pour conclure, la contribution de ce mémoire est double : d'une part méthodologique, à travers l'utilisation de la Sciences des données appliquées aux sciences sociales pour le development d'un protocole, d'autre part thématique, de par la cartographie des dynamiques d'innovation en Chine au regard de l'industrie pharmaceutique ainsi que la mise en évidence d'un nouvel aspect dans l'ensemble des dynamiques d'innovation et du processus d'innovation.

Mots clés: Dynamiques d’innovation, Chine, industrie pharmaceutique, analyse de brevets, science de données

Abstract

Surpassing the US, China is holding the position as the country where most patent applications are filed since 2015. The Chinese government plays a vital role concerning the attractivity of the Chinese economy for global innovation and the reduction of the country's dependence on foreign R&D. While China has been gaining attention for its innovation capacities, research is limited. Thus, this thesis explores the innovation dynamics in China, through the case of the pharmaceutical industry. Several streams in the literature are consulted: innovation and economic growth, R&D internationalization, open innovation and knowledge transfers. With regard to the methodological approach, a new technique coming from Data Science is being employed: text mining on a patent dataset. The database consists of 238,870 pharmaceutical patents of the Derwent World Patent Index between 1990 and 2017; an ensemble of unstructured data. This database is leveraged for the analysis of innovation thematics in the Chinese pharmaceutical industry over time, at means of a Latent Dirichlet Allocation (LDA) analysis. Finally, we propose the usage of an until now unnoticed variable, the similarity of patents, for the analysis of dynamics across different patent thematics across time. The Jaccard similarity index measures this similarity, which represents knowledge flows across research fields. The thesis closes on a proof of concept of the utility of the similarity variable in a study on innovation dynamics adopting an econometric perspective. We would like to show that the similarity across patent thematics helps explain the patenting patterns in subsequent years, and thus provides a new aspect of the innovation process. Finally, this thesis originates from the motivation to show how new techniques of Data Science can be used for the description of innovation dynamics never previously seen. To this end, a protocol of the present methodologies was developed that researchers can exploit for further studies.

Our analysis revealed that innovation in China focused on cancer treatments, aminos and their functioning, diseases (i.e., diabetes), antibodies/immunology, and polynucleotides. Thus, confirming Ni et al. (2017), Zhang and Zhou (2017) as well as Shen (2010) descriptions. The topic of antibodies/immunology appeared in earlier patents from 1990 to 2000 while aminos and their functioning, general diseases and polynucleotides figured among recent innovations, from 2001 to 2017. The analysis of similarities across patent classifications revealed that innovations rely less and less on previous innovations the more time passes. Finally, an econometric analysis has indicated the usefulness of the indicator of similarity for the analysis of innovation dynamics. Indeed, the similarity between different patent thematics seems to have a short-term impact on patenting behavior, since a higher patent count is associated with the similarity of one year ago, but also the similarities of one to five years ago. Hence, synergies of patent thematics of the near past have an effect on the patenting. We hypothesize that this pattern may be due to incremental innovations. The results are inconclusive with regards to a long-term effect or the effect of similarities from the distant past.

In conclusion, the contribution of this thesis is twofold: on the one hand, methodological, through the application of Data Science for Social Sciences and the development of a protocol, and on the other hand, thematic, because of the description of innovation dynamics in the Chinese pharmaceutical industry as well as the identification of a new aspect of innovation dynamics and the innovation process at large.

Keywords: Innovation dynamics, China, pharmaceutical industry, patent analysis, Data Science

Acknowledgements

This thesis was written across three continents: Asia, Europe, and North-America.

Foremost, I would like to thank my thesis supervisor Thierry Warin for his enormous support and excellent guidance. He gave me the opportunity to develop a unique set of skills in Data Science and to apply these in a research context. I have been lucky to have an advisor who continually pushed me to strive for more and who encouraged me to pursue a thesis leveraging Data Science techniques. It has been a great experience. Thank you!

Equally, I would like to thank CIRANO (Centre Interuniversitaire de Recherche en Analyse des Organisations) for offering me an excellent working environment. It has been a pleasure to learn from so many inspiring people. Without them, this thesis would not have been possible. A big thank you to my colleagues for the constant encouragements and discussions.

Last but not least, I have several people to thank for their emotional support. Thank you to my parents, Reinhart and Marlies, and my brother Tilman, for encouraging me to pursue this second degree 6,000 km away from them. Additionally, I would like to thank my boyfriend, Max, for his support all along my studies, and my friends for their encouragements during the last months. Thank you also to all the students and professors who made my time at HEC Montréal worthwhile.

This report was written and produced entirely thanks to the Data Science platform developed by Professor Warin (www.nuance-r.com) using two coding languages R and Markdown, as well as data provided by CIRANO (Montreal)'s data platform www.mondo.international.

Contents

List of Figures

List of Tables

List of Abbreviations and Acronyms

AMNE	Advanced Market Multinational Enterprise
API	Active Pharmaceutical Ingredients
AU	Australia
BMI	Basic Medical Insurance
CDE	Center for Drug Evaluation
CFDA	China Food and Drug Administration
CN	China
CRAMS	Contract Research and Manufacturing services
CRO	Contract Research Organizations
DWPI	Derwent World Patents Index
EMNE	Emerging Market Multinational Enterprise
EPO/EP	European Patent Office
EU	European Union
EUIPO	European Union Intellectual Property Office
FDI	Foreign Direct Investment
GDP	Gross Domestic Product
GERD	Global Gross Expenditure in Research and Development
GMP	Good Manufacturing Practice for Drugs
HAC	Heteroskedasticity and Autocorrelation
HMO	Health Maintenance Organizations
ICH	International Council for Harmonisation of technical Requirements for Pharmaceuticals for Human Use
IPB	Intellectual Property Bureau
IPR	Intellectual Property Rights
JP	Japan
KIPO	Korean Intellectual Property Office
KR	South Korea
LDA	Latent Dirichlet Allocation
MNC	Multinational corporation
MNE	Multinational enterprise
NIC	Newly industrialized country
NIPS	National Intellectual Property Strategy
NIS	National Innovation System
NLP	Natural Language Processing
NPDS	National Patent Development Strategy
OCED	Organisation for Economic Co-operation and Development
OTC drugs	Over the counter drugs
PCT	Patent Cooperation Treaty
R and D	Research and development
SFDA	State Food and Drug Administration
SIN	System Integration and Networking models
SIPO	State Intellectual Property Office of the People's Republic China

TCM	Traditional Chinese Medicine
TRIPS	Agreement on Trade Related Aspects of Intellectual Property Rights
UK	United Kingdom
UNESCO	United Nations Educational, Scientific and Cultural Organization
US	United States of America
USPTO	United States Patent and Trademark Office
VEM	Variational Expected Maximization
WIPO	World Intellectual Property Organisation
WTO	World Trade Organisation

1 Introduction

Known as the world's manufacturing powerhouse, China has recently attracted a lot of attention for its innovation capabilities. Not only are Chinese firms increasingly listed in the Fortune Global 500 list, occupying 115 positions of 500 in 2016, but China also gained considerable attention for its innovation capabilities, when it surpassed the US in terms of patent applications in 2015 (Ge 2017; Jalfin 2017). Since then newspapers recurrently make China's innovation capabilities a subject of discussion (e.g., Ovide 2017). Academics have devoted limited attention to China's innovation, mostly describing the national innovation system of China or elaborating on China's recent patent surge. Thus, to our knowledge, innovation dynamics in China have not received a lot of attention yet, in particular from a quantitative perspective. Ultimately, this thesis aims to provide insights into the specificities of China's innovation in the pharmaceutical industry, more specifically its innovation dynamics through patents as a proxy. When looking at innovation dynamics, we deem it reasonable to choose an industry known for its innovativeness as unit of analysis. Thus the pharmaceutical industry was chosen. It is "the second most important sector in shifting R&D activities to China and India" (Bruche 2009b: p. 278). Moreover, it is the most research-intensive and innovative sector of manufacturing (Achilladelis and Antonakis 2001).

Ultimately the motivation of this thesis is the following: *How can we describe innovation dynamics in the pharmaceutical industry in China, leveraged by Data Science techniques?* So, what is Data Science and why do we use it? Data Science is a discipline that uses tools such as scientific methods, processes, algorithms, and systems to gather data, process it, and derive insights from it. At our times, accessibility and availability of data are growing and thus computational tools for the exploitation of data, are not only beneficial for firms, but also for Science. In particular, in Social Sciences there is a new *wave* of studies that "leverages advances in computational techniques for the inference in a degree of detail and complexity that was previously unthinkable", explain Marta Stelmaszak and Philipp Hukal from the London School of Economics (Stelmaszak and Hukal 2017: para. 7). While the use of computational tools is growing in the Social Sciences overall, business scholars are now on their way to leverage these new techniques. Thus, this thesis aims to highlight the benefits of Data Science techniques for researchers. It is thanks to Data Science techniques that we could analyze 238,870 patents related to pharmaceuticals between 1990 and 2017, using a very innovative quantitative methodological approach. Indeed, Data Science enables us to analyze innovation dynamics to an extent and level of detail, not seen yet. In fact, these computational techniques allow us to extract a significant amount of data from patents, which has not been possible before.

The patents analyzed are made available by the Derwent World Patent Index, of Clarivate Analytics, covering over 34 million patent families and more than 71 million patent documents worldwide from more than 40 patent offices. Quantitative in nature, the study develops a protocol for how Data Science techniques can be leveraged to capture innovation dynamics through a patent analysis in China. The patent analysis of 238,870 patents – wherein 69,923 patents filed among others in China – allows identifying China's emergence as an innovator and its specificities regarding patent thematics. This part of the analysis is covered by the analysis of unstructured data extracted through text mining from 479 html files containing patents in their entirety of a size of around 7 MB each.

Three questions have been developed that capture different aspects of innovation dynamics and lead to the identification of the usefulness of a new variable: the similarity between patent thematics. The first question who will reveal the patent thematics is: *Where lies the focus of innovation in the Chinese pharmaceutical industry? For the whole time span of 1990 to 2017? For 1990 to 2000? For 2001 to 2017?* This question will be answered in the first step of our protocol, whilst classifying the Chinese patents through an unsupervised learning method, an LDA analysis, applied on the patent abstracts. The resulting classification is one derived out of the data, eliminating potential human-inaccuracy that can result from inconsistency from a human-made classification process. Ultimately, we will compare the computationally developed classification and the existing classification into thematics from the Derwent World Patent Index. This step constitutes a proof of concept insofar as it reveals the potential of a computational classification that researchers could rely on for further research. It extends the knowledge we have on specificities of pharmaceutical innovation in China from expert opinions through an empirical perspective.

This thesis proposes the use of a new factor for innovation dynamics, the similarity between patent thematics. We propose that knowledge flows across thematics and research fields are of importance when analyzing innovation dynamics. Thus, the second question examines: *What kind of dynamics do we find between patent thematics/classes? More precisely: What is the relationship of patent classes across time?* The similarity between patent thematics represented by patent classes will be analyzed through time, leveraging an automated analysis of text similarity, the Jaccard similarity measurement, developed by Paul Jaccard in 1902 (Jaccard 1902). Measuring similarities across patent thematics captures knowledge flows across research fields, an aspect of innovation dynamics that has not receives attention yet. Overall, this step will allow capturing the innovation dynamics of an enormous number of patents that we would otherwise not be able to read. We will assess how much a patent class relies on other patent classes, and hence other research fields. And most of all, the analysis allows to identify whether this reliance is dependent on time and how so. This question provides a first insight into the usefulness of the variable similarity.

Finally, the analysis will rely on an econometric perspective to assess the impact of the similarity across patent thematic classes on the patent count in subsequent years. It investigates the following question: *What is the relationship between the similarity of patent classes and the patent count when taking into account the time lag?* Thereby, it allows to identify how long it takes for synergies across patent thematics to result in new patents. It is revealing an interesting new aspect of innovation dynamics, and thereby constitutes a proof of concept, highlighting the potential explanatory power of the variable similarity in the innovation process.

The thesis is built on the following chapters: Chapter 2 describes China's innovation landscape and provides a comprehensive overview of the pharmaceutical industry. It reveals recent improvements to China's innovation landscape and developments in the pharmaceutical industry that foster the development of China's innovative capacities. Chapter 3 starts off with a systematic literature review on innovation in the pharmaceutical industry, highlighting the relevance of our topic for research in the respective industry. Furthermore, it reveals the streams of literature this thesis has to touch upon. Thus, chapter 3 explains why China is interested in developing its innovation capacities, retrieving new models of innovation and trends in R&D internationalization that explain how China became part of a global innovation network. Moreover, it explains how knowledge can flow between firms, R&D units and external entities. Chapter 4 highlights the literature gap that this thesis aims to fill, revealing a lack of studies on innovation in developing countries, in particular China, and the necessity of the variable similarity that essentially represents knowledge flows on the modular level. We will elaborate on the methodological approach, the data description and each step of the protocol in Chapter 5. Afterward, the thesis outlines the results (Chapter 6), the discussion (Chapter 7) as well as the conclusion (Chapter 8).

Finally, the contribution of this thesis is mainly twofold. On the one hand, the thesis demonstrates a novel approach of leveraging Data Science techniques for the development of a protocol on the analysis of innovation, particularly for the analysis of the dynamics of innovation in the pharmaceutical industry. On the other hand, we outline new insights into the innovation capabilities of China and propose a new aspect of innovation dynamics, the similarity between patent thematics. All in all, it contributes to the academic literature on innovation management, knowledge pipelines, and R&D internationalization. Finally, it caters to managers decision-making process with regards to innovation management.

2 Panorama: Innovation in China and the pharmaceutical industry

The following chapter aims to present a comprehensive overview of China's innovation infrastructure and the pharmaceutical industry. On the one hand, the regulatory landscape that touches upon innovation in China will be presented, with a particular emphasis on intellectual property rights due to the unit of analysis of the present thesis. Moreover, classic indicators for China's innovation capacities will be highlighted, that have received attention from scholars and the media likewise. This part will highlight the relevance of not only our research focus on China, but also of the patent proxy. On the other hand, the panorama will describe the pharmaceutical industry, its functioning, its drivers and the specific case of the Chinese industry. It reveals the pharmaceutical industry's global character and changing structure that renders the industry so interesting for research on innovation dynamics in China.

2.1 Innovation in China

2.1.1 Innovation infrastructure in China

2.1.1.1 Policies

China has been known as a manufacturing place throughout the last three decades, with an acceleration since its WTO entry. Its economic growth and cheap labor have attracted many investments. However, despite its growing role as an economic superpower, it is faced with rising costs and "[…]stringent environmental regulations[…]" that make the "[…] manufacturing sector less competitive than in countries paying lower wages and offering less environmental protection" ("UNESCO Science Report: Towards 2030" 2016: p. 621). While GDP per capita has constantly been rising during the last decade, the growth rate has declined since 2007, from 14.2% in 2007 to 7.4% in 2014. Thus, China is faced with several challenges compelling it to change its economic development model from a "[…] labour-, investment-, energy-, and resource-intensive one into one that is increasingly dependent upon technology and innovation" ("UNESCO Science Report: Towards 2030" 2016: p. 621). As a consequence, in recent years, China has been receiving increasing attention from scholars and the public due to its innovation capacities. The governmental efforts, however, go back two decades. In 1995, the Chinese government introduced the *Revitalizing the Nation through Science and Education* strategy, which affirmed that China attempted to grow economically by developing their knowledge capabilities and resources (Gassmann, Beckenbauer, and Friesike 2012).

Initially, China focused on attracting foreign companies' business. Several measures were established as a means of encouraging foreign R&D investment. For instance, the approval process for FDI was liberalized, and high-tech parks with incentives such as tax relief were built. Several points highlight their reasoning behind attracting foreign R&D centers. On the one hand, foreign R&D centers encouraged the development of China's higher education system, since job opportunities were created. On the other hand, Chinese scientists living abroad were incentivized to move back to China, transferring their knowledge to locals. Apart from providing capital investment and knowledge transfer, China's market value could rise due to the close relationship with international foreign firms (Gassmann, Beckenbauer, and Friesike 2012).

In 2006, we could see a shift in policies when China introduced the *15-year National Outline for Medium- and Long-Term Science and Technology Development Planning* that emphasized the importance of domestic innovation capabilities. To this end, the Ministry of Science and Technology, the National Development and Reform Commission and the Ministry of Finance implemented instruments such as tax incentives, financial support, technological investment, and governmental procurement (Someren and Someren-Wang 2013). The outline intended in particular to decrease the reliance on imports of technology by 30% by 2020. Moreover, it set incentives and targets for patents, as did the *National Intellectual Property Strategy* (NIPS) from 2008 and the *National Patent Development Strategy* (NPDS) (Schmid and Wang 2017). While these gouvernmental interventions promote the dislodgement from foreign multinationals, they continued to play an essential part in the development of the Chinese national innovation system.

China officially encouraged domestic innovation in 2006 in the *15-year National Outline for Medium- and Long-Term Science and Technology Development Planning*, but had made efforts to independent themselves from the economic dependence on foreign firms in specific technological fields as early as 1986 with the *Spark Program* and the *863 Program* (Haour and Jolly 2014). Indeed, at the beginning of the 1980s, China launched several programs that meant to improve China's competitiveness in science and technology. These programs are the aforementioned *Spark Program* and *863 Program* but also the *Key Technologies R&D Program*, *973 Program* and *Torch program* ("National Programs for Science and Technology" 2017).

The UNESCO Science Report 2016 presents the priorities concerning distribution by field of China's *863 program* for high-tech R&D and the priorities for China's *973 program* for key basic R&D. The following figures (Figure 1 and Figure 2) illustrate the distribution of the number of new projects by fields and the distribution of budget for new projects by fields for 2012 ("UNESCO Science Report: Towards 2030" 2016).

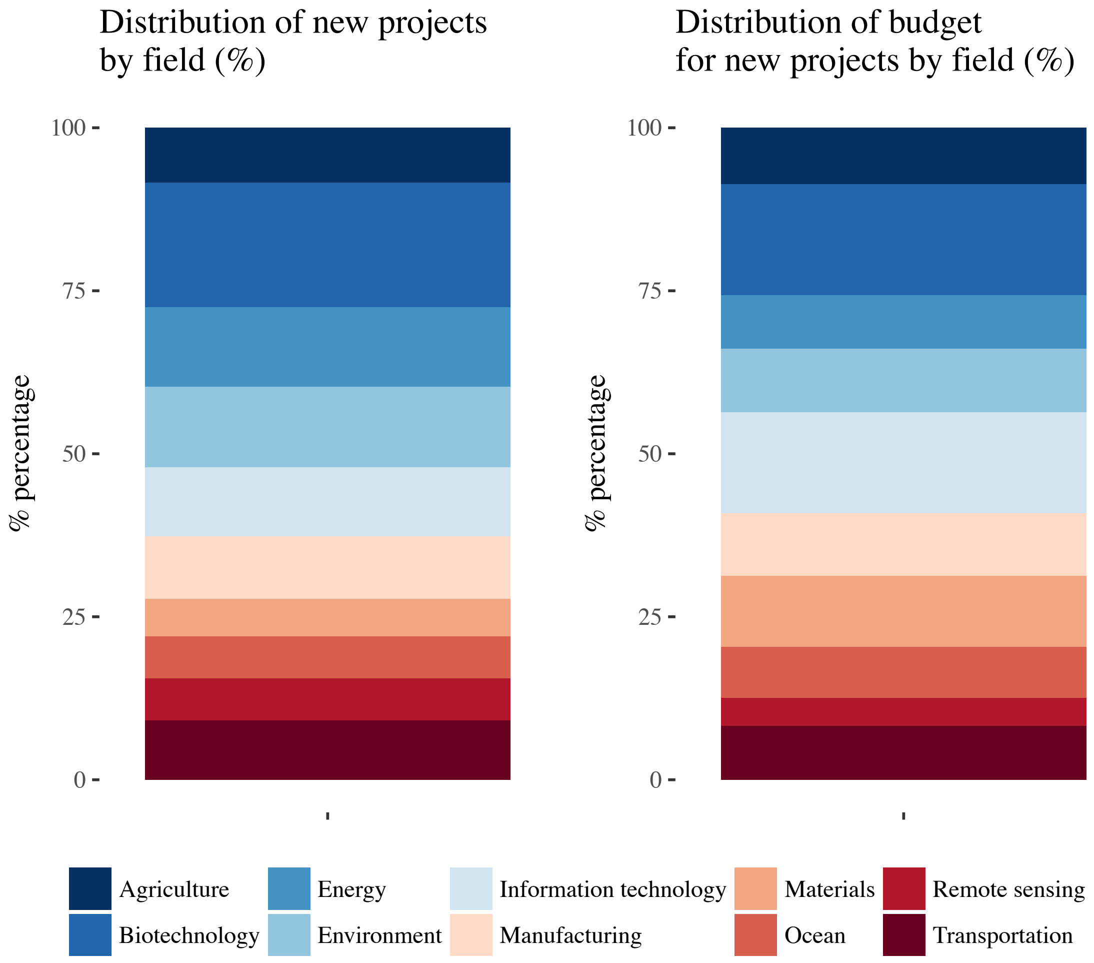


Figure 1: Priorities of China's national program for high-tech R&D in 2012 (863 Program)

Source: "UNESCO Science Report: Towards 2030" 2016; based on the Annual Report of the National Programmes of Science and Technology Development by the Planning Bureau of Ministry of Science and Technology, 2013

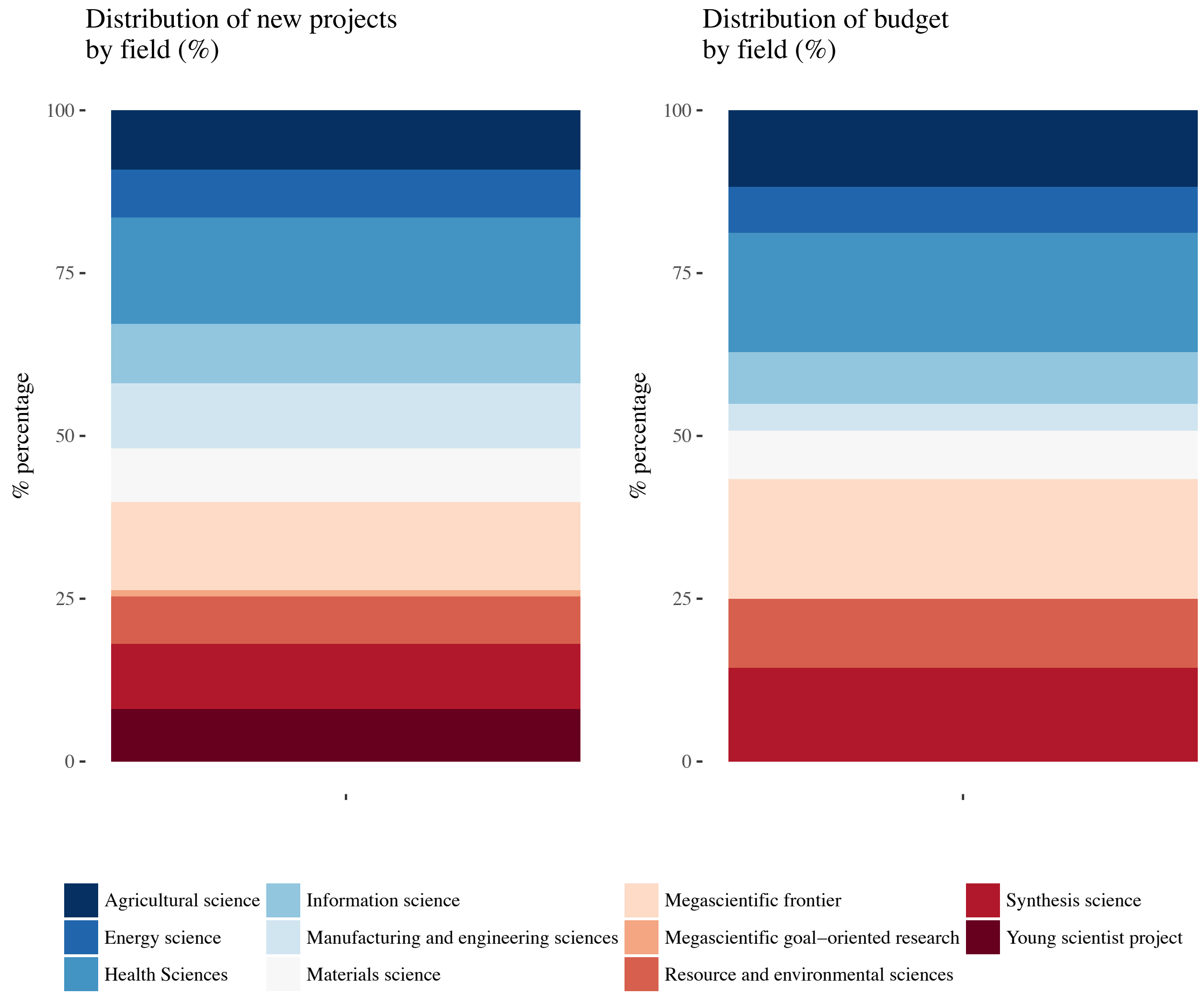


Figure 2: Priorities of China's national program for key basic R&D in 2012 (973 Program)

Source: "UNESCO Science Report: Towards 2030" 2016; based on the Annual Report of the National Programmes of Science and Technology Development by the Planning Bureau of Ministry of Science and Technology, 2013

With regards to high-tech R&D, biotechnology stands out as a clear emphasis of the *863 Program* (Figure 1). 19.1% of the projects and 17.1% of the budget were allocated to projects from this field in 2012. Key basic R&D, in turn, focused on Health Sciences with an allocation of 16.4% of the projects and 18.3% of the budget in 2012 (Figure 2). Regarding the distribution of budget for key basic R&D, most of the budget was, however, allocated to the field of the mega scientific frontier (18.4%) in 2012. These distributions indicate that there has been a lot of investment in R&D from the Chinese government related to the pharmaceutical industry.

Altogether, China's efforts have resulted in the creation of designated international innovation centers, such as Beijing and Shanghai ("A Closer Look at the 13th Five-Year Plan" 2016). They are being leveraged in China's 13th 5-year plan that implements incentives for certain industries, among others biotechnology. It set the goal that strategically identified emerging industries should account for 15% of total GDP by 2020 ("A Closer Look at the 13th Five-Year Plan" 2016). These designated innovation centers and the previously mentioned concentration on tech parks might be the reason why scholars have exposed a disparity in innovation performance between Chinese regions (Fan, Peilei, and Lu 2012; Li 2009). Thereby, the coastal regions of China are attracting most of the innovation activities ("UNESCO Science Report: Towards 2030" 2016). Guan and He (2007) refer to the disparate regional patenting activity by stressing that the areas of Beijing, Guangdong, Hong Kong and Shanghai contributed to 55.8% of the total patenting activity from 1995 to 2004. A firm-dominant regional innovation mode can explain this disparate innovation efficiency, instead of a university- and research-dominant (Li 2009).

In conclusion, the Chinese government has made efforts to foster foreign and – in recent years more and more – domestic innovation. The pharmaceutical industry is thereby one of the industries that were identified as valuable. Nonetheless, so far not the whole of China could benefit from such a program, but only a few regions.

2.1.1.2 Intellectual property rights

Intellectual property rights protection is a vital part of regulations affecting innovation, in particular in the pharmaceutical industry. While innovations need to be made available to the public, firms need the protection as an incentive to innovate and remain competitive.

Intellectual property not solely stands for patents but encompasses different kinds of intellectual property. The World Intellectual Property Organization lists patents, trademarks, industrial designs, utility models, plant varieties and newly geographical indications as forms of intellectual property in their World Intellectual Property Indicators report from 2017 ("World Intellectual Property Indicators 2017" 2017). In the following, the focus lies on patents as a form of intellectual property since the paper consists of a patent analysis, which deems most appropriate when analyzing the pharmaceutical industry. Without patent law, innovations could be easily copied or imitated and sold for a lower price not having to account for the high research and development costs.

China has experienced a considerate strengthening of its intellectual property rights (IPR) protection and enforcement in recent years (Zhang and Deng 2008). The changing regulations on intellectual property rights protection play an essential role in foreign MNEs' strategy development. MNEs initially hesitant to locate R&D activities in developing countries, like China, nowadays shift more and more activities abroad. In particular, activities that were initially from the lower end of the R&D activities – Bruche (2009a) considers them asset-seeking – shifted towards new product development with the changing IPR environment. However, intellectual property rights are essential for domestic innovators as well. Strong regulations on protection and enforcement ensure that innovators locate their R&D activities in China, which in case of pharmaceuticals helps balance the large generics industry.

While IPR protection has evolved due to the government's efforts to comply with the WTO regulations since 2001, China had adopted modern patent laws as early as 1984 (Fan 2014; Zhang and Deng 2008). Nevertheless, the protection before 2014 did not hinder firms in China from manufacturing foreign products under a new name (Jalfin 2017). Moreover, China was for a long time "[...] one of the primary sources of global counterfeit and knock-off goods" ((Plane and Livingston 2017: para. 1). In 2015, for instance, 88% of the fake goods seized in the US (by value) originated in China (Plane and Livingston 2017). More so, the OECD and EUIPO indicate in a report that China was the number one source of counterfeit products in 2013. 63.2% of the total seizures in 2013 originated in China (OECD/EUIPO 2016). Papageorgiadis, Cross, and Alexiou (2014) equally criticized China for its weak patent enforcement.

While in the past the patent law might not have been as effective, nowadays, the scope of protection of Chinese patent law is similar to European and American law, which facilitates the handling of IPR issues for foreign MNEs (Zhang and Deng 2008). Furthermore, due to its WTO membership, China is obliged to comply to the *Agreement on Trade Related Aspects of Intellectual Property Rights* (TRIPS) (Gassmann, Beckenbauer, and Friesike 2012). This similarity between Chinese, European and American patent law is not only manifested in the scope of protection, but also the process of granting patents (See appendix Figure 38). Furthermore, pendency times for the first office action and final decision are similar between China and the US, with a total of around 20 months ("World Intellectual Property Indicators 2017" 2017). Other emerging markets, such as India or Brazil have pendency times of around 90 months.

The GP Index by Park (2008) attests to China a high score in 2005 in term of patent protection, and thus confirms the mentioned improvements undertaken by the Chinese government. While patents can be filed at the national patent office, the State Intellectual Property Office of the People's Republic China (SIPO), assignees can decide to file their patent under the *Patent Cooperation Treaty* (PCT), also. The PCT allows filing a single international patent application that automatically applies to the chosen signatory states. While China is a signatory state, Macau is excluded from the PCT ("World Intellectual Property Indicators 2017" 2017). To recapitulate, the patent law and enforcement have become more similar to its European and US counterparts overall, but the coastal metropolitan areas demonstrated especially profound improvements (Zhang and Deng 2008).

Thus, it is without surprise, that global patent statistics reveal that China has become one of the key players on the international intellectual property stage in 2015. Indeed, 30% of all patents granted worldwide were filed in China. China became the largest patent issuing office in the world with over 1 million patent applications, surpassing the US (Jalfin 2017). Due to China's surge in patenting, research has been dedicated to the exploration of this phenomenon. Adopting different perspectives, researchers have studied patent value (Gupeng and Xiangdong 2012; Schmid and Wang 2017), the patents' impact on economic growth (Nilsson 2013), the patent law and the intellectual property right system (Yueh 2009; Luan and Zhang 2011) or possible explanatory factors for the patenting surge (Hu and Mathews 2008). Looking into the recent patenting surge in China Hu and Jefferson (2009) explained that foreign direct investment is significantly contributing to the rising patenting of domestic firms. Although regulations play a role as well, the competitive environment fosters an innovative environment (Hu and Jefferson 2009). Another paper disclosed that a significant part of the surge in patenting is due to previously not actively patenting firms starting to patent (Hu, Zhang, and Zhao 2017). The Chinese government's efforts to promote innovation might foster this firm behavior.

While China is experiencing a patenting surge, some claim that China is characterized by low patent quality and granted patents that should not have been granted (Jalfin 2017). Jalfin (2017) suggested that this is the reason why the 13th Five-Year Plan (2016-2020) aims to strengthen intellectual property rights through enforcement. Strengthened IPRs increase competition, forcing Chinese patent quality to improve (Jou, Wu, and Chan 2015). The current Five-Year Plan not only strengthens IPRs but has specific goals. It aims to double on the one hand the number of patents to 12 per every 10,000 people and, on the other hand, to increase the number of international patent applications to 60,000. Besides, China seeks to raise the intellectual property royalties earned abroad to $10 billion and increase the public's perception of intellectual property protection to a satisfaction rate of 80% (Jalfin 2017).

This China's interest in strengthening intellectual property rights protection thereby does in part result of their will to move up the value chain. Then again, the protection of innovation of domestic institutions and firms against foreign competitors plays a crucial role as well. This said, Chinese companies are increasingly facing litigation issues within greater China, and abroad, wherefore China created its "own state-funded patent aggregation entities", called Ruichuan IPR Funds, established in 2014 (Jou, Wu, and Chan 2015: para. 13).

Having established that IPR protection has strengthened and that patent law is recently similar to the US and EU, conditions of the Chinese patent law will be highlighted. First, the litigation and enforcement of patents in China "is based on the Civil Procedural Law that governs the procedures for patent litigation, the Patent Law and the Implementing Regulations of the Patent Law" (Gassmann, Beckenbauer, and Friesike 2012: p. 23). Patent infringement is handled in the first instance by the Intermediate People's Court - each medium-sized city has at least one - and can be appealed to the next higher court (Gassmann, Beckenbauer, and Friesike 2012). The Intermediate People's Court having jurisdiction over the alleged infringer or the infringing activity is ruling the case (Zhang and Deng 2008). Another option is to fill it "with a local Administrative Authority for Patent Affairs, sometimes called Intellectual Property Bureau (IPB), that has jurisdiction over the alleged infringement act" (Zhang and Deng 2008: p. 1236). There exist specialized IP divisions that treat patent infringement on a case-to-case basis (Gassmann, Beckenbauer, and Friesike 2012).

While administrative enforcement is generally ruled under local government's jurisdiction, hence influencing firms' location choice within China, three specialized IPR courts have been established in Shanghai, Guangzhou and Beijing in 2014 ("China Strengthens Judicial Protection of Intellectual Property Through Specialized IPR Courts" 2017; Zhang and Deng 2008). They are responsible for ruling complex technology cases.

With regard to Chinese patent law "Article 11 of the Chinese Patent Law prohibits unauthorized making, use, offer for sale, sale or import of a patented product. The provision also prohibits the unauthorized use of the patented process (and use, offer for sale, sale or import of products directly obtained by the patented process) for production or business purposes" (Zhang and Deng 2008: p. 1235). However, exceptions exist that are in particular pertinent for biotech or drug patents as Zhang and Deng (2008) state. While the unauthorized use of patented products or processes is generally prohibited, it is allowed as long as the product or process was started to be manufactured or used before filing the patent. This exception is only valid for the person that has used or made them prior and limited to the context it is currently used in (Zhang and Deng 2008). It seems to be crucial for products that have a long development and production pipeline. Hence, this explains why Zhang and Deng (2008) consider the exceptions pertinent for biotech and drugs. While in the past, China did not harmonize with the EU and US definition of a novelty invention, it complies since 2009, meaning that an invention filed abroad cannot be considered a novelty in China (Gassmann, Beckenbauer, and Friesike 2012). This constitutes a great improvement and is in particular crucial for foreign MNEs.

Although we have established that the Chinese business environment has improved in terms of intellectual property protection and enforcement, it is still facing certain threats. Certainly, there is a threat of direct infringement of an IPR, but there are also threats that are "within the boundaries of the law", that Gassmann calls legal-licit threats (Gassmann, Beckenbauer, and Friesike 2012: p. 31). These encompass among others the "circumvention of protected technologies and products", "obscurity of interpretations of legal changes" and a "weak enforcement regime" that makes the success of a litigation case uncertain (Gassmann, Beckenbauer, and Friesike 2012: pp. 36-37). Moreover, patent holders are with few exceptions subject to prove an infringement when wanting to file a claim of infringement. A "formal discovery mechanism for evidence collection", however, impedes patent holders to prove their claims (Zhang and Deng 2008: p. 1238). Nonetheless, data shows that IP disputes of foreign firms, for instance, have win rates that have rosen to around 80% and injunction rates to around 98%. Moreover, foreign firms fair just as well as privately-owned Chinese firms in IP litigation disputes (Rana, Marcus, and Fan 2018). Thus, while certain threats remain, there have been considerable improvements.

Finally, the court rulings on the municipal level can be considered problematic (Zhang and Deng 2008). China is recurrently struggling with corruption. In fact, it is on rank 79 of 176, in the Corruption Perceptions Index 2016 ("Corruption Perceptions Index 2016 - Transparency International" 2016). The country thereby achieved a score of 40 on a scale from 1 to 100, highlighting the country's struggle with corruption. Not without reason, it is crucial for Chinese firms and foreign MNEs to engage in *Guanxi*, "that is, personal, financial or political connections that allow one to gain advantages over those who do not possess such relationships" (Zhang and Deng 2008: p.1237).

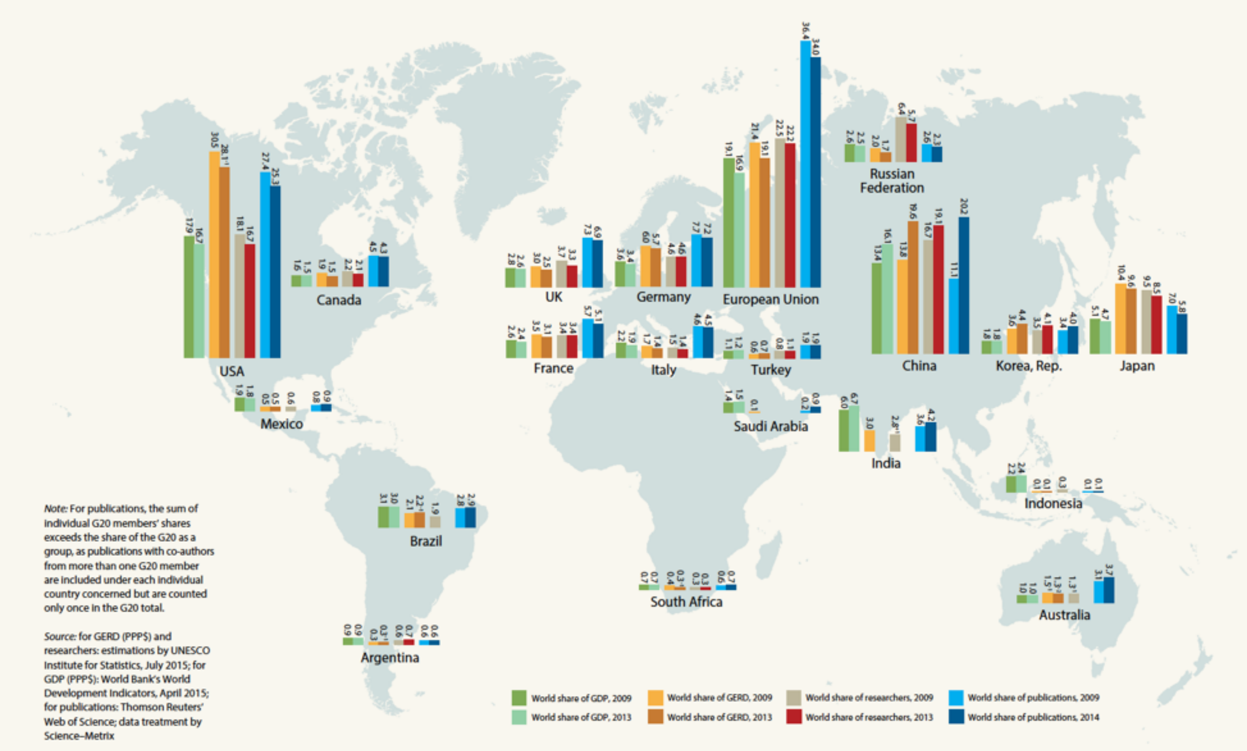
2.1.2 Indicators of China's innovation capacities

The following section will give an overview of China's innovation capacities, elaborating on the classic innovation indicators. While many scholars and newspapers emphasize the rising number of patent applications as evidence for China's growing innovation capabilities (e.g., Veugelers 2017), others conclude that China's efforts to produce frontier innovation failed (e.g., Schmid and Wang 2017). How do the indicators of innovation capabilities compare to each other? Classic indicators will be disclosed that assess the explanatory power of a patent analysis. It is, however, important to note, that the numbers and graphs presented do not picture a holistic and unambiguous measurement.

GERD, world share of GDP, world share of researchers and world share of publications

Innovation capabilities have been measured by several indicators that can be classified into innovation input and innovation output indicators (Fan 2014). The innovation input of a country is traditionally measured by indicators such as R&D investment, R&D personnel and the number of tertiary university degrees while the output is represented by the number of scientific publications, patent applications or grants and the number of citations (“UNESCO Science Report: Towards 2030” 2016).

The UNESCO Science Report from 2016 offers a comprehensive overview of innovation capabilities around the world and illustrates China’s state of innovation compared to the rest of the world (Figure 3).



Note: For publications, the sum of individual G20 members’ shares exceeds the share of the G20 as a group, as publications with co-authors from more than one G20 member are included under each individual country concerned but are counted only once in the G20 total.

Source: for GERD (PPP$) and researchers: estimations by UNESCO Institute for Statistics, July 2015; for GDP (PPP$): World Bank’s World Development Indicators, April 2015; for publications: Thomson Reuters’ Web of Science; data treatment by Science–Metrix

Figure 3: World shares of GDP, GERD, researchers and publications for the G20, 2009 and 2013 (%)

Source: (“UNESCO Science Report: Towards 2030” 2016)

In terms of world share of Global Gross Expenditure in Research and Development (GERD), China was second in 2013 with around 19.6%, behind the US (28.1%) and followed by the EU (19.1%) (“UNESCO Science Report: Towards 2030” 2016). In 2009, China was still behind the EU. However, both the US and the EU have experienced a decrease in their share from 2009 to 2013. The same development can be seen for their world share of GDP, world share of researchers and world share of publications. Despite this decrease in world shares, the EU and the US surpassed China in most of the indicators. For instance, regarding the world share of GDP and the world share of publications, the US and the EU both outperformed China. While China had a world share of GDP of 16.1%, the EU had a 16.9% world share of GDP and the US a 16.7% world share of GDP in 2013. Nonetheless, these numbers are close. It is evident, that China still has to catch up, when comparing its world share of publications (20.2%) to those of the US (25.3%) and the EU (34.0%). While there remains a need to catch up, it has to be noted, that China has been able to almost double its share of publications in four years. Regarding the world share of researchers, China has surpassed the

US from 2009 to 2013, ranking second (19.1%) behind the EU (22.2%) in 2013 (“UNESCO Science Report: Towards 2030” 2016). Overall, China has improved its innovation indicators. Differences between China, the US and the EU are small, except for the world share of GERD and world share of publications, where differences are still considerable. Finally, considering that the EU is not a country, China has considerably higher shares than any other country except the US (Figure 3).

The numbers of this report highlight that China’s innovation input is somewhat close to the input of the US and the EU. In contrast, China’s output indicator, the world share of publications is still lagging behind. Nonetheless, China has improved its innovation capabilities dramatically throughout the last decades, although often criticized as being more of a quantitative nature than a qualitative (Fan 2014). Since the variety of indicators for innovation output was limited to publications, we shall deem it necessary to include other measurements.

Patent applications

By 2016 Chinese patent applications made up 42.8% of all applications worldwide. In absolute numbers, this means that China filed 1,338,503 out of 3,127,900 patents. Compared to China, the US filed 605,571 patents (“World Intellectual Property Indicators 2017” 2017). Overall, patent offices have experienced an increase in patent applications over seven years, with an annual growth of 8.3% in global patent filings in 2016. China, being the main driver of growth, experienced an increase in patent applications of 21.5% in 2016. In contrast, the US experienced an increase of 2.7%. Indeed, China’s patent applications experienced a growth from 2000 on that is not comparable to any other growth of the top 5 offices. Around 2005 it surpassed the Korean Intellectual Property Office (KIPO) and the European Patent Office (EPO) for the first time. Five years later it succeeded in outpacing the other top 5 offices as well. When having a closer look at the ratio between resident and non-resident applications in different offices, the WIPO Indicators 2017 report illustrates that residents filed most of the applications at the SIPO. In contrast, applications at the USPTO and EPO were filed by residents and non-residents in balance. When looking at the origin of applications instead than looking at offices, the WIPO Indicators 2017 report explains that “96% of all applications from China are filed in China, and only 4% filed abroad. In contrast, filings abroad constitute around 43% of total applications from Japan and the U.S.” (“World Intellectual Property Indicators 2017” 2017: p. 33). In term of utility model applications, also called petty patents, China’s share of worldwide applications is even higher. Utility models are similar to patents but have a “shorter term of protection and less stringent patentability requirements” (“World Intellectual Property Indicators 2017” 2017: p. 39). In 2016, China filed 95% of the utility applications. This corresponds to 1,4775,977 out of 1,553,300 utility model applications (“World Intellectual Property Indicators 2017” 2017). China increased its share of utility model applications by 30.9% from 2015 to 2016.

While China made up 42.8% of all applications worldwide, the whole of Asia made up 64.6% in 2016 (Figure 4). Thereby, Asia accounts for the most significant share of patent applications worldwide. Asia is followed by North America with a share of 20.5% and Europe with 11.3% in 2016. The region Asia has the largest average growth from 2006 to 2016 with around 8.5% (“World Intellectual Property Indicators 2017” 2017). Since China accounts for 42.8% of the worldwide applications, 21.8% of worldwide applications originated from the rest of Asia, thus not only China but also other Asian regions are experiencing a patenting surge.

When comparing their share with the share of North America and Europe, their part is the second largest before North America and Europe, but after China. The countries that account mostly for the other Asian countries are Japan and the Republic of Korea, which account for 318.381 and 208.830 patent applications. Moreover, in comparison to other developing countries, China's part is quite impressive.

With regard to income groups, the increasing importance of the upper middle-income group is revealed. While the group accounted for 18.3% of worldwide patent applications in 2006, it was 47.6% in 2016. In contrast to the upper middle-income group who increased its share of worldwide applications, the high-income group – consisting of countries, like the US, Switzerland, and Germany, known for their innovative capabilities (Achilladelis and Antonakis 2001) – has decreased theirs from 78.3% in 2006 to 49.6% in 2016. This decrease might reflect a shift of innovative capabilities away from high-income countries, and thus the usual suspects, towards new players in the global innovation landscape.

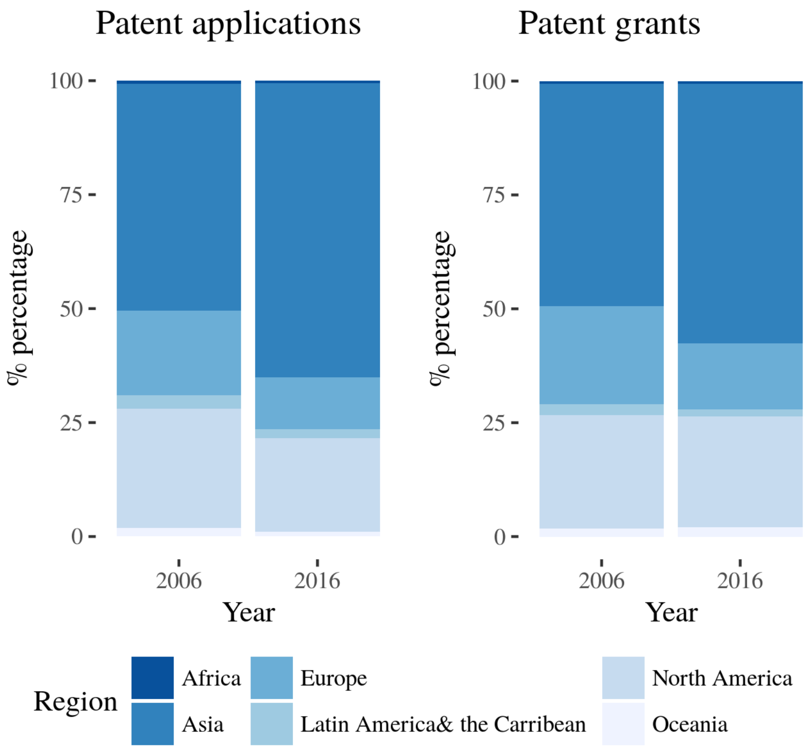


Figure 4: Patent grants and applications per region 2006 and 2016

Source: "World Intellectual Property Indicators 2017" 2017

The pattern found for patent applications is similarly reflected in patent grants (Figure 4). The Asian region accounts for the most substantial part of worldwide granted patents with 57%. In 2006, this share was still slightly below the 50% mark with a share of 48.8% of worldwide applications. Again, North America is second in terms of the share of worldwide patent grants. Its share has slightly decreased from 25% in 2006 to 24.4% in 2016. As with patent applications, the upper middle-income group has increased its share of worldwide patent grants considerably from 2006 to 2016 (from 15.4% to 35.1%). In contrast, the high-income group's share of patent grants has decreased from 81.4% in 2006 to 62.7% in 2016. Hence, both in terms of patent applications and grants the upper-middle-income group has seen considerable growth in the share of the world total ("World Intellectual Property Indicators 2017" 2017). This growth indicates a shift of innovative capabilities away from high-income countries towards upper-middle-income countries. The global innovation landscape is changing, and China, in particular, is gaining importance as a country accounting for most of the patents filed worldwide. The patenting of the whole Asian region accentuates China's increasing importance.

After having determined the patent offices with the most patent applications, it is essential to take into account the firm-level. On the micro-level, where do the firms that patent most file their patents? Does the origin of the top 100 firm applicants mirror the overall pattern of the Asian region being the most significant patent applicant? Indeed, the top 10 applicants worldwide are Asia-based multinationals. These are: Canon Inc. (Japan), Samsung Electronics (Republic of Korea), Panasonic (Japan), Toshiba (Japan), Toyota Jidosha (Japan), Mitsubishi Electric (Japan), Huawei Technologies (China), LG Electronics (Republic of Korea), State Grid Corporation of China (China), Seiko Epson (Japan) ("World Intellectual Property Indicators 2017" 2017).

The following table (Table 1) illustrates the origin of the top 100 applicants from 2011 to 2014 ("World Intellectual Property Indicators 2017" 2017). All in all, nine origins are represented. While the list mainly contains MNEs, 14 Chinese universities are present, also. Finally, on country-, as well as firm-level, China is featured as a significant innovator ("World Intellectual Property Indicators 2017" 2017). Moreover, the Asian region stands out as the region with the most significant share of the top 100 patent applicants. This is in concordance with the findings from Figure 4 where the Asian region accounted for the most substantial part of worldwide applied for and granted patents.

Table 1: Origin of top 100 applicants from 2011 to 2014

Country	**Number of App.**
Japan	40
China	26
Korea	15
US	9
Germany	6
France	1
Netherlands	1
Sweden	1
Taiwan	1

The assessment of a country's innovation capabilities requires the distinction between different types of innovators such as business enterprises, governments, higher education and private non-profit organizations. While the list of the top 100 patent applicants reveals that 14 out of the 26 Chinese top innovators are universities, this ratio is not reflected in the distribution of R&D personnel between different types of innovators. In 2015, business enterprises employed 2,9 million R&D personnel out of a total of 3,8 million. Around 0,5 million were employed by the government and 0,4 million by the higher education sector. Employment by private non-profit organizations can be neglected since it amounts to a few thousand ("OECD Statistics - Science, Technology and Patents" 2017). The UNESCO Science Report from 2016 confirms that regarding R&D expenditure the business enterprise sector is predominant (*Data from 2013*). While 76.6% of R&D expenditure are attributed to the business enterprise sector, 16.16% belong to the government sector and 7.23% to higher education in 2013. This data is consistent with the distribution of R&D personnel. Achilladelis and Antonakis (2001) revealed that the number of innovations and patents is related to the level of R&D expenditure. Hence it is not surprising, that in terms of patents and R&D expenditure most originate from business enterprises.

Scientific publications

We have shown that China ranks third after the US and Europe in terms of share of scientific publications in 2013. When having a closer look at scientific publications, it is revealed that China published mostly papers related to engineering (34.8%) in 2013, followed by chemistry (24.5%), computer sciences (21.1%), physics (19.6%), mathematics (18.0%), biological sciences (13.9%) and finally medical sciences (7.5%) (Veugelers 2017; "UNESCO Science Report: Towards 2030" 2016). Since chemistry, biological sciences and medical sciences play an important role for China's innovation capabilities in terms of scientific publications, one can assume that domestic innovation in pharmaceuticals is strong. This assumption is justified since historical documentation has shown that chemistry, biological scientific and medical achievements apply in particular to innovation in the pharmaceutical industry (Achilladelis and Antonakis 2001). However, around 24.4% of the publications are created in co-authorship, which means that scientific publications are not solely originating from China. The US constitutes the main partner with around 119,594 co-published papers, followed by Japan with 26,053 papers in the period 2008 to 2014 ("UNESCO Science Report: Towards 2030" 2016). Apart from the scientific publications that indicate a focus on pharmaceutical innovation, two of China's 13 mega-engineering programs, that were made public refer to new drug development or treatments of major diseases. Hence, it seems reasonable to assume that China is trying to foster its pharmaceutical industry.

While it has been shown that China's innovation input is comparable to the Western standard, the UNESCO Science Report explains that China is nonetheless still dependent on foreign core technologies for its innovation output. In contrast, the Global Innovation Index from 2017, listing China as the sole middle-income economy in the top 25, ranks the country overall 22nd, with an innovation input rank of 31 and output rank of 11[1] (Cornell University, INSEAD, and WIPO 2017). This ranking suggests that China's innovation output is, in fact, closer to advanced economies than its input. Although the relative ranking suggests this, when taking into account the absolute scores, one can notice that China's input scores higher than its output. Furthermore, the index attests to China a high innovation efficiency, ranking it 3rd in 2017, despite identifying institutions as a weakness of China's innovation landscape.

In conclusion, the assessment of whether the inputs and outputs of China's innovation are significantly similar to those of advanced economies depends on the measurement. Nonetheless, the different indicators revealed that China has progressed. It is evident that China has become a key player in innovation. Both the patent indicator and other indicators have highlighted China's role. Moreover, China has undertaken

[1] The Global Innovation Index 2017, distinguishes between China and Hong Kong (China); hence the ranking does not take into account Hong Kong's scores (Cornell University, INSEAD, and WIPO, 2017)

investments and increasingly published scientific publications with regard to pharmaceuticals. Thus, a look at the Chinese pharmaceutical industry is of interest.

2.2 The pharmaceutical industry

Within the constraints of this thesis, the pharmaceutical industry has been chosen as the unit of analysis due to several reasons. On the one hand, due to its propensity to patent – to secure the sale of branded drugs against generics – (e.g., Gassmann 2008), and on the other hand due to its global nature and increasing embeddedness in global innovation networks (Bruche 2009b). Pharmaceuticals are one of the three major industries where MNEs disperse R&D globally (Bruche 2009b).

Since this paper attempts to assess innovation dynamics of pharmaceutical innovation in China, the industry will be described in the following. This description will facilitate the understanding of the context surrounding pharmaceutical innovation.

2.2.1 Evolution

The emergence of the pharmaceutical industry occurred around the middle of the 19th century. It originated from two different sources, either apothecaries or chemical companies. While the latter moved into the wholesale production of drugs or dye, the former did research and discovered medical applications for their products. For instance, two of the leading companies in pharmaceuticals nowadays, Merck and Pfizer, formerly were known from another field. Merck started as an apothecary, while Pfizer as a producer of organic chemicals (Daemmrich and Bowden 2005).

After its emergence in the 19th century, the industry enjoyed its golden era in the mid 20th century. Numerous inventions lead to a healthier society. Among others, female deceases due to infections during childbirth declined by more than 90%, and illnesses such as tuberculosis and pneumonia could be cured. In the 1940s the US became the leader of the industry, producing over half of the world's pharmaceuticals and accounting for one-third of international trade in medicines (Daemmrich and Bowden 2005). Among others, the wartime support for research that accelerated the development of products was responsible for their position. Firms, as well as academics researched, constantly holding a close relationship. This cooperation can be considered a form of innovation network. Networks, however, were restricted to national collaborations. Trans-national innovation networks had not been documented yet. Furthermore, in the mid 20th century, the industry saw the rise of regulations. Firms were, for instance, assigned responsibility for product testing. At times, firms marketed their products mainly to physicians who prescribed the drugs to patients (Daemmrich and Bowden 2005).

In the 1960s, US firms were facing threats from European competitors. Due to different regulatory bodies, these were faster at launching new products. Thus, US firms started diversifying into consumer products such as fragrances and cosmetics. As a consequence of several tragedies in Europe, firms started investing more in preclinical and clinical testing (Daemmrich and Bowden 2005).

Meanwhile, US firms were experiencing decreasing new drug approvals in the US that forced them to act. They diversified even more, for instance into diagnostics and household goods (Daemmrich and Bowden 2005). Furthermore, internationalization constituted a countermeasure. Hence, they sold their drugs increasingly overseas, forming research networks with European, South American and Asian labs in that event. The trans-national innovation network was born. Nonetheless Achilladelis and Antonakis (2001) explain that despite internationalizing, the firms kept their vital operations in the home country. Firms carried out parts of the drug development process abroad, since national governments of the host-countries required them to do clinical trials in their countries or else firms risked not getting approval for the drugs (Ramirez 2006).

Starting in 1984 with the US Food and Drug Administration (FDA)'s Act authorizing the approval of generics without preclinical or clinical testing if referring to the data of the originator, branded-name manufacturers saw competition from generic drug manufacturers. Since then they have become a considerable part of the industry, dependent on the innovations of the brand-name manufacturers. The latter ensuring their survival by patenting innovations and delaying the production of generics.

Finally, it is important to mention that since the 1980s the field of biotechnology is becoming an integer part of the industry. Increasingly large pharmaceutical companies and biotechnology firms collaborate to leverage one's respective complementary capabilities (Chiaroni, Chiesa, and Frattini 2008). However, with the slowdown of economic growth in the US in the 1980's, American pharmaceutical companies divested their non-pharmaceutical businesses and opted for mergers with or acquisitions of other pharmaceutical companies resulting in big pharma companies operating globally. These mergers resulted in the birth of blockbuster drugs that could be launched worldwide (Achilladelis and Antonakis 2001). Blockbuster drugs refer to drugs with "annual sales ranging from US$ 300 to US$ 3000 million" (Achilladelis and Antonakis 2001: p. 565). The mergers and acquisitions, not limited to the American companies, which were often cross-border, helped firms to gain access to new technologies, products, and markets. This way, fixed costs could be distributed (Bruche 2009b). The diversification resulted in adaptability to a changing competitive environment and explained the survival of large pharmaceutical companies founded almost a century ago (Achilladelis and Antonakis 2001). All in all, since having become more international, the industry is embedded in a "framework of international regulations aimed largely at liberalizing trade. Thus, the effects of national policies on the intensities of the driving forces for TI [technological innovations; author's note] became marginal." (Achilladelis and Antonakis 2001: p. 572).

At the beginning of the 21st century, composed of mainly American and European MNEs, the global pharmaceutical industry experienced difficulties due to low new drug outputs, long research and development cycles, rising costs and potential patent expiration. Thus, firms disaggregated their pharmaceutical value chain; big pharmaceutical firms kept their high value-added innovations in-house and outsourced the low value-added business to developing countries (Huang 2012). Thus, while previously the internationalization process in R&D was focused on the triad region of developed countries, the 21st century saw a shift towards developing countries (Bruche 2009b). In general, open innovation schemes gained ground (Gassmann 2008). Boutellier, Gassmann, and Zedtwitz (2008) confirmed the internationalization of R&D of MNEs in the 1990s and early 2000s in their book *Managing global innovation: uncovering the secrets of future competitiveness* from 2008. These recent developments could challenge the traditional business model of firms, which rely on blockbusters for growth (Gassmann 2008). Indeed, Achilladelis and Antonakis (2001) explain that firms increasingly patent incremental innovation, due to growing competitivity and the increasing pace of innovation. To summarize, the pharmaceutical industry has become global and connected, increasingly integrating developing countries and diversifying their products.

2.2.2 Structure and production pipeline

In the following chapter, the recent structure and the dynamics of the pharmaceutical industry will be highlighted, which clarifies the industry's innovation processes. The pharmaceutical industry while being highly profitable, has become very capital intensive. With technological progress, the development of new drugs has increased, but costs have risen as well. Due to these rising costs and increasing competition in the industry, and due to generic manufacturers, companies increasingly specialize in one area of pharmaceuticals. They currently concentrate on biologics, cancer treatment, rare diseases and biotechnology. While the manufacturing of brand-name products concentrates around large multinationals, generic manufacturers are rather local ("From Vision to Decision - Pharma 2020" 2012). These generic manufacturers pressure branded manufacturers since they leverage the expiration of patents and bring similar products to market at a lower price. Conversely, some companies, like Novartis, enter the market of generics themselves, counteracting this threat. Therefore, Achilladelis and Antonakis (2001) question the traditional structure of branded-name drugs versus generics. They call into question most firms' reliance on blockbusters for future growth, but nevertheless reveal that radical innovations are commercially more successful than following the lead of a competitor (Achilladelis and Antonakis 2001).

On another note, alternatives to pharmaceutical products challenge the industry. Indeed, they face a threat of substitution through treatments like surgery or homeopathic treatments. Homeopathic and herbal remedies accounted for 7.5 billion US dollars in retail sales in 2017, compared to 6.28 billion US dollars in 2011 ("U.S. Retail Sales of Homeopathic and Herbal Remedies 2011-2017" 2018). While not comparable to the 459 billion US dollars of revenue of the pharmaceutical industry in the US in 2016, the sales of homeopathic and herbal remedies have been growing. Health insurers, however are reluctant to cover these. Thus, the industry has a low level of threat through substitutes since pharmaceutical drugs are acknowledged for their beneficial use and not every drug can be compensated by alternative treatments (Gassmann 2008).

While the theat of substitution is low, the industry experiences a strong risk of new entry of potential competitors, despite existing barriers. In pharmaceuticals economies of scale play an important role with regards to R&D, marketing, and sales. These economies of scale are achieved by the global market presence of large multinationals. Likewise, the long development cycle of new drugs and uncertain success pose a threat to new entrants. In fact, it takes between 8 to 10 years from the R&D to the commercialization of radical innovations (Achilladelis and Antonakis 2001), or 9 to 13 years (Ding, Eliashberg, and Stremersch 2014). Firms often try to delay the patenting of innovations, so that the commercialized product is protected as long as possible. While highly regulated and relying on a brand's image, the outsourcing and consolidation of R&D facilities have made it less difficult for new firms to enter the industry (Gassmann 2008; Turk 2016). Contract Research Organizations (CRO) have expanded into the pharmaceutical market (Gassmann 2008). Nevertheless, the entry is rendered difficult by the limited access to distribution channels. Overall, there is evidence that CROs and smaller biotech firms, in particular, defy the barriers and enter the market (Gassmann 2008). Hence, despite these strong barriers, the market is facing a strong risk of new potential competitors.

On the other hand, the rivalry between established competitors is strong. Manufacturers profit from patented products and the first-mover advantage since they can achieve a higher market share. In fact, no single company controls more than a 6 % share ("Market Share Top Pharma Companies Rx Drugs Sales Globally

2024” 2018). And the top 20 drug manufacturers control a little less than 65% of the global market (Figure 5). Firms “try to improve their position by means of price competition, acquisitions, advertising battles and new product introductions” (Gassmann 2008: p. 28). In view of this strong rivalry and the risk of entry of new competitors, pharmaceutical firms might change their business model, diversifying into generics or changing their innovation processes, since the firm’s survival is dependent on innovation.

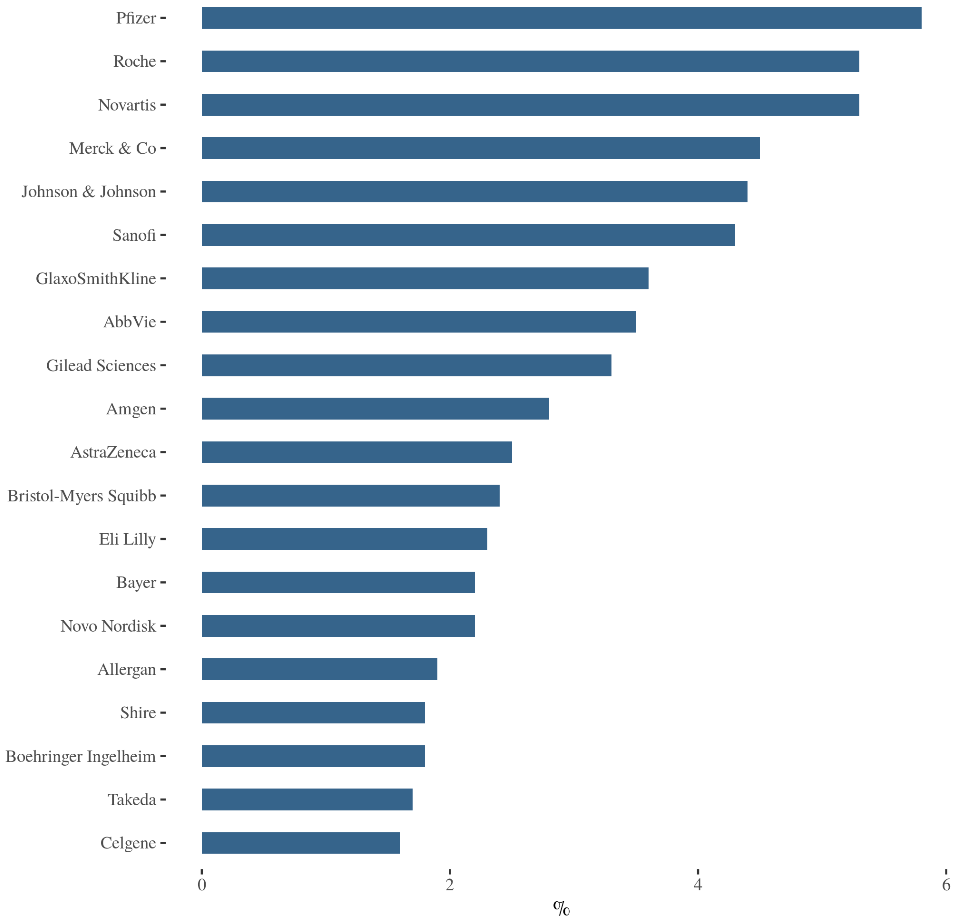


Figure 5: Top 20 pharmaceutical companies worldwide based on prescription drug market share in 2017

Source: “Market Share Top Pharma Companies Rx Drugs Sales Globally 2024” 2018

The industry is facing a certain level of dependence from suppliers. For instance, pharmaceutical manufacturers rely on the supply of raw material, which they need for the production of drugs. Because of regulations, there are high standards for raw material as well as the production process in general. Some firms produce some material synthetically themselves, lowering their dependence and ensuring their quality standards themselves. Nonetheless, suppliers have a certain degree of power in the industry, especially when handling rare material. Apart from suppliers of raw material, biotechnology firms, manufacturing plants, local marketing partner and patients can be considered suppliers as well. Indeed, patients are “supplies” for clinical trials. While generally suppliers do not manifest a strong negotiation power, biotech firms do (Gassmann 2008).

Pharmaceutical firms sell their pharmaceutical products to different customers: consumers, physicians, pharmacies, hospital boards, governments, public health insurers, etc. Therefore, the pharmaceutical distribution is highly complex and fragmented (Turk 2016). While consumers can choose over-the-counter products themselves, pharmacists, physicians, pharmacies, hospitals, and governments are gatekeepers between the consumer and the manufacturer, with the power to choose the drug the consumer is going to use. Hence, the latter are essential stakeholders, more so, considering that the global prescription market is expected to increase from 780 billion USD in 2016 to 810 billion USD in 2017, while the pharmaceutical spending overall decreased from 2005 ("ISIC 2423 of Pharmaceuticals, Medicinal Chemicals and Botanical Products" 2017). Health insurance companies in general and public health insurance in particular, play an increasingly important role. In mature markets, governments started encouraging the prescription of generics, since they are cheaper than branded drugs. They rae not willing to pay the price-premium, which goes for China as well (Chervenak 2005). In fact, instead of condoning drugs from branded manufacturers as better and paying the price premium, some started asking for evidence of the added value of the branded drugs ("From Vision to Decision - Pharma 2020" 2012). Moreover, health insurers play an active role in setting the prices for drugs, since they are a large buyer. One might even say they have monopoly power to a certain degree; they have a strong negotiation power in particular in "countries with nationalized healthcare and tight price controls" (Gassmann 2008: p. 26). This power poses a problem to drug manufacturers since they are reliant on mature markets paying a certain price to enable them to sell the same drug for cheaper in developing countries, following their moral responsibility. But finally, due to the recent development of co-payments for branded prescriptions, the consumer is again gaining more importance and moves into the focus of pharmaceutical companies for prescription drugs as well as OTC drugs. In fact, in the US, some patients have formed so-called Health Maintenance Organizations (HMOs) that can influence price negotiations through their aggregated power (Gassmann 2008). In conclusion, prices can be regulated by different entities: HMOs, and national health care systems. The latter have an enormous impact on prices, since health insurance organizations reimburse part of the product prices (Gassmann 2008). Generally, governments can regulate prices through federal authorities (except for the US and New Zealand).

When explaining the different dynamics around the pharmaceutical industry, one has to highlight the influence of regulations. They are crucial to the industry, not only due to IPR protection that ensures the backing of high research and development costs, but also due to the sensitive nature of the industry. As the recipient of health care, the public has an interest in regulating certain parts of the innovation process. Public laws and regulations impact IPRs, as well as R&D regulations and product registrations, price regulations and national healthcare systems (Gassmann 2008). Regulatory entities, such as the Food and Drug Administration (FDA) in the US or the CFDA (Chinese Food and Drug Administration) manage the authorization and registration of new drugs. The quality, efficacy, and safety of the new drug are evaluated before the entities grant the drug's registration in its respective market (Gassmann 2008). This process takes between 1.4 and 1.7 year in Germany, Australia, Spain and the US. While drugs generally have to pass this process ensuring their safety and benefit, there has been resistance of the public against new technologies such as bio- and gene-technology. Highly-developed industrial countries, in particular, have faced the public's resistance and regulate these technologies restrictively. As a consequence, pharmaceutical companies locate R&D activities in these research fields abroad where regulations are more favorable (Gassmann 2008). This is the case for China, where for instance gene-editing trials were allowed on humans while these trials were not allowed in the US (Rana, Marcus, and Fan 2018).

Lastly, with regards to IPR protection, one has to mention that there are international patent laws as well as national patent laws regulating the industry's innovations. The patent law, although harmonized between many countries, nevertheless differs across countries. China's patent law has for instance become more and more similar to the American and European patent law. For some research fields like gene-technology or new technologies they have however not been harmonized yet (Gassmann 2008). This can be considered beneficial or disadvantageous for society or national industries. China for instance benefits from lax regulations with regard to gene-technology.

In conclusion, the pharmaceutical industry remains a profitable but complex industry despite recent developments threatening the pharmaceutical companies monopolistic market power. Indeed, both new entrants and the rise in generic manufacturers pose a competitive threat (Gassmann 2008). This competitive threat is changing the industry that increasingly specializes and relies on collaboration in the innovation process. This is one of the reasons of the shift of R&D activities to developing countries. Figure 39 in the appendix summarizes the Five Forces, after Porter (1985) that describe the industry's dynamics.

2.2.3 Innovation drivers of the pharmaceutical industry

To our knowledge, few authors have studied innovation patterns in the pharmaceutical industry and its drivers, in particular in emerging countries. Achilladelis and Antonakis (2001), as well as Ni et al. (2017), however, proposed an analysis of the said topic. They agree with regards to the impact of national environments on the innovation process and describe the different entities involved. Hence, they follow the new literature on national innovation systems (NIS) that describes the systemic nature of innovation (Lundvall 1992). In chapter 3 this literature and several innovation models will be covered, likewise.

Achilladelis and Antonakis (2001) analyzed the dynamics of technological innovation in the pharmaceutical industry from the beginning of the 19th century to 1990 in their paper "The dynamics of technological innovation: the case of the pharmaceutical industry" published in 2001. They identified seven driving forces for innovation in the pharmaceutical industry: scientific & technological advances, government legislation, societal needs, market demand, raw materials, competition as well as company scientific, technological & market specialization. Scientific and technological advances refer to the scientific achievements of R&D organizations, universities, institutes or firms. The scientific achievements thereby influence each other's innovation. This is the typical source of innovation, reflected in the first model of innovation developed in the 1950s, the *technology-push model*. Government legislation influences the innovation process through the aforementioned price regulations and intellectual property rights regulations. In general, due to its nature, the pharmaceutical industry is the most regulated manufacturing industry and therefore exposed to a variety of regulations (Achilladelis and Antonakis 2001). Societal needs influence the innovation process of firms insofar as the societal needs identify prospective fields for innovation. In the course, societal needs, meaning market demand, has to be assessed by marketing departments, so that the high costs of research can be compensated. Scholars have developed a model in the 1960s, the *demand-pull model* that describes innovation to be solely originating from societal needs. Later on this view was reconciled in a more systemic approach, which is descrobed here. the *National Innovation System* approach. Another driver Achilladelis and Antonakis (2001) mentioned was raw materials. They impact the availability and price of material that for their part influence the choice of research venues and possible fields of research. Finally, competition among firms as well as scientific, technological & market specialization of firms influence a firm's innovation process, since firms R&D activities might be constrained to certain fields of research they have a competitive advantage in. More so, competition creates time pressure, fostering an increased pace of innovating. The authors find out that the "intensities of these driving forces and their synergies varied over time and thus determined the rate of technical change. As they were also influenced by the national environment, they were largely responsible for the competitive advantages of national pharmaceutical industries." (Achilladelis and Antonakis 2001: p. 585). However, while up to the 1970s the driving forces were profoundly affected by the national environment, henceforth the national environment's force decreased due to the globalization of the pharmaceutical industry (Achilladelis and Antonakis 2001). Nevertheless, the geography of innovation still plays a role. Countries can implement certain incentives for innovation, as we understood in the case of China. Relationships, however, are becoming more complex with trends of open innovation models, such as outsourcing of R&D, joint ventures, strategic alliances and out-licensing (Gassmann 2008).

Ni et al. (2017) who offer a recent analysis of China's pharmaceutical innovation system, describe the innovation system by identifying different agents (Figure 6).

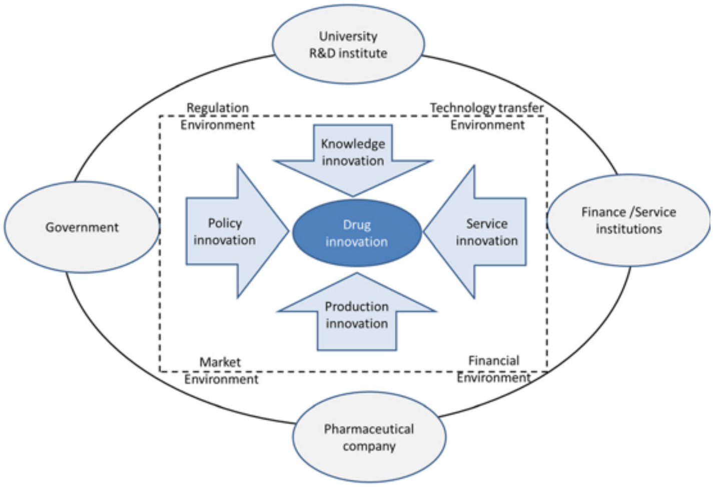


Figure 6: The pharmaceutical innovation system

Source: Ni et al. (2017): p. 2

In general, these agents and their functions overlap with the description of Achilladelis and Antonakis (2001). Ni et al. (2017) explain that:

> "[...] R&D organizations, governments, pharmaceutical companies, finance and service institutions, [are, remark author] responsible for knowledge innovation, policy innovation, production innovation, and service innovation, respectively. These innovations link together and generate

> new medicine discovery under a favorable regulation, market, finance, and technology transfer environments" (Ni et al. 2017: p. 2).

For Ni et al. (2017) the interactive character of the new business model of pharmaceutical R&D is essential. In this vein, they follow the National Innovation System approach that postulates the dynamic character of interactions between entities (Godin 2009; Lundvall 2009). When analyzing the pharmaceutical innovation, partnerships between the entities are integral and foster innovation. Nonetheless, entities can hinder innovation in the system of pharmaceutical innovation. A regulatory system with long approval times can discourage pharmaceutical innovation. Finance and service institutions are equally able to foster but also discourage innovation. Due to the pharmaceutical industry's high R&D costs, proper financing is crucial for efficient R&D. Venture capital and investment levels, for instance, are an essential component of the innovation system. While the US experienced investments in life sciences from around 711 venture capital and private equity funds, China only received investments from around 89 that were mostly one-time investments (Ni et al. 2017). All four actors, the government, the pharmaceutical companies, finance/service institutions and universities/R&D institutes build the market, financial, regulatory and technology transfer environments, in whom pharmaceutical innovation is embedded. Ultimately pharmaceutical innovation is grug innovation for them. But this new drug innovation not only originates from production innovation of a pharmaceutical company but can equally originate or at least be fostered by policy innovation from governments, service innovation from financial/service institutions and knowledge innovation from universities/R&D institutes. This argument emphasizes the interactive nature of innovation and refutes the classic linear models of innovation (Maclaurin 1953).

While Ni et al. (2017)'s concept of the innovation system and its drivers of pharmaceutical innovation is close to Achilladelis and Antonakis (2001)'s concept, they differ in a few points. In contrast to Achilladelis and Antonakis (2001), Ni et al. (2017) mention finance and service institutions, who were not taken into account separately by the former. These institutions are insofar of importance since they designate whether a firm can carry out the research that requires their financing. Moreover, while Achilladelis and Antonakis (2001) agree in general that drug innovation is affected by policy innovation, knowledge innovation, service innovation and production innovation, they explain that knowledge creation is not solely originating from universities and R&D institutes. Knowledge innovation thereby refers to basic scientific research needed for the innovation of products. According to Achilladelis and Antonakis (2001), recently a large part of knowledge innovation derives from pharmaceutical firms. Ni et al. (2017), however, identify universities and R&D institutes as sole sources.

In conclusion, scholars have emphasized that innovation in the pharmaceutical industry is driven by several factors and actors, going beyond science and research or market demand. Due to the system nature of innovation that is based on dynamics between different entities, national environments might have lost some of their importance. They remain, however, a non-negligible factor due to globalization, as is understood in the Chinese pharmaceutical industry who actively encourages its domestic players to innovate.

2.2.4 Global dynamics - performance, consumption and multinationals

The following section describes the significant key players of the global pharmaceutical industry on country- and firm-level enabling us to assess the global dynamics and China's position therein. The identification of key players will be done by looking at the production and consumption of pharmaceutical products, as well as the top innovating pharmaceutical firms.

The top 5 pharmaceutical markets in terms of revenue in 2016 were the US (459,000 Mio USD), China (117,000 Mio USD), Japan (89,000 Mio USD), Germany (44,000 Mio USD) and France (33,000 Mio USD) ("Revenues of the Top 10 Pharmaceutical Markets Worldwide in 2016" 2017). While the US was the country with the highest revenue in 2016 with a clear margin, China was second with around a quarter of the US revenue. China was the only middle-income country among the top 5 markets. When comparing the top 5 pharmaceutical markets to the top 5 producing countries, one can notice that there is a discrepancy. The data does not feature the same top 5 countries. Actually, regarding production output in 2012, Japan, Switzerland, Ireland, Germany and France figure as top 5. China does not figure in the top producing countries; it is, however, a huge customer of pharmaceutical products (Warin 2018). Thus, China has to import pharmaceutical products. In fact, China imported products worth 21,000 million USD in 2016. The Top 5 importers were the US (93,000 million USD), Germany (49,000 million USD), Belgium (35,000 million USD), UK (33,000 million USD) and Switzerland (25,000 million USD) ("Trade Map - List of Importers for the Selected Product in 2016 (Pharmaceutical Products)" 2017). The Top 5 importing countries are congruent with the Top 5 exporting countries ("Trade Map - List of Exporters for the Selected Product in 2016 (Pharmaceutical Products)" 2017). However, while the US imports most with around 93,000 million USD, Germany exports most with around 77,000 million USD. China mainly imported from Germany with products of a value of around 6,000 million USD, followed by the US with 3,000 million USD ("Trade Map - List of Supplying Markets for the Product Imported by China in 2016" 2017). The data revealed that China is an important market for big pharma companies right after the US. The country is however highly reliant on imports of pharmaceutical products since it does not produce enough on its own.

Due to the pharmaceutical market's nature, the market is not that sensitive to external shocks, e.g., a financial crisis ("Global Sector Report - Pharmaceuticals" 2017). Nonetheless, the annual pharmaceutical spending grew in the pre-crisis years of 2.3% each year, and experienced a slowdown in the 2003-2009 period. In the period 2009 to 2015, the spending dropped by 0.5% per year on average across the OECD. 57% of all retail spending in OECD countries was covered by government financing or compulsory insurance schemes so that the governments' pressure for the reduction of public spending after the financial crisis affected the industry. In recent years, however, some markets have experienced a steeper growth of pharmaceutical expenditure, for instance, Germany, Switzerland and the US ("Health at a Glance 2017 - OECD Indicators" 2017). This data excludes "[...] pharmaceuticals consumed in hospitals and other health care settings - as part of an inpatient or day case treatment [...] (data available suggests that their inclusion would add another 10-20% to pharmaceutical spending)" ("Health at a Glance 2017 - OECD Indicators" 2017: p. 186). While only some OECD members' markets are experiencing steep growth again, emerging countries are considered the growth markets the firms have to keep a close eye on. China's sales, in particular, are expected to increase from 66.9 billion USD in 2011 to 175.8 billion in 2020 ("From Vision to Decision - Pharma 2020" 2012). Thus, pharmaceutical multinationals, which mainly originate from advanced economies have an interest in locating their activities in emerging economies, China in particular. For the Chinese economy this is a development they try to exploit.

The global pharmaceutical industry generally consist of a few multinational firms, responsible for the production of branded drugs, and several fragmented generics manufacturer, that mostly operate on the country-level. Some large branded manufacturers, however, started producing generics themselves, as to gain a share of the generics market growth ("From Vision to Decision - Pharma 2020" 2012). Pfizer for instance bought Hospira, a leading generic manufacturer in the US for 16 billion US dollars in 2015, ensuring its share of the generics market ("Generic Pharma Industry Top Acquisitions U.S. 2015" 2018). This is an interesting new venue of revenu for firms since the generic market share in North America has been growing from 52% in 2006 to 70% in 2016. The share in Asia/Australasia was equally high with 71% in 2016, but 74% in 2006 ("Generic Market Share: Change in Unit Volume Worldwide by Region 2006-2016" 2018). Apart from patented drugs and generics, pharmaceutical firms produce active pharmaceutical ingredients (API)s, also. These are required for drug production. Despite the branded manufacturers venturing into the generics market, the development of new drugs remains crucial for the companies growth. Thus, firms invest large sums into R&D.

In 2016, the Top 10 pharma companies in terms of R&D spending originated all from advanced economies, as is illustrated by the following figure (Figure 7). While Merck, Roche, Novartis, Pfizer, Sanofi, Johnson & Johnson, and GlaxoSmithKline persist in the Top 10 firms in terms of prescription sales as well, AbbVie, Gilead Sciences and Amgen replace the others. All Top 10 firms either in terms of R&D spending or prescription sales are from advanced economies. In fact, no Chinese firm is listed neither in the Top 10 nor the Top 50 pharmaceutical firms in 2016 ("Top Pharma Companies by Rx Sales and R&D Spending 2016" 2017).

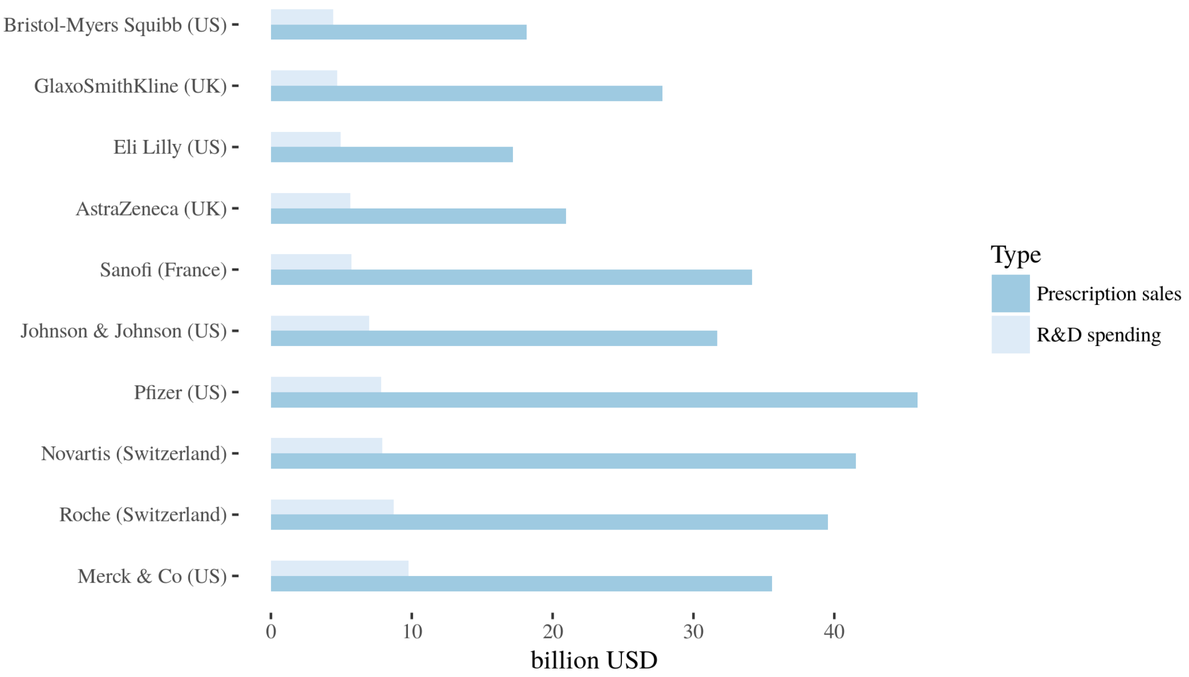


Figure 7: Top 10 R&D spending pharmaceutical firms in 2016

Source: “Top Pharma Companies by Rx Sales and R&D spending 2016”, 2017

However, when looking at revenu, Chinese firms are part of the top 100 pharmaceutical and biotech firms in 2015 (“Top 100 Pharmaceutical & Biotech Companies (Global)” 2015). The Chinese firms related to the ISIC code 2100 *Manufacture of pharmaceuticals, medicinal chemical and botanical products* are: Huadong Medicine Co., Ltd. (rank 60), Yunnan Baiyao Group Co. Ltd (rank 63), Guangzhou Baiyunshan Pharmaceutical Holdings Co. Ltd (rank 66), Harbin Pharmaceutical Group Co. Ltd (rank 73), Tasly Pharmaceutical Group Co. Ltd (rank 81), Shanghai Fosun Pharmaceutical (Group) Co. Ltd (rank 84), Humanwell Healthcare (Group) Co. Ltd (rank 89), Jiangsu Hengrui Medicine Co. Ltd (rank 94), CSPC Pharmaceutical Group Ltd (rank 95) and Zhejiang Hisun Pharmaceutical Co. Ltd (rank 98) (“Top 100 Pharmaceutical & Biotech Companies (Global)” 2015). Thus, although Chinese firms do not figure in the top 50 of R&D spending pharmaceutical firms in 2016, they are a significant part of the top 100 pharmaceutical and biotech firms in 2015 in terms of revenue. Hence, it is of importance to analyze the Chinese pharmaceutical industry closer. While in terms of innovation, indicators are conflicting with regard to the extent of the Chinese importance, as we can see with the number of patent applications and top R&D spending firms, it is generally understood that in case of the pharmaceutical industry China has become a player on the global pharmaceutical market.

2.2.5 China’s pharmaceutical industry

While overall, emerging economies are responsible for the pharmaceutical industry’s projected growth, China is considered the biggest growth “pharmerging” market according to IMS Health (Huang 2012). Notwithstanding its importance, the market is subject to challenges. Foreign and domestic firms have to consider these challenges but also its opportunities when doing business in the Chinese pharmaceutical market.

There exist several entities in China that are crucial for the pharmaceutical industry. The aforementioned intellectual property rights and patent laws constitute some of these entities. As has been understood, the Chinese patent law has seen improvements and is currently close to US law, constituting an opportunity for the industry’s growth. The same goes for the China Food and Drug Administration, responsible for the regulation of food and drug safety (“CFDA” 2017). Since 2013 this institution streamlined the regulation processes under the State Council of the People’s Republic of China. Established at the beginning of the 2000s as the State Food and Drug Administration (SFDA), the regulatory body rendered the industry more transparent and standardized procedures (Chervenak 2005; Zhang and Deng 2008). Pharmaceutical companies operating in China have to pay close attention to the CFDA. *The Drug Administration Law of the People’s Republic of China*, *the Regulations for Implementation of the Drug Administration Law of the People’s Republic of China*, *the Regulations for the Supervision and Administration of Medical Devices* and *the Regulations on Administrative Protection for Pharmaceuticals* have to be respected by domestic and foreign firms in China alike. The CFDA recently applied for membership to the International Council for Harmonisation of technical Requirements for Pharmaceuticals for Human Use (ICH), and henceforth figures as one of eight regulatory authorities and six industry associations across the globe, such as the US FDA, that opt for efficient, standardized drug development practices (Mullin 2017). As a consequence, this international cooperation ensures the development and registration of safe, effective and high-quality pharmaceuticals in China. These recent developments with respect to the CFDA constitute an opportunity for local and foreign pharmaceutical firms alike since procedures became more transparent and harmonized.

Apart from that, the Chinese pharmaceutical industry is facing several other opportunities. As mentioned before, China is an essential market of growth for the pharmaceutical industry, since sales are expected to grow from 66.9 billion USD in 2011 to 175.8 billion in 2020. The most significant part of sales is expected to originate from sales of generics since consumers pay drugs out of their pocket (“From Vision to Decision - Pharma 2020” 2012). It has been understood that China is promoting innovation in high-technology industries and this applies to the pharmaceutical industry as well. “The government is dedicated to provide a platform for domestic pharmaceutical companies to foster original innovation, nurture high technologies and transmitting pharmaceutical industry into the higher valued segments of the global value chain” (Huang 2012: p. 125).

Apart from the government actively pushing for the industry’s growth, China offers a cost-advantage with respect to several factors. Low costs for labor, clinical trials, and raw materials are attracting foreign manufacturing pharmaceutical firms. China is distinguished by a large number of western-educated scientists, also. These assure foreign multinationals that scientific standards will be met. In fact, the Chinese government implemented programs since the mid-1990s aimed at Chinese talent abroad. These programs, for instance, the *Thousand Talents Program* introduced in 2008 incentivize Chinese entrepreneurs with patents, executives and scientists to return to China (“UNESCO Science Report: Towards 2030” 2016). Incentives include subsidies of around RMB 0.5 to 1 million and housing provided by the host institution. To be eligible for the program Chinese have to fulfill certain criteria, such as being a full professor from any institution or an associate professor. These criteria changed over time and in 2012 a *Ten Thousand Talents Program* was introduced, trying to improve the weaknesses of the Thousand Talents program ("UNESCO Science Report: Towards 2030" 2016). From then on "young scientists and engineers aged 40 years and under who hold a doctorate from a well-known foreign university, have at least three years of overseas research experience and hold a formal appointment at a well-known foreign university, research institute or company" were eligible as well ("UNESCO Science Report: Towards 2030" 2016: p. 630). While initially requiring the recruits to work full-time at a Chinese institution for five years, the length was reduced to six months and later on even short-term two-month employment was a valid criteria. This change makes one contemplate whether the programs are successful in retaining the recruits in China.

Furthermore, China's high population numbers attract pharmaceutical firms due to the big population for clinical trials (Huang 2012). The population offers a wide disease spectrum and an aging population, also (Ni et al. 2017). Due to the large population and market growth, China-specific diseases can become the target of multinational's R&D activities. The aging population is in need of health care services, including pharmaceutical products. But not only an aging population but an increasing healthcare awareness also are responsible for market opportunities (Ni et al. 2017). On another note, the Chinese government is planning "to develop the national healthcare system and to enhance the Basic Medical Insurance (BMI) coverage from approximately 65% of the population to 90%" (Ni et al. 2017: p. 6). Thus, pharmaceutical expenditure is expected to increase. However, the government is encouraging generics, not branded-drugs (Chervenak 2005). Hence, global branded-name manufacturers could be struggling. Despite encouraging generic drugs, the government is nonetheless fostering the development of indigenous innovation. Thus, generic drug manufacturers are increasingly venturing into R&D (Zhang and Zhou 2017). Innovation efforts most likely focus on finding treatment for liver disease, gastric cancer and cardiovascular diseases (Ni et al. 2017; Zhang and Zhou 2017; Shen 2010). The latter is constituting the leading cause of death in China (Shen 2010). Furthermore, diabetes is expected to gain attention from a research perspective, since almost 10% of the population are affected (Shen 2010). China is increasingly affected by the so-called "diseases of affluence" (Shen 2010). Apart from these diseases, the growing biotech sector in China is expected to contribute to the development of innovative medicines (Zhang and Zhou 2017). Chervenak (2005) explains that the biotechnology firms in China focus on the areas of gene therapy, antibodies, and TCM modernization. The latter including "the HTS of herbal preparations and other techniques aimed at the discovery of active components, [which; remark author] is thought by some to be a competitive advantage for China's drug discovery." (Chervenak 2005).

With regards to R&D investment, China has experienced growth at times, when other countries reduced their expenditure due to the economic recession. While the government supports the industry in the form of subsidies, for instance through the *Key Drug Innovation* project from 2007, that granted subsidies to research (Ni et al. 2017), it undertakes large investments in infrastructure that improve the distribution of pharmaceutical products, also ("A Closer Look at the 13th Five-Year Plan" 2016). Pharmaceutical innovation is, furthermore, making efforts to leverage the traditional Chinese medicine (TCM) and the distinct biological resources of China (Ni et al. 2017).

Finally, an advantage for domestic and foreign firms alike is that the Chinese Academy of Sciences is trying to foster collaboration between firms, universities, R&D institutes and administration as well as accelerate research. In fact, domestic firms and MNEs increasingly cooperate in terms of R&D. This reflects a general trend in the industry towards innovation networks and open innovation models (Chesbrough 2003; Gassmann and von Zedtwitz 1999).

Overall, the Chinese industry is facing a lot of opportunities making it so promising for MNEs and for the local economy's growth. Nonetheless, it is confronted by several challenges. Huang (2012) states that the industry is challenged by foreign competition with other developing countries, India in particular, as well as public distrust. Moreover, although close to knowledge and innovation capabilities of Western countries in new domains such as biopharmaceuticals, the country is still catching-up in older domains. In addition, firms in China have experienced several issues when developing drugs. They face problems with both the financing of drug innovation and with drug marketing. Indeed, only those pharmaceutical manufacturers allowed to produce drugs in China were granted the right to market these drugs in China (Ni et al. 2017). Finally, an understaffed Center for Drug Evaluation (CDE) resulted in prolonged drug approval times. While the standard time was 12.3 months, it could extend up to an indefinite time (Ni et al. 2017).

In conclusion, the pharmaceutical industry in China is characterized by several opportunities that firms do not seem to want to miss on out. Nevertheless, China is struggling with challenges proper to an emerging market: the quality of innovation is questioned, and catching-up in regard to some domains has yet to be achieved. Overall, however, the industry is growing and increasingly counting on innovation. The Chinese government is playing an important role in this development, actively encouraging first foreign R&D, then domestic innovation, while updating their regulations to comply with the American and European counterparts. The panorama of this thesis has shown that China has become an innovator, and so for pharmaceutical innovation. Nonetheless, while learning a lot on the government's efforts in that regard, less is known on the dynamics and specificities of this innovation. In the following we will consult the topic from an academic perspective, consulting the relevant literature that explains China's development.

3 Theoretical Framework

3.1 Algorithmic systematic literature review : Innovation in the pharmaceutical industry

We performed a systematic algorithmic literature review, as to facilitate the identification of the relevant literature streams for this thesis on innovation dynamics in the pharmaceutical industry in China. Thus, this Data Science approach lays the groundwork for the development of the theoretical framework. It has allowed to identify whether China has surfaced as a topic in the literature on innovation in the pharmaceutical industry or whether there exists a gap in the literature. The computational Data Science approach enables us to "read" all articles relevant to the research topic and to identify the core concepts and streams that are relevant. Ultimately the most important papers will be identified and read, based on this prior identification. Consequently, Data Science allows to certify the chosen streams in the literature that are relevant when analyzing innovation in the pharmaceutical industry.

With regards to the operationalization of the algorithmic systematic literature review, 1458 academic articles were collected from Web of Science. While it would have taken a lifetime to read all these articles, the Data Science approach allowed us to "read" them in a few minutes[2]. The terms selected for the topic search box of the academic aggregator were ("pharmaceut*" AND ("innovate*" OR "R&D" OR "research and development" OR "patent" OR "IPR" OR "intellectual property right*")). The selection of keywords ensured that papers mentioning not only innovation but also R&D or intellectual property rights were included (OECD/Eurostat 2005). Moreover, several spellings were included by using the * that takes into account several endings of a word. Finally, the spelling out versions of R&D or IPR were considered as well. The systematic literature review considers all years available on Web of Science, as of October 30th, 2017 for literature in economics, management, and business.

Using vosViewer, a software developed to analyze bibliometric data, an overview of the field of innovation in the pharmaceutical industry was created (Waltman and van Eck 2013). The dataset considered keywords present in at least three documents. Keywords that replicated the overall topics of research (i.e., pharmaceut*) were omitted. The methodology of van Eck et al. (2006) was replicated and produced the following figures (Figure 8 and Figure 9).

[2] The algorithmic systematic literature review was possible thanks to the protocols put in place at CIRANO.

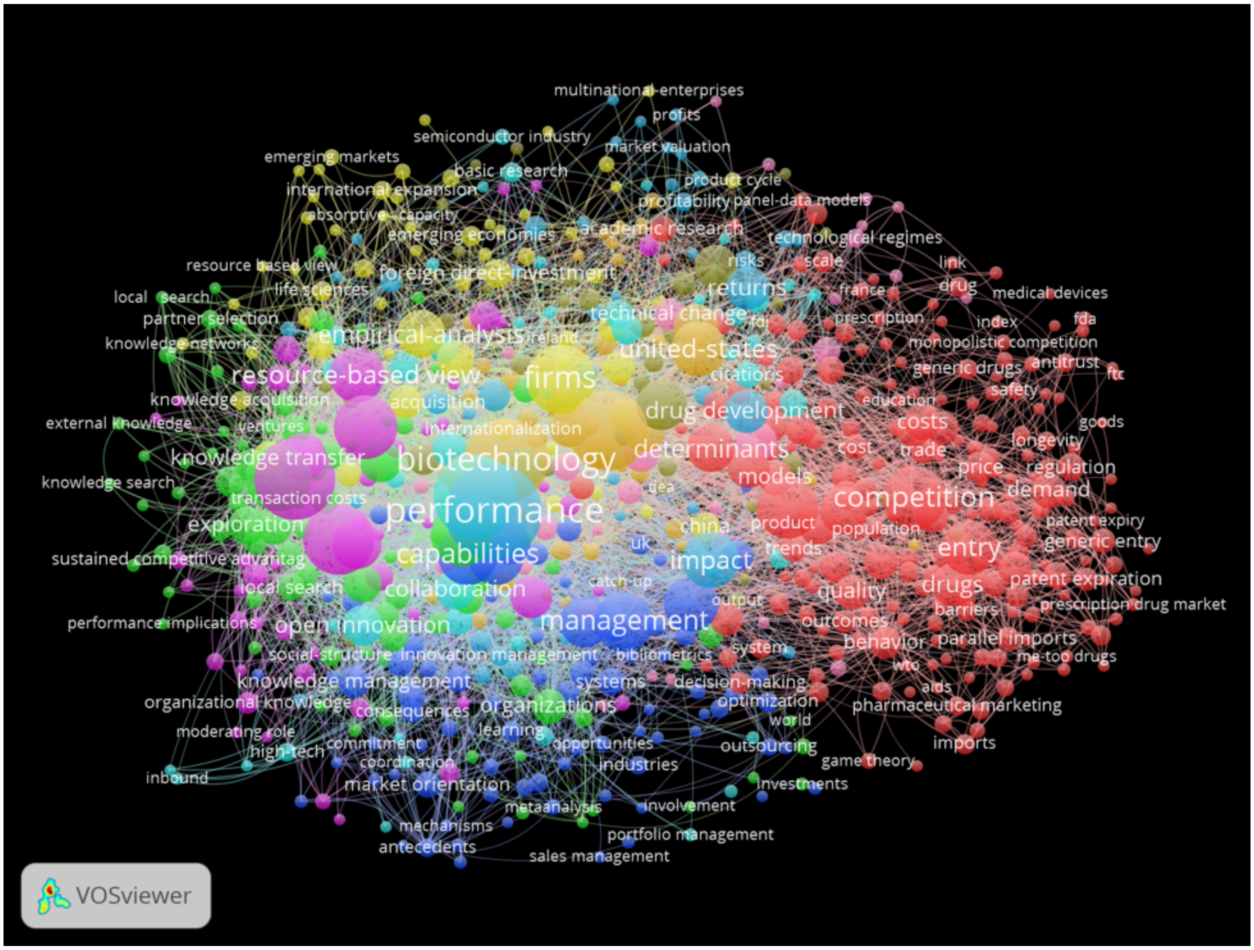


Figure 8: Cartography of the keywords extracted from the systematic literature review on innovation in the pharmaceutical industry

Figure 8 illustrates that the academic literature clusters around different streams in the literature, these are:

- industrial organization,
- MNEs & specific industries,
- management literature,
- innovation & knowledge management.

When focusing on the different clusters, one can identify a group of keywords on innovation networks. It discusses among others “collaboration”, “open innovation”, “partner selection” and “knowledge transfer”. The left side of the cloud contains the literature on innovation networks. It is fairly interlinked between different clusters identified by the software. In contrast, the right-hand side of the cloud represents the literature on industrial organization and is less interlinked (in red). This part of the cloud centers around “competition”, “regulation”, trade-related keywords and intellectual property rights. Finally, we learn that the US has gained a lot of attention from scholars, which is not surprising, considering that it traditionally lies at the center of the global pharmaceutical innovation map (Hadengue, de Marcellis-Warin, and Warin

2015). In contrast, emerging markets and China, are mentioned but have not gained as much attention by scholars. China lies however at the center of the cloud, indicating its involvement with different streams.

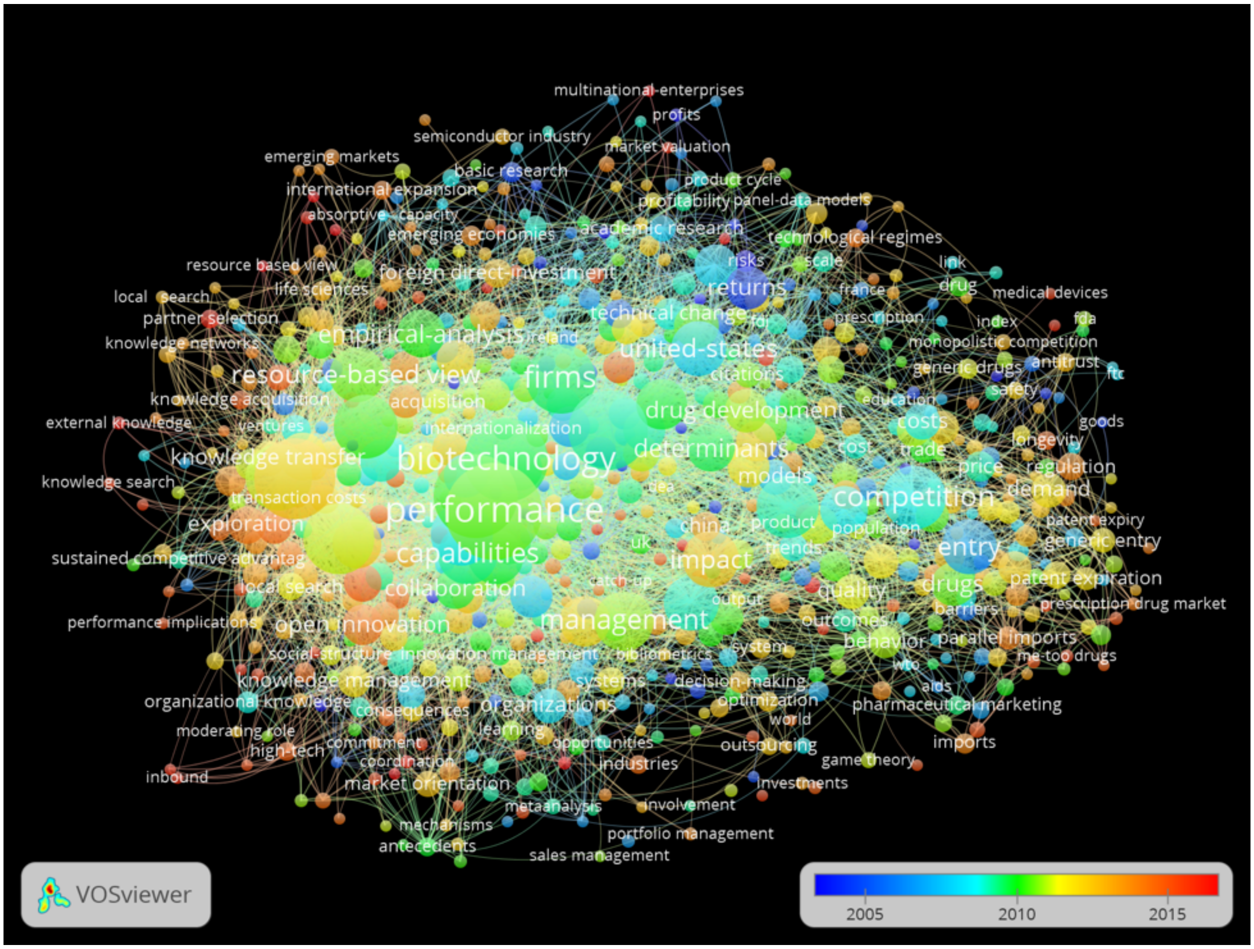


Figure 9: Cartography of the keywords extracted from the systematic literature review on innovation in the pharmaceutical industry, color-coded by year of mention

Figure 9 represents the same cloud of keywords that are now color-coded according to the year of mention. It illustrates that recent literature has been focusing on innovation networks. Moreover, the cloud reveals that the recent literature is located at the edge of the cloud and due to the small size of the circles it is not as prominent yet, with a few exceptions. In addition, China is disclosed as being a relatively new topic in the literature, as well.

To summarize, the algorithmic systematic literature review highlights the relevance of the Chinese focus when looking at pharmaceutical innovation. The method of using algorithms to verify the validity of the literature review is promising, enabling researchers to either check the validity of their literature review and research question in a simple way or to facilitate the identification of relevant streams in the literature.

From the algorithmic systematic literature review, we learned that the analysis of innovation in the pharmaceutical industry and China requires us to consult different streams of literature. Generally speaking, the theoretical frameworks evoked will first focus on economists and sociologists who explain the importance of innovation for economic development and try to describe the process of innovation. This constitutes groundwork that has to be taken into account when analyzing innovation dynamics in a developing country such as China. It explains the different efforts explained in the panorama that the Chinese government has implemented. Such as tax reliefs or high-tech parks. More specifically, we will touch upon theories on innovation and economic growth, innovation models as well as the concept of National Innovation Systems (NIS). This part contextualizes what we have learnt from the panorama from an academic perspective. Moreover, the literature on reverse innovation as well as the open innovation concept from innovation management scholars will be evoked, two reasonably new concepts that are of importance in the context of innovation in a developing country, since they might explain increasing innovative capacities (e.g., Chesbrough 2003; Govindarajan and Trimble 2012).

Then, the work of international business scholars, looking into R&D internationalization and innovation networks will be evaluated. These are of interest, since Chinese policies have shown that the country relied on foreign R&D, to begin with (Gassmann, Beckenbauer, and Friesike 2012). Indeed, scholars attribute R&D internationalization and the intertwined knowledge spillovers the power to increase innovation capabilities in developing countries (Gassmann, Beckenbauer, and Friesike 2012). Focusing on foreign MNEs' R&D internationalization, global innovation networks are created that connect different actors across countries (e.g., Boutellier, Gassmann, and Zedtwitz 2008; Hu et al. 2015; Pilat et al. 2009). These connections, however, can be of a different nature. The literature on knowledge flows/pipelines agrees that innovation is nowadays mainly a transnational process (e.g., Lundvall 2007). Most importantly, this part highlights the benefits of knowledge flows and leads towards our proposition that knwoledge flows across patent thematics, represented by the similarity measurement, have a positive impact on the patenting behavior.

In the following, we will discuss the different concepts relevant for this thesis on innovation dynamics in the Chinese pharmaceutical industry and explain critical empirical studies.

3.2 Innovation: developing countries moving towards innovative capabilities

Several scholars have attempted to define the term innovation; this thesis will revive the proposed definitions by the literature since it is crucial for the understanding of the context. The term innovation which has seen new forms such as open or reverse innovation in the last decade, started its conceptualization back in the 40s where several authors had a try at defining it. Schumpeter (1934) was one of the first trying to define innovation. He explained that innovation is "the commercial or industrial application of something new - a new product, process or method of production; a new market or sources of supply; a new form of commercial business or financial organization" (quoted from Suroso and Azis 2015: p. 387). In general, the vast literature agrees to distinguish between both invention and innovation (Schumpeter 1934; Freeman 1982). In effect, an invention did not even insinuate an innovation for Freeman (1982), since innovation implied a commercial transaction for him. Roberts (1988) agreed with Freeman on that. All in all, they concurred that innovation does not have to be a product, but can take the form of a process or method.

Over the years, the literature has started to distinguish between different types of innovations, for instance, product and process innovations (e.g., Edquist, Hommen, and McKelvey 2001) or incremental and radical innovation (Freeman 1982).

In what follows, we will consider the literature on Economics of Innovation, since it explains the recent developments in China and helps us explain the relevance of this thesis for the international business field.

3.2.1 Innovation and Economic growth

A considerable amount of literature explores the relationship between innovation and economic growth. Scholars advocating the importance of innovation for economic growth have developed so-called “innovation-based” growth models, wherein two branches are being distinguished. The Schumpeterian model (Schumpeter 1934) and the product-variety model of Romer (1990). As so-called “innovation-based” growth models both advocate the importance of innovation for economic growth and thus the move towards a knowledge-based economy.

The Schumpeterian model

One of the most important scholars that put forth the role of innovation in economics is the Austrian economist Schumpeter. Numerous scholars have been referring to him over the years. Schumpeter (1950) criticized the neoclassical approach to economics not being sufficient. In his view, innovation is the driver of economic development: “The fundamental impulse that sets and keeps the capitalist engine in motion comes from the new consumers’ goods, the new methods of production or transportation, the new markets, the new forms of industrial organization that capitalist enterprise creates.“ (Schumpeter 1950: p.83). Innovation thereby not only refers to original innovation but also to innovation it induces in”imitators” that improves the original innovation. This process fosters economic development (Schumpeter 1934). Thus, China’s economic development and growth is linked to its innovative capabilities, explaining the government’s efforts to first attract foreign R&D and then promote domestic innovation.

Another critical postulate of Schumpeter is his belief that innovations tend to cluster in certain industries and time periods. This postulate can be explained by the concept of “creative destruction” that Schumpeter (1950) put forward. Derived from Marx, the term refers to the proposal that, in a capitalistic system, innovation is a disruptive force, creating economic growth while destroying the old economic structure and incessantly creating a new one. This is a crucial distinction from the product variety model. Furthermore, it clarifies Schumpeter’s position towards the patenting of innovation that he considered necessary for the safeguarding of an investment in an economy whose structure is uncertain (Schumpeter 1950). Along similar lines, Kline and Rosenberg (1986) argued that the economic significance of innovations changes throughout time since innovations change themselves. Achilladelis and Antonakis (2001) refered to Schumpeter’s notion of “creative destruction” in their analysis of the dynamics of technological innovation in the pharmaceutical industry, explaining that most firms while being exposed to creative destruction in the industry have been able to adapt to the new economic structures so far and could even prosper from these changes.

With regards to market competition and innovation, Schumpeter (1934) put forward that market concentration is beneficial for the creation of innovation since an innovative environment requires large internal R&D

resources and abundant capital. In a competitive environment, firms are facing market uncertainty and hence innovate less. These hypotheses have been ground for several studies of theoretical and empirical nature (Aghion et al. 1998; Dasgupta and Stiglitz 1980). Aghion and Griffith (2005) dedicated the book "Competition and Growth – Reconciling the Theory and Evidence", published in 2005, to this debate. They revealed that the theoretical literature tends to adopt Schumpeter's view, while empirical pieces of work associate competition with greater innovation output. Given this discrepancy, they elaborated on the Schumpeterian model by distinguishing between pre-innovation rents and post-innovation rents. Propounding an "escape competition" effect or "rent dissipation" effect for low initial levels of competition and the Schumpeterian effect for higher levels of product market competition, they revealed that evidence pictures an inverted-U relationship between innovation and competition. Meaning that at low initial levels of competition firms that have not been innovative have strong reasons to innovate and hence become innovating firms. In contrast, at high initial levels of competition, sectors that have not been innovative experience low incentives that retain them in the non-innovating state for a long time (Aghion and Griffith 2005). Hence, the relationship is positive at low initial levels of competition but becomes negative at high initial levels of competition. This finding applies to the macro as well as industry level. With regards to the firm level, however, increased competition increases the technological gap between leaders and followers. While neck-and-neck firms are encouraged to innovate, laggard firms are discouraged (Aghion et al. 2005). Aghion and Griffith (2005)'s elaboration of the Schumpeterian model could explain why the Chinese government incentivizes its domestic firms to innovate in an increasingly competitive environment.

The product-variety model

Like Schumpeter (1950), Romer (1990) criticized the neoclassical approach to economics. Therefore, he developed the product variety model that extends neoclassical models by highlighting the importance of human capital in the development of economic growth. It construes that human actions result in technological change that motivates the continued capital accumulation. Both technological change and accumulated capital explain an increase in productivity, defined as output per hour worked. The model owes its name to his explanation that innovation creates new products that in turn generate productivity growth. Finally, he stated that this model "is essentially the one-sector neoclassical model with technological change, augmented to give an endogenous explanation of the source of technological change" (Romer 1990: p. 99). Among others, Romer (1990)'s paper initiated the field of endogenous growth theory. While an improvement in contrast to exogenous growth theory, the model saw limitations due to the neglect of obsolesce of products.

Ultimately, Schumpeter's model, as well as the product-variety model term China's efforts sound, since a new doctrine *innovation economics* was born. This new doctrine refuted the neoclassical economists' argument that economic growth resulted from capital accumulation. Instead, it explained economic growth by innovation in the case of knowledge-based economies. Most likely, the increasing competition with other developing countries, with cheaper manufacturing locations, incited the Chinese government to foster innovation and the move towards economic growth through innovation.

3.2.2 North-South approach

As explained, the relation between innovation and economic growth is of interest due to China's efforts to move towards a knowledge-based economy. In fact, authors have put forward innovation-based growth models when exploring the differences in cross-country growth. For instance, Posner (1961) explained that two sources, innovation, and imitation, the former enhancing the difference in economic growth and the latter decreasing the difference, can explain these differences in economic and technological development. The former difference is often referred to as the technological gap between countries or the north-south discrepancy (Fagerberg 1994; Fagerberg 1996). According to his rationale, northern countries were said to innovate, paying high wages, while the South was said to imitate exploiting low wages. In order to stay ahead of the South, the North was obliged to continue innovating and to adapt its pace to the South's efforts of catching-up. Several models emerged thanks to Posner's rationale. Vernon (1966)'s theory of the product life cycle can be invoked in the context of describing innovation dynamics between the North and South, also. His theory conjures technology transfers from innovating countries to those with large markets and/or a low-cost advantage since the standardization of products entails that process innovation, economies of scale and low material costs gain importance in sequence. At this stage, less developed countries might offer competitive advantages as production location (Vernon 1966). Thus, this theory offers a potential explanation for how less developed countries, like China, were able to create their capabilities to innovate. In the same vein, Fagerberg and Verspagen (2002) have shown that southern countries are able to surpass the sole state of imitating and can grow by innovating. Looking at the Asian newly industrialized countries (NIC), they explained that this rapid growth could be attributed to the scope of diffusion before 1983, but was thereafter replaced by innovation-induced growth (Fagerberg and Verspagen 2002).

In conclusion, the theoretical literature, as well as empirical works, show how innovation can explain and can overcome cross-country differences in economic development. Similar to the NIC, China is moving towards innovation-induced growth. While formerly counting on the low-cost advantage, the country is facing competition from other southern countries and thus likely catches-up with northern countries.

3.3 Models of Innovation

Since the introduction of Economics of Innovation, several scholars have set out models of innovation explaining the process of innovation. Thereby scholars retained in part a macroeconomic perspective, focusing on the role of the state and innovation's importance for economic growth, but also adopted a firm-specific view because of its growth being equally related to innovativeness (Achilladelis and Antonakis 2001). We have learnt that the Chinese government has played a significant role in fostering innovation from firms through its regulations. Nonetheless, the state is not solely responsible for innovation, thus the following chapter will provide an overview of innovation models that explain the process of innovation firms adopt. While these models can be employed at all times, certain years saw predominant models. The dominant perceived models will be presented in the following, according to the classification of Rothwell (1994) that is dominant in the literature. In general, scholars focused on linear models from the 1950s to early 1970s, characterizing innovation as a linear process with sequential stages, either putting importance on R&D or the marketplace, but they soon shifted towards multi-dimensional models, such as coupling models, that took into consideration feedback loops between the different stages in the innovation process as well as a multitude of factors. More specifically, Rothwell (1992) identified five generations of innovation models, starting with the linear models of technology-push from the 1950s to mid-1960s and the demand-pull model from the late 1960s to the early 1970s. The multi-dimensional models follow these: the coupling models from the early 1970s mid-1980s, the integrated model from the mid-1980s to the 1990s and finally the system and network model from the 1990s on (Rothwell 1993).

3.3.1 Linear model

The model perhaps best known in innovation economics is the linear model that postulates a unidirectional linear process of the creation of innovation (e.g., Ames 1961; Maclaurin 1953; Mansfield 1968; Rogers 2003; Schmookler 1966). Its common three stage form, from basic research to applied research to development, came up in the mid 20th century. Economists added the production and diffusion to the model, as a fourth stage in the 1950s. In the mid-century, this original model came under criticism from sociologists and economists that argued in favor of introducing the factor need or demand into the model. While shifting the focus on another factor, the new model upheld the perspective of a single explanatory factor. The juxtaposition between the former original "technology-push" and the latter following "demand-pull" model came into existence (Godin 2006; Godin and Lane 2013).

The technology-push model

One of the first economists to develop a theory on the process of innovation was Maclaurin (1953) who developed Schumpeter's ideas. He analyzed technological innovation as a process composed of several stages thereby highlighting the importance of research. His theory was later called the linear model of innovation. The five stages he identified are called (1) "the propensity to develop pure science", (2)" the propensity to invent", (3) "the propensity to innovate", (4) "the propensity to finance innovation" and (5) "the propensity to accept innovation" (Maclaurin 1953: p. 98). These steps were chosen due to their measurability. In the 1960s several models of innovation emerged. Ames (1961), elaborating on Schumpeter's model agreed with

Maclaurin on the continuous flow of information and its linear properties. It is this continuous flow that can induce changes in the economy. Despite drawing on Schumpeter (1950)'s model, Ames (1961) did not agree with Schumpeter (1950) on the source of innovation. In contrast to Schumpeter (1950) who identified entrepreneurs as the source of innovation, Ames (1961) refuted this static distinction between managers and entrepreneurs. According to him, managers could become entrepreneurs and hence contribute to the creation of innovation. Moreover, he elaborated that innovations no matter if small or big both contribute to economic growth. Nonetheless, he was skeptical of the term innovation that seemed to be applied too easily according to him (Godin 2006). Apart from Maclaurin (1953) and Ames (1961), there are other economists, such as Mansfield (1968) "that defined innovation as a sequence from research or invention to commercialization and diffusion" (Godin 2006: p. 657). The literature on the linear model is however not reserved for economists, the sociological literature on the diffusion of innovation contributed to its development as well. For instance, Rogers (2003) book "Diffusion of Innovations" distinguished four stages in the process of innovation: innovation, communication/diffusion, consequences on the social system, and consequences through time (Rogers 2003). He later integrated needs/problems in his initial model (Rogers 2003), hence integrating the demand-pull model's rationale.

In conclusion, several economists and some sociologists in the 1960s developed linear models adopting a technology push approach that postulate that innovation stems from scientific discoveries and go through distinct stages in the process of innovation. While at first glance this might seem like the model pharmaceutical firms adopt, this model would not hold up, when thinking of incremental innovations that improve existing innovations. Moreover, we have learnt that the Chinese pharmaceutical industry is expected to focus on topics such as diabetes, since this is a prominent disease in China (Ni et al. 2017; Zhang and Zhou 2017; Shen 2010). Thus, innovations originating from this disease troubling the Chinese population cannot be considered as originating from scientific discoveries. Researchers realized this fact and thus came up with a new model, the demand-pull model.

The demand-pull model

While in some cases, innovation stems from scientific breakthroughs, firms also draw on existing knowledge, combining it with each other or with new aspects to cater to a new commercial need. Hence the process of innovation can be incited not solely by scientific achievements but by a commercial need the firm discovers, also. This perspective came up in the 1960s. Policymakers began emphasizing the importance of market needs. Moreover, since competition became fiercer in the 1960s, firms began to attach importance to marketing and the rationalization of technological change (Rothwell 1994).

Scholars citing the demand-pull model identified Schmookler (1966) as the main representative of this model in the economics literature (e.g., Rogers 2003; Godin and Lane 2013). Drawing on US patent data records and important inventions throughout the world since 1800, Schmookler (1966) refuted the common belief that science and research induce inventions. According to him, economic factors were of importance in the process of innovation. Indeed, the data yielded by his study strongly supported that it is a technical problem or opportunity that predominantly had acted as a catalyst. In science-based industries, however, it sometimes originated from scientific discoveries. Going a step further, he explained that often inventions themselves are a stimulus for other inventions, more than science. Moreover, the study found that "the number of capital goods inventions in different industries tends to be distributed among them in proportion to capital goods sales" (Schmookler 1966: p. 206). In other terms, the number of inventions correlated with investment in the previous year. While there are inter-industry differences, these can be explained by "the combination of a richer knowledge base underlying the product technologies and possible shifts in the characteristics desired in inventions" in some industries (Schmookler 1966: p. 212).

Along similar lines, Kline and Rosenberg (1986) discussed market demand not only fostering innovation but also science. During the creation of products, problems can arise whose relay can initiate science. Thus, "the demands of innovation often force the creation of science." (Kline and Rosenberg 1986: p. 287). In their empirical study on companies in the UK, Carter and Williams (1957) found for 60 cases that R&D initiated only 25% of them while the rest was from different kinds of demand-pull, such as commercial foresight, a short-term need of the sales department or consumer pressure.

Proponents of the technology-push or the demand-pull model debated strongly on the active or reactive role of R&D and the importance of the marketplace. The juxtaposition takes a prominent role in the literature on innovation economics. Nonetheless, some scholars advocated a complementary approach and the reconciliation of this juxtaposition (Rothwell and Robertson 1973). In case of the pharmaceutical industry, it has been shown that innovations can originate from both scientific discoveries and the marketplace, thus the juxtaposition does not hold up.

Both models have experienced numerous criticism due to their linearity and lack of cooperation between the stages in the innovation process (Rothwell 1994). Achilladelis and Antonakis (2001), criticized that the models were insufficient for the theory of the dynamics of technological innovations in their paper on the dynamics of technological innovation in the pharmaceutical industry. Despite the accord on this significant deficiency, the models are still applied due to their applicability in statistics (Godin 2006). Some scholars that criticized the unidimensionality of the models tackled the neglect of feedback loops and unidimensionality by developing multi-dimensional models.

3.3.2 Multi-dimensional model

Due to the recurring criticism of the linearity and singularity in the factor explaining innovation, scholars developed several multi-dimensional models. These models would reconcile the finding that pharmaceutical innovation can originate from scientific discoveries and from a demand. While called multi-dimensional and reconciling aspects of the different linear models, Godin (2013) criticized that these models often remain linear in their approach as well.

The coupling model/chain-linked model

The coupling models became prominent in the early 1970s, extending the debate to governments and corporate consultants (Godin 2013). Most of all, they dealt with the lack of feedback loops between the stages in the innovation process and integrated R&D as well as marketing to the innovation process (Rothwell 1992). Historically the development of the coupling model is anchored in times of two oil crises where demand stagnated, and firms were forced to understand the factors of successful innovation (Rothwell 1994). As a consequence numerous empirical studies were launched looking into successful innovation, revealing that the linear models were inaccurate representations of the reality (e.g., Rothwell et al. 1974; Utterback et al. 1976). Thus, the coupling model puts forward a more realistic representation of the process of innovation in firms (Rothwell 1992).

The coupling models that had been named differently by different authors included more factors in the process, such as management, marketing, communication, entrepreneurship, and finance (Godin and Lane 2013). Inherent to them was the importance of the flexible links between the different stages. Rothwell and Robertson (1973) for instance put forward the importance of the individual and the communication between departments. According to them, individuals function as gatekeepers. Finally, not only the contact between different departments is of importance but also between "the innovating organisation and its environment, commercial and marketing, as well as scientific and technical" (p. 223). Rothwell (1992) provided the drawing of a coupling model that illustrates these points (Figure 10).

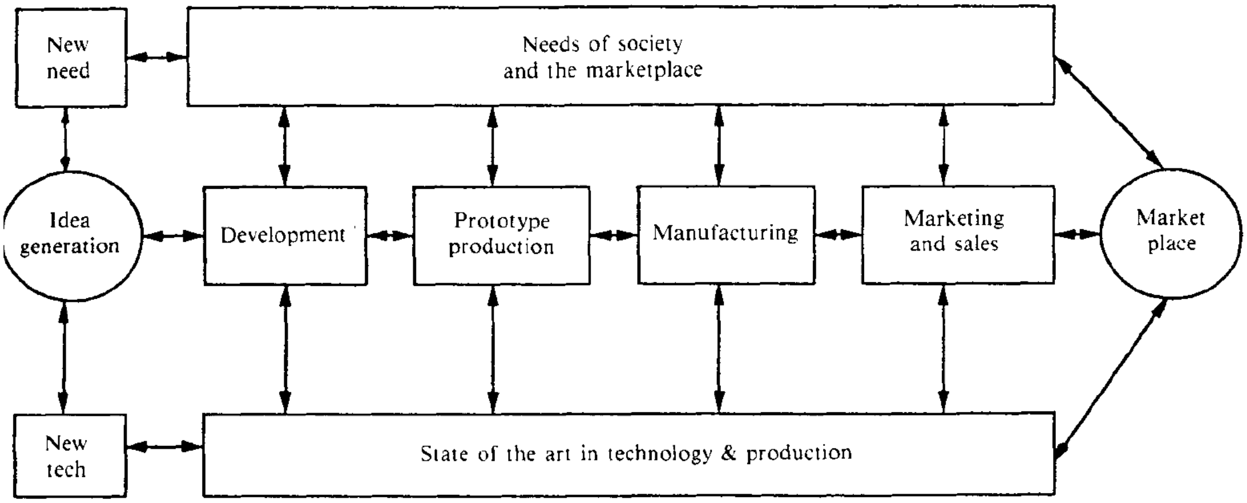


Figure 10: The coupling model

Source: Rothwell (1992): p. 222

As Figure 10 confirms, extra-organizational factors are introduced to the innovation process: the marketplace and the scientific and technological community. New ideas can thereby derive from three different sources, a new need of the society, a new scientific or technological achievement or from the inside of the firm. Thus, the model reconciles the *technology-push* and the *demand-pull* models. The linkages between the different factors in the coupling model are bidirectional, reflecting the feedback nature of the model.

The chain-linked model of Kline and Rosenberg (1986) is another prominent example of a coupling model. In the same vein as Rothwell (1992), they described the typical process from design to development to production and marketing, introducing feedback links. Their model's name relates to the linkages between science and innovation. These linkages extend throughout the process. They specified that each stage in the process thereby requires a different kind of science. For the design or invention stage, for instance, Kline and Rosenberg (1986) identified pure, long-range science. In concordance with Rothwell (1992), the linkage between science and innovation goes so far, as there are links back from innovation or its products to science. Kline and Rosenberg (1986) concluded that most often analytic designs or inventions were the initiating steps for innovation.

While coupling models provided the long-after thought rapprochement towards a more realistic representation, they were still based on sequential stages, although not necessarily continuous, and therefore saw their successive replacement by the integrated model at the end of the 1980s (Rothwell 1992).

The integrated model

Superseding the coupling model is the integrated model, also called the fourth generation by Rothwell (1992). This model saw its emergence in the 1980s at times of economic recovery where firms concentrated on their core competencies and discovered the importance of technology (Rothwell 1994). Technological discoveries brought along the shortening of product life-cycles and thus required firms to speed up the development process by adopting a parallel development approach (Taferner 2017). At times, Japanese companies emerged as innovators whose success could not be explained by imitation. Two specific features could explain their success: the integration and parallel development of the Japanese product development (Rothwell 1994). The Japanese firms had developed the integrated model, replacing the sequential processes. The model set forth a parallel development process with integrated development teams. While some might argue that a parallel development process is not adequate for some sectors such as pharmaceuticals, Rothwell (1994) extended the applicability of this model to them in the form of functional overlap.

Due to the Japanese firms' success, firms adopted their conduct, lean production and horizontal collaboration in the form of strategic alliances or joint ventures, key features of the integrated model. Linkages with suppliers and customers were central, also (Rothwell 1994). Finally, the integration of R&D and manufacturing was emphasized (Rothwell 1992).

While this model can be considered the prominent model in our times, Rothwell (1994) saw the emergence of a new generation of innovation models in the 1990s.

System integration and networking models (SIN)

In the early 1990s, Rothwell (1992) saw the up-spring of a new generation of innovation models, the *systems integration, and networking model* (SIN), which he identified as a model of the future. Continued strategic networking, technological accumulation and the increased speed to market that the fourth generation model was developed in, portrayed the 1990s (Rothwell 1994). The regulatory environment prospered. Thus, the SIN model is similar to the integrated model insofar as it can be considered a "somewhat idealized development of the integrated model, but with added features, e.g., much closer *strategic* integration between collaborating companies." (Rothwell 1992: p. 236“). Rothwell (1992) identified”the *electronification of innovation* as most significant feature of the fifth generation model, characterized by an increased use of expert systems as a developmental aid, simulation modelling partially replacing physical prototyping, linked supplier/user CAD systems as a part of a process of co-development of new products, and closer electronic product design/manufacturing links" (pp. 236-237). Central to the model is the thought that adopters increase their speed of development and gain greater efficiency (Rothwell 1994). Firms with "internal organizational features, strong inter-firm vertical linkages, external horizontal linkages and [...] a sophisticated electronic toolkit" can achieve this (Rothwell 1994: p.15).

All in all the characteristics of the fifth generation model, SIN, are:

- "greater overall organizational and systems integration (including external networking);
- flatter and more flexible organizational structures, including devolved decision making;
- fully developed internal databases;
- electronically assisted product development;
- effective external electronic linkages" (Rothwell 1994: pp. 24-25)

Only the successful handling of the electronically gathered information across the system of innovation, ensures firms to exhaust the benefits of the SIN model.

It is important to mention that while Rothwell (1994) described several generations of models, all types can occur nowadays. Table 2 summarizes these generations of innovation models and provides dominant authors that can be consulted for further information on the models. It is, however, of importance to recognize that the four generations of innovation models solely represent the prominent type at a certain time. Indeed, Rothwell (1994) explains that depending on the type of innovation a different model than the prominent one might be better suited. In case of radical innovation, a parallel process might be unwise in virtue of too many technological uncertainties. Depending on the industry cycle different models of the innovation process might be best applicable, also. Nonetheless, the purely technology-push model should be neglected since the market needs remain central to success at all stages (Rothwell 1994). Finally, sectoral differences can account for the use of different innovation models by firms. In this line of thought, pharmaceuticals tend towards the technology-push model, although Rothwell (1994) suggested that the coupling model would suit it best. It is said that China's population is experiencing a growing number of sick people because of diseases such as cancer and diabetes (Ni et al. 2017; Zhang and Zhou 2017; Shen 2010), and that pharmaceutical firms in China are focusing on these areas for the development of innovation. Thus firms clearly do not adopt a technology-push model for innovation. They seem to consider market demand. Nonetheless, the Chinese regulatory environment is the one fostering innovation proper to China, in the first place, thus these linear models are not adequat when classifying the innovation model used predominantely in China. The multi-dimensional models again, take into account other factors and feedback loops but do not take into account the role of the state in the process of innovation either. But, the state seems to play a role, considering that after the introduction of the *15-year National Outline for Medium- and Long-Term Science and Technology Development Planning* in 2006, which promotes domestic innovation in China, patenting rose considerably. Hu, Zhang, and Zhao (2017) explained that this surge was partly due to firms patenting that were not patenting before, hence we can assume the Chinese state to play a role as initiating innovation in firms. Thus, another approach to innovation has to be put forward, the National Innovation System (NIS) approach, that reconciles the presented firm-strategic models with the role of the government in the process of innovation.

In conclusion, it has been established that firms are adopting different types of innovation models with a trend towards a model that functions as a network and leverages electronic tools. In this regard, the literature on innovation models is in line with the literature on innovation systems and management practices that proclaimed *open innovation* practices of firms since the 2000s. Nonetheless, the role of the government and its linkage to firms has not received attention in these models, which is a crucial aspect in the process of innovation in China. In what follows, we will learn more about the National Innovation System (NIS) approach, that accords a role to national governments in the process of innovation.

Table 2: Models of innovation

Linear Models

Year	**Name**	**Author**	**Description**
1950s	”Technology-Push” model	Maclaurin (1953)	Science and research incites the process of innovation
1960s	”Demand-Pull” model	Schmookler (1966)	A commercial need incited the process of innovation

Multi-dimensional models

Year	**Name**	**Author**	**Description**
1970s	Coupling/ chain-linked model	Rothwell(1992) and Kline and Rosenberg (1986)	Sequential stages from design to development to marketing with feedback links
1980s	Integrated model	Rothwell (1992)	Parallel development process with integrated development teams
19990s	System Integration and networking models (SIN)	Rothwell (1994)	Electronification of innovation, organizational and system integration

3.4 National Innovation System (NIS)

At the same time as Rothwell (1992) described the emergence of the integrated innovation model in the 1980s adopting a firm strategic focus in his studies, the 1980s saw the emergence of a new concept in innovation, the concept of the *National Innovation System* adopting a systemic approach to innovation. This new concept of the national innovation system focuses on the dynamics between entities of a system of innovation, in contrast to mathematical economic models or firm strategic innovation models (Godin 2009; Lundvall 2007). It is of importance when looking at innovation in China, since the Chinese government has encouraged its firms to innovate via numerous programs and regulations. Hu, Zhang, and Zhao (2017) explained that indeed firms seem to innovate more due to these policies. When one wants to understand China's innovation dynamics it is crucial to understand the system innovating entities are embedded in. This does not only include the Chinese government but also the CFDA, for instance, we have learnt about in chapter 2.

Going back to Freeman (1987), the *National Innovation System's* fundamentals were evoked earlier than the 1980s. Godin (2009), for instance, argues that its rationale is inherent to the OECD's work since the 1960s. Nonetheless, the framework under its name *National Innovation System* was born in the 1980s. It arose out of concern that existing economic frameworks and theories neglected "the dynamic processes related to innovation and learning when analyzing economic growth and economic development" (Lundvall 2007: p.96). Moreover, the policy focus on science-based innovation had been exposed as insufficient by studies that showed the interactive nature of innovation (Kline and Rosenberg 1986; Rothwell et al. 1974). Hence, scholars recognized the need to develop an analytic framework taking into account the interactive nature of innovation. While the chain-linked model from Kline and Rosenberg (1986) constituted a step towards the idea of a concept of NIS, the latter embodies a wider perspective. Godin (2009) explains that the NIS approach added new terms to the preexisting systems approach of the OECD, namely: "globalization of research activities, networks of collaborators, clusters, and the role of users" (p. 493). Lundvall (2007) distinguishes in his NIS framework between a core and a wider setting that encompasses the aspects mentioned by Godin (2009). The core setting refers to "firm[s] in interaction with other firms and with the knowledge infrastructure" (Lundvall 2007: p. 102). In contrast, the wider setting refers to "national education systems, labor markets, financial markets, intellectual property rights, competition in product markets and welfare regimes" (Lundvall 2007: p. 102). For scholars, it was essential to try to provide a holistic framework since they assumed the impact of innovation on economic performance being dependent upon people's relations within and across organizational borders as well as upon changes in people (Lundvall 2007). This refers back to the interactive nature of the system approach which is here being described as socially embedded. Finally, the concept saw its strengthening with the popularity of Porter (1990)'s competitive advantage of nations (Godin 2009). Achilladelis and Antonakis (2001) based their identification of the driving forces of pharmaceutical innovation on several frameworks: Porter (1990)'s concept of the competitive advantage, as well as the literature on the national innovation system and linear, as well as multi-dimensional models. The seven driving forces they identified were: (1) scientific and technological advances, (2) raw materials, (3) market demand, (4) competition, (5) societal needs, (6) government legislation, as well as (7) company scientific, technological and market specialization[3].

[3] For a detailed description of the seven driving forces in the pharmaceutical innovation consult chapter 2.

Two currents can be identified in empirical studies (Lundvall 1992), since they have either focused on measuring the performance of NISs (e.g., Samara, Georgiadis, and Bakouros 2012) or comparing and describing them (e.g., Xiwei and Xiangdong 2007). Moreover, the literature has focused dominantly on the analysis of NIS in the North, while developing countries have not received much attention yet (Lundvall 2007). This is however of importance since innovation systems in the South differ to the ones in the North. In the South, systems are being built and not yet strong and diversified as in the North (Lundvall 2007). Although there has been relatively little research done on NISs in the South, some studies focused on NISs in China, or on Regional Innovation systems; the latter is a concept that emphasizes the systemic nature of innovation as well, however, focuses on the regional level. The initial focus on the national level of the system approach is comprehensible since the concept was developed in opposition to neo-classical frameworks that focused on the national level (Lundvall 2007).

Concerning China, the vast literature focuses on the National Innovation System (NIS) in China and its impacts on China's innovation capabilities (e.g., Sesay, Yulin, and Wang 2018; Zhang and Zhou 2015). With regards to economic performance, the literature shows consensus on the importance of the NIS on economic growth in China (Fagerberg and Srholec 2008; Sesay, Yulin, and Wang 2018; Tang and Hussler 2011). In their study on the Chinese NIS and its contribution to the catch up of the Chinese economy, Tang and Hussler (2011) explained that China consists of two complementary innovation systems: the FDI-based innovation system and the indigenous innovation system. Both affect China's catching up to advanced economies positively. Nonetheless, the indigenous innovation is considered less effective in influencing this process, despite government efforts in promoting the latter and departing from their dependence on foreign investment (Tang and Hussler 2011). Thus, this entails that foreign R&D from big pharma companies is crucial for the Chinese economy.

While the NIS framework has the firm level at its core, ample research on the NIS in China focused on the role of the state (Chan 2015; Sun and Liu 2010; Xiwei and Xiangdong 2007). The Chinese government takes a so-called top-down approach to innovation, having introduced several policies fostering innovation by universities, research institutes and enterprises (Chan 2015). While initially, more university and research institute centered, innovation is becoming increasingly enterprise-centered in China (Sun and Liu 2010; Xiwei and Xiangdong 2007). Both domestic and foreign firms are among the innovating firms. Foreign firms contribute through their R&D activities and linkages to local universities (Jin, Wu, and Chen 2011). Jin, Wu, and Chen (2011) characterized these linkages as international university-industry collaborations and explained their contribution to the NIS in their case study on Brainbridge in China. In the context of their study, they raised the question of whether the NIS should not become global due to the international networks. Along similar lines, Altenburg, Schmitz, and Stamm (2008) argued that the National Innovation System approach is not sufficient when assessing the catching-up process of China. Transnational relationships have to be taken into account when one attempts to gain a holistic understanding of Chinese innovation capabilities. Indeed, Lundvall (2007) criticized as well, that the relation between NISs and globalization has not been researched enough, in particular in the context of developing countries that often rely on the adoption of foreign technologies for their development, this is a crucial topic for research. Pietrobelli and Rabellotti (2011), as well as Jin, Wu, and Chen (2011), are one of the few looking at integrating Global Value Chains or R&D globalization into the literature on NISs.

As has been mentioned, few authors have adopted a regional perspective on systems of innovation. Scholars adopting this perspective in China have exposed regional disparities in innovation capabilities and outputs (Crescenzi, Rodriguez-Pose, and Storper 2012; Fan, Peilei, and Lu 2012; Li 2009; Shang, Poon, and Yue 2012).

In conclusion, there is a consensus on the importance of the development of the innovation system for the economic development of developing countries (e.g., Fagerberg and Srholec 2008; Lundvall 2007; Lundvall 2009). Countries that build a strong NIS can benefit economically and catch up to the North. Thus, China's efforts in fostering innovation are understandable. When analyzing innovation the system and environment perspective has to be taken into account. A solely firm-specific perspective would not allow to understand why firms in China innovate more and more. It is left to integrate this concept of the National Innovation System with globalization and firm-level managerial concepts such as the open innovation concept (Lundvall 2007).

3.5 Open Innovation

Before the 20th century, innovation management followed the resource and knowledge-based view of the firm (e.g., Grant 1996). According to this theory, firms were supposed to focus on their core competencies. Diversification should be limited to related technology fields. In this vein, the decentralization of R&D was rejected by scholars of the resource- and knowledge-based-view of the firm (Chesbrough 2006). The 1990s mostly saw representatives of this view, and firms largely adopted a closed innovation model. They relied on the ideas they generated, developed and commercialized inside the boundary of the firm (Chesbrough 2003). This view however changed and firms began opening their innovation process to external entities as well as to the decentralization of R&D. This brings us to a paper published by Cohen and Levinthal (1990) in 1990, where the open innovation model was referred to first to some extent, when discussing the notion of absorptive capacity and explaining that internal R&D not only develops new and improves innovations but also absorbs external knowledge. The external knowledge thereby has to be related to the existing one for the absorptive capacity to have an effect. Otherwise, firms have to make an effort to create absorptive capacity for the new unrelated knowledge, which can be detrimental (Cohen and Levinthal 1990). Albeit not refuting the importance of a firm's core competencies, open innovation increases the return on innovation capabilities for a firm (West and Gallagher 2006). The vital point is that with the emergence of open innovation models knowledge was being transfered across departemental, firm or country-borders. The adoption of this new model might be responsible in part for the creation of China's innovation capabilities.

Indeed, as mentioned, in the 20th century, the concept of open innovation had gained ground since scholars noticed a shift in innovation management by many firms. They observed that "firms commercialize external (as well as internal) ideas by deploying outside (as well as in-house) pathways to the market" (Chesbrough 2003: pp. 36-37). Linkages with external institutions, such as competitors, suppliers, universities, etc., gained ground. These can take the form of licensing, collaborative R&D or crowd-sourcing, etc. The new model benefited from the knowledge of external actors and enabled the faster pace of innovating according to Chesbrough (2017). Due to the increasing competition in the pharmaceutical industry, firms might have started adopting this model ensuring a faster pace of innovating but also the sharing of large R&D investments.

While Chesbrough's (2003) theory of open innovation proclaims the open character of the model and the interaction between internal and external establishments, Pénin (2011) argued that there is a different – more open – type of innovation, which he named open source innovation. It is more open insofar as the knowledge is disclosed to all. Moreover, its nature is more interactive. The open source innovation mostly applies to software firms but has been pushed forward elsewhere, also (Pénin 2011). Most firms, however, are slowly adopting the open innovation model. The pharmaceutical industry was transitioning from the closed model to an open model beginning of the 21st century (Chesbrough 2003; Pilat et al. 2009). Open innovation is in support of decentralized R&D since it allows the sourcing of foreign knowledge. Thus, the concept is closely intertwined with firms internationalizing R&D to access foreign knowledge (Dunning 1998; Kuemmerle 1997). Although offering the advantage of outsourcing the cost-related risk of expensive R&D activities in the pharmaceutical industry, intellectual property right issues arise. The nature of knowledge-sharing implies that intellectual property has to be accessible and thus strong protection might be harmful (Pilat et al. 2009). Nonetheless, it remains necessary to provide innovators with some leverage. The literature shows no consensus on the role of IPR. While some argued that IPR are necessary for knowledge transfer (e.g. Chesbrough 2003; Chesbrough 2006; Wang, Vanhaverbeke, and Roijakkers 2012; West 2007), others criticized that IPR refute the principle of openness (e.g., Pénin 2011; von Hippel and von Krogh 2006). Hagedoorn and Zobel (2015) dedicated an empirical analysis to the analysis of the role of IPR in firm-to-firm open innovation models as well as to the governance mode of the cooperation. Their survey among professionals indicated unanimous that IPR protection is perceived as crucial to the open innovation model. The control over the intellectual property is thereby only one part of the reasoning. Firms consider their IPR as part of their attractiveness factor for potential open innovation partners (Hagedoorn and Zobel 2015). Furthermore, IPRs can encourage the specialization of firms, since they can leverage their IP through open innovation models (Chesbrough 2006; after Lamoreaux and Sokoloff 1999).

On another note, universities are of interest, when considering open innovation models, since firms rely on their scientific publications. However, there is a trend of commercialization and patenting in university research. Thus, this departure from the common open science model could inhibit firms from leveraging university knowledge. Fabrizio (2006) analyzed "the relationship between the growth of university patenting in a technology area and both the exploitation of openly published scientific research in firm's inventions in that technology area and the pace of knowledge exploitation evidence in a firms' patents" on the basis of the pharmaceutical and biotechnology sector (Fabrizio 2006: p. 146). He found out that the patent surge by universities in the US did not restrict the exploitation of its research to only a few firms. The industry does, in fact, rely more heavily on university research. Nonetheless, the pace of knowledge exploitation in industry inventions is slowing down with increasing university patenting. In the case of pharmaceuticals, Cockburn and Henderson (1998) found out that firms that collaborate with universities achieve a greater patent output of important innovations. This finding is in line with the view that geographic proximity to a university intensifies the absorptive capacity of a firm (Zucker, Darby, and Brewer 1998).

While the concept of open innovation is grounded on the assumption that external knowledge has a positive influence on innovation capabilities, Jiang, Branzei, and Xia (2016) pointed out that these are Western scholars and that in case of China, dependence on external knowledge limits indigenous innovation. Finally, NIS play a role in open innovation models. Since the efficiency of knowledge flows in an open innovation model depends on the NSI (Wang, Vanhaverbeke, and Roijakkers 2012).

In conclusion, there is evidence that firms have increasingly adopted open innovation models, leveraging the knowledge from external actors. In return, external establishments have had the opportunity to access internal firm knowledge. Knowledge can pass across firm borders and entails an increased pace in innovation activity since knowledge external to an institution can be exploited. The role of IPR, however, is debatable. While not necessarily linked to economic geography, globalization has seen the emergence of open innovation models across borders. In this vein, developing countries, like China, can create linkages to the Western world. This, in turn, has implications for both sides. All in all, Chinese innovation capabilities are expected to be affected by these developments. Open innovation models might explain the exponential growth in innovation capabilities.

3.6 Frugal and Reverse Innovations

When talking about innovation in the context of an emerging economy, some scholars have focused on the concepts of frugal and reverse innovation (Govindarajan and Trimble 2012; Zedtwitz et al. 2015; Zeschky, Winterhalter, and Gassmann 2014). These concepts are of interest in this paper since they reveal that emerging countries cannot be underestimated in terms of the creation of innovation. Indeed, these concepts are founded on the assumption that emerging economies have certain features different from advanced economies, generating a certain type of innovation that is inherent to them.

Radjou et al. (2012) published a book on Jugaad Innovation, demonstrating how entrepreneurs from emerging economies defy the difficult circumstances in their countries. While Jugaad Innovation, in reference to the Hindi word for an innovative fix, is his term for innovation formed under adverse conditions, the Chinese call it *zizhu chuangxin*. Mostly pursued in emerging markets, it is not limited to them. Indeed, they advocate for the inclusion of the, what they call, “marginal consumer”, for instance, ethnic minorities. As an example of a Jugaad or frugal innovator HTC and Haier are mentioned who both develop low-cost, high-value products (Radjou et al. 2012). Jugaad innovation, also known as frugal innovation and sometimes referred to as innovation for the bottom of the pyramid is an important concept in the context of this thesis considering that we are looking at innovation in China. Indeed, this concept proclaims that frugal innovation is currently inherent to emerging markets and that advanced economies can learn from them. Thus, China would be expected to generate specific innovations according to its proper settings.

The premise that advanced economies can learn from emerging markets is apparent in the concept of reverse innovation. Instead of arguing in favor of frugal innovation as Radjou et al. (2012) do, research on reverse innovation focuses on the flow of innovation from emerging markets or developing countries to advanced economies (Govindarajan and Trimble 2012).

This concept is relatively new in the academic world since companies were following a glocalization approach in the past (Arasaratnam and Humphreys 2013). They adapted products that were initially developed for advanced economies to the context of a more impoverished country. Adaptation thereby took form in the removal of specific costly features. Fundamentals, however, remained the same and could pose problems. Medical devices, which require constant energy supply, for instance, can be considered challenging in environments without reliable energy supply (Arasaratnam and Humphreys 2013).

Prahalad (2006) published a book on innovation for the bottom of the pyramid that can be considered a precursor for the concept of reverse innovation. He explained that the bottom of the pyramid, representing over four billion people living on less than 2$ per day, could be recognized as a private sector market firms can benefit from. While considered a precursor, frugal innovation is not fundamental for reverse innovation. Instead of a concentration on low-income consumers, as with frugal innovation, the spatial context is imperative for reverse innovation (Hadengue, de Marcellis-Warin, and Warin 2017). Hadengue, de Marcellis-Warin, and Warin (2015) developed a study focusing on reverse innovation in the pharmaceutical industry and confirmed global sourcing of knowledge. Emerging countries are increasingly chosen by multinationals that voice intent for reverse innovation. This analysis focused on China’s emergence as the innovation center for pharmaceutical MNEs. While their content analysis of specialized press confirmed the firms’ intent for reverse innovation, it remains unproven whether pharmaceutical MNEs have put it into practice.

To conclude, the concepts of reverse and frugal innovation rest on the assumption that emerging or developing economies have something specific to them that distinguishes their innovation and hence renders it lucrative for local firms but also foreign MNEs that can exploit this specificity through reverse technology transfer. In this vein, the concepts show that innovation has become a thematic for developing countries which cannot be neglected. Both concepts essentially postulate that there is pharmaceutical innovation originating from China, exploiting the country's specific features conducive for innovation.

3.7 Knowledge flows: fostering innovation in developing countries

Taking into account the literature on open innovation and R&D internationalization it is evident that knowledge flows play a role in the creation of innovation. The literature on open innovation and R&D internationalization agrees that firms form inter-organizational networks that transfer knowledge and can span across firm boundaries and borders. Patent citations often measure these ties between entities (Vanhaverbeke 2006). In contrast to this micro-perspective (intra- and inter-organizational knowledge flows), knowledge flows have to be reflected upon from the macro-perspective as well. The macro-perspective adopts a regional or national point of view on knowledge networks, which has to be considered in the context of the present paper on innovation in China. From this perspective, the geographic concentration of firms or establishments in so-called industry clusters is of interest. An extensive literature has emerged on this topic (e.g., Bathelt, Malmberg, and Maskell 2004; Porter 1990; Tallman et al. 2004) and has recognized the importance of knowledge spillovers for industries.

The literature on knowledge pipelines has to be taken into account in this thesis, since it provides indications with regard to the development of China's innovative capabilities in the pharmaceutical industry. Moreover, it is a concept adopted on the firm-, cluster- and country-level but has not been applied on actual innovations. The dynamics of innovation with regard to the exploitation of embedded knowledge in existing innovations or from other research fields is not clear. In the following, the exisitng research topics on knowledge flows will be presented.

3.7.1 Types of knowledge flows

When examining subsidiary-headquarter knowledge transfer and R&D structure, it has been mentioned that firms increasingly collaborate with external R&D partners adopting open innovation models, and thus building intra- and inter-firm innovation networks. While the absorptive capacity of a firm is a condition impacting the knowledge flow (Cohen and Levinthal 1990), it depends on the nature of the knowledge, also. Indeed, knowledge flows vary between tacit and explicit knowledge. While tacit knowledge is unspoken and difficult to transfer, explicit knowledge is detained (Grant 1996). Both types are inherent to the creation of innovation. In the same vein, Jindra, Giroud, and Scott-Kennel (2009) argue that the subsidiary's autonomy and technological competencies influence the technological diffusion via vertical linkages to domestic entities in transition economies.

Finally, scholars distinguish several types of knowledge flows. Simard and West (2006) distinguish deep vs. wide, formal vs. informal knowledge flows focusing on the features of the knowledge flow. In contrast, in the context of intra-firm knowledge flows of MNEs, the classification of knowledge flows of Mudambi (2002) adopts an actor-perspective. He distinguishes:

(a) "Lateral flows (e.g., subsidiary to subsidiary) versus hierarchical flows (e.g., parent-subsidiary or subsidiary-parent).

(b) Competitive flows (as between subsidiaries competing for resources or mandates) versus collaborative flows (as between units undertaking complementary functions in the same enterprise) versus autonomous (as between units undertaking unrelated activities under the same corporate umbrella).

(c) Transplantation (as in the parent to subsidiary flow involved in a greenfield start-up) versus supplantation (as in the infusion of fresh knowledge from the MNE to supplant an existing knowledge base in an acquisition) versus integration (as in the ongoing leveraging of a multisite knowledge base within in MNE).“(Mudambi 2002: p. 4).

His classification confirms that MNEs can willingly or unwillingly experience different types of knowledge flows. It is of importance that managers are aware of and strategically undertake efforts to influence the flows. Considering the finding of Gassmann and von Zedtwitz (1999) that MNEs move towards the R&D network for R&D organization, firms seem to be increasingly aware of knowledge flows and the potential local knowledge can have when integrating it. While managers seem to be aware of these flows between subsidiaries and headquarters, inside or across firms, it is unclear how innovation exploits existing knwoledge embedded in innovations.

In conclusion, networks are formed at the inside of the firm's boundaries as well as with external entities and can be differentiated according to the type of knowledge flow and knowledge type. Global knowledge networks have gained attention by scholars, in particular. Moreover, extensive literature adopting a geographical perspective has emerged that parts from the micro-perspective focusing on the MNE linkages. In line with the micro-perspective on the MNE, Dunning (2009) reveals the importance of spatial clusters for firms: "[...] the structure and content of the location portfolio of firms becomes more critical to their global competitive positions." (Dunning 2009: p. 16).

3.7.2 Role of geographical proximity

While some scholars argue that global knowledge networks are gaining importance, the literature on innovation theory and economic geography indicates, that knowledge is localized. Knowledge is localized since innovation processes are related to institutional settings such as property right protection. Furthermore, innovation is interactive (e.g., Bruche 2009a; Patel and Vega 1999; Samara, Georgiadis, and Bakouros 2012). Hence there is a spatial aspect inherent to innovation. As has been mentioned, MNEs leverage locational attributes when locating themselves in clusters (Gugler, Keller, and Tinguely 2015). Sourcing knowledge from different locations enables MNEs to maintain a competitive edge. However, sourcing knowledge from different locations is not void of complications. Already in 1999 transaction costs economics, going back to Williamson, had highlighted the advantage of geographical proximity that decreases transaction costs (Williamson and Masten 1999). Thus, the emergence of clusters is comprehensible. Not only are knowledge spillovers easier with geographical proximity due to the often tacit nature of knowledge but also due to the availability of common resources such as skilled labor (Criscuolo, Narula, and Verspagen 2005; Furman et al. 2005). In line, Gugler, Keller, and Tinguely (2015) assumed that the location within a cluster fosters a firm's ability to generate innovation.

Vanhaverbeke (2006) confirmed that clusters and regional innovation systems are essential for open innovation. Hence, the geographic context plays an important role. Both the findings of Jaffe, Trajtenberg, and Henderson (1993) and Cooke, Braczyk, and Heidenreich (2004) confirmed the relevance of the geographical context. The finding that knowledge flows more easily to closer entities explains firms' clustering around universities and knowledge centers. Indeed the role of public research entities is recurring in several studies

(Cockburn and Henderson 1998; Furman et al. 2005; Perri, Scalera, and Mudambi 2017). Public spillovers are said to provide emerging economies firms with fundamental knowledge critical for the development of innovation capabilities and are said to drive private sector productivity (Furman et al. 2005). On another note, the importance of geographical proximity is revealed when looking at the co-inventorships in the pharmaceutical industry (Perri, Scalera, and Mudambi 2017). Then again Grimes and Du (2013) described how some foreign MNEs behave in China regarding innovation and reveal that these foreign MNEs prefer to keep their core R&D close to home. Thus geographical proximity plays a role. Nonetheless, dispersed R&D can be beneficial; in the case of China, it is often market access that drives foreign MNEs to patent innovations in China itself (Grimes and Du 2013). Thus, while geographical proximity plays a role, some arguments convince foreign MNEs to innovate further away. Ultimately, this is how China is gaining access to global innovation networks.

3.7.3 Knowledge flows and innovation networks

While the literature agrees on the positive impact of clusters and innovation networks on innovation (e.g., Mudambi and Swift 2011), some suggest that there is a "risk of becoming closed to outside knowledge and becoming overembedded" (Simard and West 2006: p. 230; Uzzi 1997). Simard and West (2006) list several strategies to prevent the over-embeddedness, such as forming weak ties, the diversity of ties and institutions and exploiting structural holes. Weak ties, for instance, enable the firm to stay flexible and change its orientation, thereby counteracting symptoms of path-dependency. Thus innovation networks have to be of a dynamic character.

The literature on R&D internationalization and innovation networks is increasingly paying attention to emerging markets such as China. Several studies look at co-inventor networks and the role of emerging markets (Hu et al. 2015; Perri, Scalera, and Mudambi 2017). The large literature agrees on the importance of international connectivity for emerging countries. In particular, MNEs have been known for creating linkages to domestic actors, leveraging knowledge flows. In their empirical analysis Perri, Scalera, and Mudambi (2017), however, revealed that in addition to fostering linkages to MNEs, domestic actors in emerging countries should foster linkages to advanced economies research institutions. They found evidence that research institutions generate wide-ranging connectivity and more fundamental knowledge, whose knowledge flows foster the emerging country's technological upgrading. Their empirical analysis looked at the co-inventor networks of US patents associated with the Chinese pharmaceutical industry. In a similar vein, Hu et al. (2015) analyzed the international collaboration network in pharmaceutical patents granted by the US Patent and Trademark Office (USPTO) between 1996 and 2013. Initially the US was at the center of a mono-centric network, which is understandable since it represented 40% of the world market. Nowadays, the network has developed towards a star-like one. The authors explained this development by an increase of co-inventorships from geographically close countries. Finally, while China emerged as a country in the network, connected through Japan, it is not a significant actor in the network. China might not have surfaced as a key player, since the country might have limited interest of filing patents at the USPTO (Hu et al. 2015). For instance, MNEs internationalizing R&D to China target local diseases whose treatment would not be patented in the US since there is no occurrence of the local disease. Nonetheless, the study has shown, that China maintains knowledge flows and linkages with Japan that is geographically closer and a significant player in the global pharmaceutical industry.

In a recent working paper from 2018, Turkina and van Assche (2018) look at global connectedness and local innovation and reveal that "[...] [i]nnovation in knowledge-intensive clusters disproportionately benefits from enhancements in their constituent firms' horizontal connectedness to foreign knowledge hotspots" (p. 1). In general, the innovation performance of a cluster can be encouraged by horizontal or vertical connectedness. Thus, while geographical proximity plays a role for innovation – here in clusters – it can benefit from external connections as well.

In conclusion, firms that are embedded in innovation networks can benefit from the geographical proximity of entities. These networks characterized by knowledge flows can differ according to the type of knowledge, flow, and linkage. Firms can exploit these different types for their strategic purposes. While geographical proximity is beneficial, there are factors in favor of a global innovation network and to the detriment of the argument of geographical proximity. China, for instance, features several factors that convince foreign MNEs from the Western world to innovate far away from home. Thus knowledge flows across borders and China can gain access to Western knowledge, possibly influencing its innovation capabilities. It is undisputed that knowledge flows are a given part of today's innovation process. While having been analyzed at the headquarters - subsidiary level and macro-level, when looking at clusters and countries, the knowledge exploitation of innovation in terms of sourcing knowledge from different research fields, has not been worked on. We can only think of Mudambi (2002) who refered to a similar notion, when talking about collaborative knowledge flows that source knowledge from complementary fields across different subsidiaries. But independently from whether it is across subsidiaries or at the inside of the same subsidiary, might there not be innovations that rely on existing innovations or innovations created thanks to the knowledge exchanged between different research fields? This is an aspect we will dedicate part of our analysis to.

3.8 Innovation through MNEs: R&D internationalization

While originally, a firm's innovation management centralized R&D at home, over the last three decades, firms are increasingly internationalizing their R&D activities. In particular, since the year 2000 developing countries are gaining attention (Bruche 2009b). In fact, China aimed to exploit this development by introducing an economic growth strategy through technology and innovation that targeted foreign R&D investment. The government expected foreign R&D investment to promote the domestic creation of knowledge through knowledge transfers. However, since 2006, policies have been implemented, aiming to shift the concentration from foreign R&D investment towards the creation of domestic innovation ("UNESCO Science Report: Towards 2030" 2016; Gassmann, Beckenbauer, and Friesike 2012). Nonetheless, it is essential to consider the literature on R&D internationalization since it helps to understand China's emergence as an innovator and the creation of global innovation networks. Moreover, it indicates how different R&D units in a company relate to each other and how their cooperation looks like. It will provide indications on what kind of strategies firms adopt on how different R&D units cooperate when developing innovations. Since this thesis aims to look at innovation dynamics and proposes a new variable, the similarity between innovations, that measures synergies between different thematics of innovations and thus knowledge flows between research fields, this literature on R&D internationalization has to be consulted.

3.8.1 Definition of R&D internationalization

Many scholars have developed definitions of R&D internationalization or the globalization of research, as Ramirez (2006) calls it. The latter, for instance, explains that it "refers to the development of an international division of labor in research through the existence of coordinated and integrated international intra- and inter-firm research networks" (Ramirez 2006: p. 144). This definition differs from Patel and Vega (1999)'s, that refers to the percent of research undertaken by firms in a host-country as the globalization of research. Unlike Patel and Vega (1999) and Ramirez (2006), Janne and Cantwell (1999) distinguish between internationalization and globalization. They follow Patel and Vega (1999)'s definition when using the term internationalization, and Ramirez (2006)'s definition for the term globalization. While slightly different, both concepts base on the postulate that research is located in a foreign country.

3.8.2 Enabler of R&D internationalization

R&D internationalization as a concept is closely intertwined with the management literature on MNEs since R&D was restricted to MNEs in the 1980s and 1990s (Gassmann and von Zedtwitz 1999). This former restriction explains that MNEs are the primary drivers of R&D internationalization according to the OECD (2008). While in the past MNEs kept their R&D units at home, the increasing competition in the global environment rendered the internationalization of R&D lucrative. MNEs saw pressure to access new knowledge and accelerate the time from development to market (Kuemmerle 1997). Both Dunning (1998) and Porter (1990) helped promulgate the understanding that the location choice of a firm is important for its competitive advantage.

Boutellier, Gassmann, and Zedtwitz (2008) are among those that investigated the enablers of R&D internationalization. Boutellier, Gassmann, and Zedtwitz (2008), relying on different academic papers, categorized the different enablers for R&D internationalization into input-oriented enablers, output-oriented enablers, external enablers, efficiency-oriented and political/ social-cultural enablers (p. 50) (see Table 3 for details). Moreover, they emphasized that R&D internationalization is frequently "a by-product of decisions or developments outside the traditional scope of R&D" (Boutellier, Gassmann, and Zedtwitz 2008: p. 50). They refered to mergers and acquisitions that include for instance the acquisition of foreign R&D units.

Table 3: Enablers of R and D internationalization

Group 1		Group 2	Group 3	
Input-oriented	**Output-oriented**	**External**	**Efficiency-oriented**	**Political/ socio-cultural**
- Information and communication networks	- Improving local image	- History	- Improving flexibility through new organization	- National and legal conditions
- Center-of-innovation	- Customer-specific development	- Peer pressure	- Critical mass	- Patenting laws
- Infrastructure	- Closeness to lead users	- Tax optimization	- Reduced project failure risk through parallel development	- Protectionist barriers
- Qualified human resources	- Local values	- Acquisition of parent company	- Making use of many time zones	- Predictable labor relations
- Adaptation to local production processes	- Market and customer proximity	- Merger	- Proximity to production, marketing, distribution	- Local content
- Scientific community			- Learning curve	- Legal restrictions
- Tapping informal networks			- Reduction of development cycle time	- Acceptance
- Country-specific cost advantages			- Overcoming logistic barriers	- Subsidies
			- Lower R and D personnel costs	- Taxes

Source: Boutellier, Gassmann, and Zedtwitz (2008): p. 50

Bruche (2009b), looking at China and India's emergence in MNE's innovation networks, explained that in the case of China, firms were initially moving their R&D to China to exploit their assets. Existing products and manufacturing processes were sought to be complemented and supported, hence lower-value added R&D was off-shored. Moreover, a new market was seized. Recently R&D activities of foreign MNEs in China are moving up the value chain. Bruche (2009b) listed Bayer Health Care in China, for instance, and explained that such initiatives "are not only trying to tap important emerging knowledge bases in specialized areas but also demonstrating that they are 'good corporate citizens' to the respective national health authorities - a move that may support their positions in the local market" (p. 278). Although Bruche (2009b) showed that activities are moving up the value chain, he stated that the MNE's shift of R&D to China is at an early stage. The paper concluded that "a 'new geography' of the MNC innovation value chain seems to be emerging, [but; remark by author] it is still largely characterized by a hierarchical organisation of the R&D value chain" (Bruche 2009b: p. 280).

While MNEs experienced pressures to internationalize their R&D, there are challenges to the geographic dispersion, also. For instance, the distance between R&D units and the tacit character of knowledge challenges the management of R&D in MNEs (OECD 2008; von Zedtwitz and Gassmann 2002). Firms with internationalized R&D activities face a more complex R&D organization and additional costs (Gassmann and von Zedtwitz 1999). Thus, it might be more efficient for companies to centralize their R&D activities than dispersing them worldwide (Boutellier, Gassmann, and Zedtwitz 2008).

3.8.3 Organizational Economics and R&D internationalization

Boutellier, Gassmann, and Zedtwitz (2008) see the decision in favor of international R&D as the rejection of central R&D. They explained that R&D internationalization is offset by factors supporting centralization and factors challenging international R&D (Boutellier, Gassmann, and Zedtwitz 2008). Nonetheless, in some cases, enablers of R&D internationalization overlay these factors. From history, Boutellier, Gassmann, and Zedtwitz (2008) concluded that it is mostly caused by "mergers and acquisitions, localization needs, and the need for skills" (p. 55). While Boutellier, Gassmann, and Zedtwitz (2008) assessed the possible reasons for a decision against or for the internationalization of R&D, Gassmann and von Zedtwitz (1999) concentrated on identifying different types of R&D organization in MNEs. They determined five types according to "the dispersion of R&D activities and the degree of cooperation between individual R&D units" (p. 231) by means of "195 semi-structured research interviews in 33 technology-based firms between 1994 and 1998" (p. 235). These are ethnocentric centralized R&D, geocentric centralized R&D, polycentric decentralized R&D, the R&D hub model and the integrated R&D network. The first two models represent a centralized organization although with an increasing orientation towards international markets. While no physical internationalization has taken place, the geocentric centralized R&D model requires an international awareness of the employees. Workshops abroad can teach employees this international orientation and awareness. In contrast, the latter three models are characterized by R&D units in host countries. In the case of the polycentric decentralized R&D, these units are mainly established to ensure local responsiveness, hence product adaptations. This strategy is based on the knowledge-based view which argues that multiple locations of R&D enable a firm to access new knowledge and to adapt its existing knowledge to new markets, overall improving its innovation capabilities (Grant 1996). This model, however, is challenged due to its lack of coordination between units. This is the reason why Gassmann and von Zedtwitz (1999) anticipate the extinction of this model. The R&D hub model constitutes a coordinated network of R&D units with the home-based R&D unit at the center that is the main location of R&D and retains the worldwide lead in relevant technological fields (Gassmann and von Zedtwitz 1999). Foreign R&D units concentrate their activities on specifically assigned areas. The authors explained that this model is often adopted by centralized firms that experience the internationalization of resources. Unlike the R&D hub model, the integrated R&D network does not constitute of a domestic central R&D unit controlling all activities. Instead, "central R&D evolves into a competency center among many interdependent R&D units which are closely interconnected by means of flexible and diverse coordination mechanisms" (Gassmann and von Zedtwitz 1999: p. 243). In this vein, Mudambi (2002)'s notion of collaborative knowledge flows is evoked. According to the integrated R&D network, each local R&D unit establishes local competencies, which are leveraged for the benefit of the whole network. Not only, did Gassmann and von Zedtwitz (1999) state that pharmaceutical companies are adopting this R&D organizational structure by assigning specific fields of research to units, but they also revealed that there is a general trend towards the R&D integrated network. In support of global efficiency, firms reduce their R&D units to fewer, coordinated leading centers. These few leading centers are located in geographic areas with excellent technological knowledge or lead markets (Gassmann and von Zedtwitz 1999). Overall, firms seem to increasingly adopt an international orientation with specialized R&D units as well as coordinate and integrate these units better. While it is argued that the cooperation between different specialized research fields is beneficial for the whole network, empirical studies have not looked at it yet. Thus, this thesis presents a proof of concept revealing the explanatory power of a variable of similarity between patents and the impact it has on innovation output. The similarity across patent thematics thereby might reflect the trend towards leading centers of excellence that cooperate across borders.

The trend towards leading centers of excellence focalizing on specific areas is a move away from the classification of Kuemmerle (1997) that R&D sites are either home-base augmenting or home-base exploiting. We consider it a move away, since the home R&D center loses its significance. The terms home-base augmenting and home-base exploiting refer to the coordination and the knowledge transfer between the home R&D unit and the foreign units. In the case of a home-base exploiting approach, the knowledge flows from the home-base R&D center to the foreign R&D sites. In contrast, in the case of the home-base augmenting approach, the foreign R&D sites are a means of accessing knowledge from foreign competitors and research institutions, which is ultimately lead back to the home-base R&D center. While both the works from Gassmann and von Zedtwitz (1999) as well as Kuemmerle (1997) focused on the knowledge flow between R&D units, the classification of Kuemmerle (1997) might be outdated because of firms adopting more flexible and interdependent organizational structures. Nonetheless, while some authors refered to the types of R&D organization from Gassmann and von Zedtwitz (1999; e.g., Boutellier, Gassmann, and Zedtwitz 2008), others, such as Awate, Larsen, and Mudambi (2015) and Criscuolo, Narula, and Verspagen (2005) adopted Kuemmerle (1997)'s home-base exploiting or augmenting approach. Criscuolo, Narula, and Verspagen (2005) highlighted the need to reconcile the home-base exploiting and home-base augmenting distinction. They postulated that both the innovation system of the home base region as well as host base region and the firm's technological resources influence the knowledge sourcing of affiliates. In a study on the patent citation patterns of EU MNEs in the US and US MNEs in Europe, they revealed that the MNEs heavily rely on home region knowledge sources while simultaneously sourcing from the local knowledge base, also. This holds, in particular, true for pharmaceutical MNEs but might be questionable in the sector of biotechnology (Criscuolo, Narula, and Verspagen 2005).

Over the past decades, most research on R&D internationalization has focused on advanced economies MNEs (e.g., Bruche 2009b; Grimes and Miozzo 2015; von Zedtwitz and Gassmann 2002), however more recent literature has introduced the perspective of emerging markets R&D (e.g., Altmann and Engberg 2016; Awate, Larsen, and Mudambi 2015). While the literature in innovation management saw the birth of frugal innovation and innovation for the bottom of the pyramid, the rationale gained some attention from the literature on R&D internationalization, also (Altmann and Engberg 2016). Traditionally arguing in favor of the development of frugal innovation in the host-country, a case study by Altmann and Engberg (2016) challenges this premise. Depending on the transferability of the technical knowledge, a home-base exploiting R&D approach might be more promising. Nonetheless, there are other factors, such as market knowledge, that have to be taken into account as well, before deciding on a home-base exploiting R&D approach.

So far there is evidence that R&D is undertaken by both foreign MNEs and domestic firms in emerging economies creating global innovation networks. Different strategies with regard to organizational structure exist that imply different degrees of local knowledge creation in subsidies. Recently the literature supports the argument that emerging economies such as China create specialized knowledge that foreign MNE headquarters tap into by adopting an R&D network approach. The following chapter will synthesize the literature that examined the impact of R&D internationalization on innovation output, since China initially counted on foreign R&D to increase its innovation capabilities and innovation output.

3.8.4 R&D internationalization and innovation output

Underlying the literature on R&D internationalization and in particular its aspects of R&D organization and knowledge flows rests the question whether internationalized R&D entails a better innovation output (e.g., Leiponen and Helfat 2011). While a competitive environment fosters the R&D internationalization process and scholars identified the objectives of different types of R&D organization, these do not have to lead to better innovation outputs. Indeed, Leiponen and Helfat (2011) find that multiple R&D locations are positively associated with imitative innovation, not new-to-market innovation. They interpret that multiple locations are sought by firms in pursuit of imitative innovation, since the variable multi-location correlated with greater external knowledge sourcing, in their sample of the Finish manufacturing sector (Leiponen and Helfat 2011).

The findings of Penner-Hahn and Shaver (2005) indicate that internationalization of R&D increases innovation output in case of Japanese pharmaceutical firms under the premise that the firms already have research capabilities in the respective technology. Instead of focusing on the extent of innovation output, Singh (2008) investigates the impact of a firm's R&D geographic dispersion on innovation quality. He reveals that R&D internationalization has an adverse effect on the value of innovation. This finding indicates that the challenge of integrating knowledge across regions is not overcome by firms yet.

The organizational structure of internationalized R&D is expected to have an impact on the innovation output as has been indicated in the previous chapter. Decentralization or centralization favor different kinds of innovation. While a centralized approach entails more generic innovation that has a subsequent impact on technological evolution, a decentralized approach concentrates on "[...] solving narrower, business-unit specific challenges[...]" (Argyres and Silverman 2004: p. 935). Hence, given the trend towards decentralized R&D networks, it can be expected that firms generate specific innovation in the context of the R&D unit and the national/regional environment. Since Argyres and Silverman (2004) argue that this entails innovations of a more narrow character, focalized on specific topics, decentralized R&D might rather foster incremental innovations than radical innovations.

Many researchers have looked at foreign versus domestic innovation, arguing that foreign firms are faced with a liability of foreignness (e.g., Barnard 2010; Klossek, Linke, and Nippa 2012;2010; Zaheer 1995). However, by drawing on the concept of foreignness, Un (2011) argued in favor of the advantage of foreignness. Analyzing foreign and domestic subsidiaries in Spain, he revealed that foreign subsidiaries are more innovative than domestic firms. He assumed that it is due to the integration in a global innovation network. Furthermore, he hypothesized that in the context of developing countries and advanced market MNEs (AMNEs), the findings might be even more relevant. Domestic firms might not be able to match the innovativeness of products transferred from developed countries.

In recent years, emerging market MNEs (EMNEs) have emerged and become the center of studies. In 2015, Awate, Larsen, and Mudambi (2015) published a case study that compared R&D internationalization of an advanced market MNE (AMNE) to an emerging market MNE (EMNE). In view of global innovation networks, they assumed that knowledge levels between R&D units would differ and hence would incite catch-up processes between R&D units. This study revealed that EMNEs and AMNEs differ in their catch-up processes. In the case of the AMNE, Vestas, the subsidiaries specialize and support headquarters. The patent citations, however, reveal that despite its intentions, Vestas did not leverage the knowledge created in the subsidiaries. In contrast, Suzlon, the EMNE, exploits the local knowledge of their subsidiary locations. Headquarters sources their knowledge and develops the knowledge further for catch-up purposes (Awate, Larsen, and Mudambi 2015). On the intra-firm level, researchers have shown that the internationalization of Chinese firms can be beneficial for their innovation capabilities (Jiang, Branzei, and Xia 2016; Ying, Liu, and Cheng 2016).

Several authors have dedicated studies to research on knowledge flows, innovation networks, internationalization strategies and R&D internationalization in China and their meaning for China's innovation capabilities (e.g., Lin, Zhang, and Wang 2013; Hu et al. 2015; Jiang, Branzei, and Xia 2016; Motohashi 2015; Shi et al. 2014). While there is consent on the positive effect of China being integrated into global innovation networks form a theoretical perspective (Bruche 2009b; Bruche 2009a), Chinese researchers reveal in empirical studies, that external knowledge is not beneficial for indigenous innovation (Fu and Gong 2011; Jiang, Branzei, and Xia 2016; Sun and Liu 2010). Fu and Gong (2011), for instance, showed that joint indigenous R&D activities at industry level are a driver for the catching-up of domestic firms, while technology transfers from foreign firms might only help in the initial stages of the process. Lin, Zhang, and Wang (2013), however, found positive spillover effects from foreign firms to domestic firms in China. Thus the literature and empirical studies show no consensus on the effects of knowledge transfers on indigenous innovation in China.

In conclusion, the effect of R&D internationalization on innovation output has been tackled from different points of view. There is, however, no unanimous accord on its effect. This applies to the literature on R&D internationalization in the context of China's emergence in global innovation networks as well.

4 Literature Gap

The preceding literature review has started with an explanation of China's interest in innovation, and essential understandings about the process and models of innovation. This context led us to have a closer look at newly employed innovation models and concepts. These innovation concepts postulated an increasingly connected innovation landscape spanning across borders. In this vein it was of interest to evaluate the existing literature on R&D internationalization and knowledge flows.

It has been revealed that developing countries are slowly becoming of interest to researchers (e.g., Bruche 2009b; Huang 2010). In particular, on China there has been a slowly increasing body of literature (e.g., Fan 2014; Sun 2000). Nonetheless, research on China and its innovation capabilities remains superficial, descriptive and inconsistent. While the economics literature explained China's efforts to foster indigenous innovation, the literature on R&D internationalization explained China's emergence in MNE innovation networks (e.g., Bruche 2009b). Indeed, the Chinese government fostered this development by initially attracting foreign R&D. It is assumed that this was supposed to initiate knowledge transfers to local firms. Since 2006, however, the government tries to foster indigenous innovation. In line with the literature on R&D internationalization, the literature on open innovation models and knowledge pipelines postulates the flow of knowledge across firm departments and across firm-boundaries (e.g., Mudambi 2008; Mudambi, Piscitello, and Rabbiosi 2014; Perri, Scalera, and Mudambi 2017; Simard and West 2006). These flows are mostly considered beenficial for the innovation output. The innovation process is nowadays regarded as a network or system with several actors working together (e.g., Pilat et al. 2009). This is equally supported by the literature on National Innovation Systems that confirms from a theoretical perspective the important role the Chinese government plays for the country's emergence as innovator (e.g., Fagerberg and Srholec 2008). Overall, developing countries and China gain increasing attention from scholars. China's innovation efforts are however mostly analyzed from a macro-perspective or a firm-strategic perspective in a descriptive way (e.g., Hu and Mathews 2008; Lundvall 2009). On the one hand, its national innovation system has gained attention from scholars or its emergence in global innovation networks, and on the other hand frugal innovation concepts.

Overall, the notion of collaboration or network has emerged throughout the literature review. Be it from a firm-strategic perspective or a country perspective, knowledge flows are present. Nonetheless, the notion of cooperation has not received attention yet, when looking at a specific industry and its innovation dynamics. An industry that usually focuses on different research fields might try to leverage knowledge flows across these research fields. The literature on R&D internationalization has slightly touched upon this topic with the integrated R&D network (Gassmann and von Zedtwitz 1999), as well as Mudambi (2002)'s notion of the collaborative knowledge flow across subsidiaries. Both concepts focus however on the knowledge flow between subsidiaries. Finally, we propose the usefulness of a new variable "similarity", measuring this aspect of knowledge flows independently from the subsidiary perspective. This variable "similarity" will allow to show how innovations rely on existing innovations from other research fields. Finally, we assume that it can explain the innovation output in terms of patent count since knowledge can flow between research fields, creating synergies that are measured by the similarity and ultimately impact the patenting behavior.

Not only have scholars not looked thoroughly at this aspect, but also the description of an industry's innovation and its key areas is seldom considered a topic for research. This is a promising avenue for research since we have to rely on expert statements for this so far (e.g., Ni et al. 2017; Zhang and Zhou 2017; Shen 2010). A more detailed description of innovation dynamics in an industry in a developing country such as China would be beneficial for future research on innovation in developing countries. So far, innovation dynamics in the industry have not received a lot of attention yet, which might be due to data availability or applicable methodologies.

Given the development of new methodologies such as Data Science techniques, it seems promising to have a look at China's innovation dynamics. To our knowledge, no researcher has leveraged Data Science techniques to analyze innovation dynamics in the form of a wide-ranging patent analysis. In fact, the literature review has shown, that the pharmaceutical industry has not been analyzed from this perspective a lot. Nonetheless, it is very interesting due to its global character, and propensity to innovate and patent. Moreover, it resides in an increasingly competitive environment, forcing firms to increase the pace of innovation and might, therefore, foster cooperation not only across subsidiaries but generaly across research fields. All in all, the pharmaceutical industry is promising for the analysis of innovation dynamics and allows to leverage a patent analysis based on Data Science. For the first time, we will be able to gain insights into this topic from a vast, unique patent base.

5 Data-Science-based methodology

5.1 Research proposition

The contextualization and the theoretical framework of this thesis have provided ample support to the relevance of this research. From a theoretical as well as practical perspective, it is of interest to identify innovation dynamics in the pharmaceutical industry in China. Innovation dynamics have incited research since decades, scholars have however failed to describe them for a specific industry from a quantitative perspective. The existing papers focus on data collected manually by researchers and are of a descriptive nature. To our knowledge, the closest researchers have come to analyzing innovation dynamics in the pharmaceutical industry, is a paper of Achilladelis and Antonakis (2001). Nonetheless, their research is of a descriptive nature, also. Moreover, they solely focus on product innovations and their analysis ends in 1990. Finally, they focus on identifying generations of drugs and on the distinction of clusters of groups of drugs. This thesis differs in so far, as it integrates product and process innovations and tries to identify synergies between different thematic classifications of innovation, meaning knowledge flows across research fields.

Overall, this thesis's contribution will be two-fold. It will develop a protocol, explaining how Data Science techniques can be leveraged to analyze an enormous patent database and ultimately to picture innovation dynamics in the pharmaceutical industry in China. Moreover, it proofs the relevance and usefulness of the new variable "similarity", meaning synergies between research fields. This new variable has not received attention from scholars yet, but might yield interesting results with regard to the relationship of thematics across time and innovation output. Data Science techniques are a new set of skills leveraged by business professionals and researchers equally that allows us to do so. Researchers from the international business field have proven to be hesitant to exploit the new methodologies' benefits. Thus, this thesis will leverage new tools enabling us to analyze innovation dynamics to an extent and with a level of details that was not possible so far. Thereby, the data collection process and the data treatment will be explained in detail so that the analysis is made transparent and reproducible.

Overall, this thesis' motivation is the elaboration of a describtion of innovation dynamics in the pharmaceutical industry in China, leveraging Data Science techniques. Moreover, the thesis aims to prove the usefulness of a measurement of similarity that reflects synergies between research fields. Three research questions have been developed that capture innovation dynamics. It would be of utter difficulty to find answers to these aspects of innovation dynamics without computational methods. The three questions are the following:

(1) *Where lies the focus of innovation in the Chinese pharmaceutical industry? For the whole time span of 1990 to 2017? For 1990 to 2000? For 2001 to 2017?*

While experts have made statements with regard to this question (Zhang and Zhou 2017) and Achilladelis and Antonakis (2001) have developed manually a list of all pharmaceutical drugs invented, these analyses are not complete and mainly not empirical. The present research not only aims to look at new original drugs but also at new components that later on might come into play when developing new drugs, all from a quantitative perspective. From the literature it is known that experts believe China's recent innovation to focus on cancer-related drugs, liver disease, cardiovascular disease, diabetes and biotech (Ni et al. 2017; Zhang and Zhou 2017; Shen 2010). With regards to the time span of 1990 to 2000, however, scholars expect China to focus on basic, supporting research, for the likes of MNEs internationalizing their basic research activities to China. At times, the intellectual property schemes of China were premature, and MNEs refrained from patenting their innovation in China (Gassmann, Beckenbauer, and Friesike 2012). Moreover, the Chinese government did not yet foster indigenous innovation capabilities. Thus, in accord with the literature, we propose that:

> (H1.1) Overall, innovation in the Chinese pharmaceutical industry focuses on topics such as cancer, biotech, diabetes, diseases such as liver and cardiovascular diseases; this applies in particular to innovation after 2010.
>
> (H1.2) China's innovation in earlier times (1990-2000) differs from the thematics patents cover after 2000.

The second questions refers to the new variable "similarity" that we put forward when analyzing innovation dynamics. This question thereby provides first insights into the relationship between different patent thematics across time.

(2) *What kind of dynamics do we find between patent thematics/classes? More precisely: What is the relationship of patent classes across time?*

It is common knowledge that the development of innovations relies on prior innovations. This is in particular the case for incremental innovations that improve existing innovations, but can also be the case for radical innovations (Freeman 1982). To our knowledge, so far no researcher has attempted to capture the dynamics of innovations across research fields empirically, represented by patent classes and across time. Thus, we extend the existing literature on knowledge flows, when looking at similarities between different patent thematics of different years. Overall, this thesis will attempt to characterize the relationship of patent classes across time. We assume thereby that certain classes will be very similar to other classes since they might rely on the same basic research. Then again, time is assumed to play a role when describing the dynamics across classes.

> (H2) Measuring the similarity across classes and time will allow us to characterize the dynamics between classes and to assess whether time might play a role when capturing similarities across patent classes.

This second question and hypothesis are very important since they allow to illustrate how innovations rely on previous innovations and how their similarities to previous innovations evolve over time. Moreover, it lets us identify whether all classes rely on the same extent on previous innovations.

After having provided an overview of the dynamics across patent thematics, we will prove the usefulness of the new explanatory variable from an econometric perspective, when asking:

(3) *What is the relationship between the similarity of patent classes and the patent count when taking into account the time lag?*

This latter question allows to not only assess whether there is a relationship but also allows to capture what we call the innovation cycle of the pharmaceutical industry in China. We capture this cycle, when assessing the time after which similarities across classes in a given year exercise an impact on subsequent patenting. Thereby, we assume that the impact will be positive. Similarities across patent classes at a given time might reflect synergies in research and knowledge flows that are created across research topics and fields, which ultimately fosters closer cooperation between researchers working on the different topics and thus might lead to increased patenting. The trend towards R&D networks (Gassmann and von Zedtwitz 1999) and open innovation models (Chesbrough 2006) are some of the reasons why it is of interest to look at this relationship. While the analysis will try to answer the subsequent hypothesis, it is also of an exploratory character since it will assess the time lag after which similarities across classes in a given year exercise an impact on subsequent patenting. The hypothesis is as follows:

> (H3) We propose that similarities across classes in a given year have a positive impact on the patent count at one point.

In the following, we will briefly touch upon the Data Science techniques employed to elaborate on these three questions. These methods, as well as the data collection process, build the protocol that we develop. Future research could exploit this protocol for the analysis of other industries, countries or further development of the protocol.

The prerequisite step of the protocol consists of applying text mining to a patent database of 238,870 patents related to pharmaceuticals between 1990 and 2017. Text mining allows to "read" the 238,870 patents in a reasonable amount of time, (ca. 8 hours). In consequence, the derived data allows us to describe patent assignees, patent classes, etc. when wrangling the data. It also enables us to compare China to other patent-issuing authorities. Finally, based on a framing strategy, a content analysis was employed in order to describe the therapeutic groups the patent database in China focuses on.

After having extracted the information from the patents and treating the data so that we can analyze it, the protocol foresees the unsupervised (semi-supervised) classification of patents through an Latent Dirichlet Allocation (LDA) analysis. The LDA analysis answers to the question (1) *Where lies the focus of innovation in the Chinese pharmaceutical industry?*. It allows a computational topic modeling of the patents, eliminating human-made inaccuracies. Ultimately, it enables us to compare its results to the previous application of classification using a framing strategy and gives the first insight into innovation dynamics in China through a comparison of the most common words they have in common.

In a second step, innovation dynamics will be mapped out thanks to the Jaccard similarity index. Going back to Jaccard (1902), the measurement allows to measure the similarity between texts by taking into account the words and the order they are in. Due to the scientific language of patents, this approach deems promising. Ultimately similarities between patent classes will be analyzed through time, which captures innovation dynamics. Thus, the measurement of the Jaccard similarity allows to answer the second question: (2) *What kind of dynamics do we find between patent thematics/classes? More precisely: What is the relationship of patent classes across time?*.

An econometric model will be based on the Jaccard similarity, analyzing the impact of the Jaccard similarity on the patent count, utilizing lagged Jaccard similarities. The econometric model constitutes the third step and will cater to the last question (3) *What is the relationship between the similarity of patent classes and the patent count when taking into account the time lag?*. Ultimately, the econometric model allows to assess the time it takes for similarities between patent thematics to influence the patent count. Or at which period of maturity the similarities impact the patent count, meaning whether similarities of the distant past or near past impact it. In this regard, the econometric model allows to capture the cycle of innovation of the pharmaceutical industry in China.

To summarize, this thesis uses four analytic methods: a content-analysis (with and without a "framing strategy"), second, an unsupervised (semi-supervised) classification based on a Latent Dirichlet Allocation (LDA) approach, third the Jaccard similarity to measure similarity between datasets, and finally econometric tests on a panel linear model analyzing the impact of the lagged similarity on the patent count. These approaches and each's purpose will be explained at a further point. Each is thereby considered part of the developed protocol. All analysis is performed via a computing platform called Nüance-R (www.nuance-r.com) [4].

Data Science techniques make the analysis of the presented aspects of innovation dynamics possible in the first place. Manually these questions would hardly be able to be treated. In future research, our findings can be cross-referenced with findings from other industries and countries when applying our protocol. Therefore, this thesis constitutes a novelty with regards to the topic as well as the methodological choice. It has been shown that there has not been enough quantitative research on innovation in China, in particular on the pharmaceutical industry. Moreover, Data Science techniques that are increasingly applied by business professionals have not been widely adopted by business scholars yet.

5.2 Sample description

5.2.1 Data collection process

In the following, the data collection will be explained, and the corpus will be presented. This part features the first step of the protocol that is being developed to capture innovation dynamics. In itself, the data collection is an integral part of the protocol since an algorithm had been developed to extract the information from .html files. Data Science allowed to extract relevant information from 238,870 patents related to pharmaceuticals from 1990 to 2017. A more detailed explanation will follow.

Before explaining how patents were collected, however, it is important to understand that patents are considered a valid proxy for innovation and have received attention from various scholars in different fields. Despite not ignominiously being considered a valid proxy for innovation, patents have proven to be useful in empirical studies. Several authors consider them a valid proxy (Hagedoorn and Cloodt 2003). In particular, within this research project, patents can be considered an adequate proxy, since the pharmaceutical industry is an industry highly reliant on the patenting of its innovations (Arundel and Kabla 1998; Frost and Zhou 2000; Hadengue, de Marcellis-Warin, and Warin 2015). Indeed, the pharmaceutical industry is one of the industries with the highest propensity rates for patenting, for instance, 79.2% in 1998 (Arundel and Kabla 1998). Without the patenting of new drugs and compounds, hence the protection against generics, pharmaceutical manufacturers cannot shoulder the high R&D costs associated with the development of pharmaceutical innovation. Moreover, Dang and Motohashi (2015) explained that patent statistics are a meaningful proxy for innovation in China since they found that patent count is correlated with R&D input and financial output. Nevertheless, patents are not able to fully picture innovation since innovations might at

[4] Nüance-R is a platform developed by Professor Thierry Warin.

times not be patented or are not possible to be patented (Griliches 1990). The risk of the latter is minimized, since this paper focused on the pharmaceutical industry.

Patents have been collected through the Derwent World Patents Index (DWPI), that covers 14.3 million inventions from 40 worldwide patent-issuing authorities ("Derwent World Patents Index," n.d.). Web of Science runs this database that has the advantage of providing global patent data in English. The translation has been undertaken by industry experts thereby ensuring the accurate translation of the content. Given the importance of the text content for this analysis, an accurate translation is crucial.

A keyword search enabled the researcher to identify patents related to the pharmaceutical industry. The keyword used was pharmaceut* and allowed to identify 238,870 patents from 1990 to 2017 related to pharmaceuticals (retrieved October 18th, 2017). The patents were downloaded in .html and .txt formats. They thereby encompass all the innovations of the industry, product and process innovations. The information retained in the Derwent World Patent Index does not allow to identify whether the patent constitutes a product, process, radical or incremental innovation. But, the consolidation of all types of innovations is essential when analyzing the dynamics of the industry as a whole.

5.2.2 Sample analysis

Patents have received attention from scholars in various fields and have seen different methodologies employed. Madani and Weber (2016), present a comprehensive overview of patent analysis methodologies, analyzing the evolution of patent mining. They identify three stages and explain that when taking the mean publication year of papers related to patent analysis, researchers focused on bibliometrics analysis and citation analysis around 2001 and 2004, in the second stage researchers concentrated on cluster analysis around 2004 and 2006, and then again in 2007 on using network analysis to analyze patents. Finally, they show - in the third stage - that text mining and semantic analysis is the most popular methodology in 2008 (Madani and Weber 2016). This thesis is adopting text mining as a methodology to extract information from patents.

Abbas, Zhang, and Khan (2014) defined text mining as a "knowledge based process that uses analytic tools to derive meaningful information from the natural language text" (p. 5). Text mining as a methodology thereby differs from the other methodologies insofar as it is based on patent content and allows to treat unstructured data as well as structured data, extracting knowledge and information (Madani and Weber 2016). Unstructured data of a patent are texts, such as claims, abstracts or descriptions of the invention whose information is lost when analyzing solely structured data such as patent number, filing date and assignees (Tseng, Lin, and Lin 2007). Hence, applying text mining techniques allows getting a better picture of a patent. It has to be noted that text mining techniques cannot guarantee the correct representation of concepts (Abbas, Zhang, and Khan 2014). Nevertheless, it is a promising approach since it allows to take into account structured as well as unstructured data and hence the patents in their totality.

According to the literature, a text mining approach in patent analysis can be primarily categorized in the following approaches: "NLP, property-function based approaches, rule based approaches, neural networks based approaches, and semantic based approaches" (Abbas, Zhang, and Khan 2014: p. 6). Text mining used in patent analysis is "largely based on NLP, property-function based approaches, rule based approaches, neural networks based approaches, and semantic based approaches" (Feinerer, Hornik, and Meyer 2008).

This paper uses an NLP based approach, which means that “computational mechanisms to analyze and represent the textual information present in electronic documents” are used (Abbas, Zhang, and Khan 2014: p. 6). Although this approach can entail certain issues of adequate representation of the information contained, they are considered very effective in analyzing “large documents containing huge volumes of textual data” (Abbas, Zhang, and Khan 2014: p. 6).

Thus, as a first step, the database was created from multiple (480) .html files, which contain the full patent with title, abstract, publication date, assignees, inventors, etc. These .html files were extracted from the Derwent World Patent Index, according to the aforementioned method. As mentioned, the thesis relies on the power of the platform Nüance-R. RStudio’s IDE is used for accessing the platform. The tm package, allowing to text mine data was installed on the platform. Then Feinerer, Hornik, and Meyer (2008) as well as Silge and Robinson (2017) was followed, who offer a comprehensive overview of text mining techniques with R, using the tm package in R. After the collection of the .html files, the dataset was prepared and tidied following Wickham (2014)’s principle: “1. Each variable forms a column. 2. Each observation forms a row. 3. Each type of observational unit forms a table.” (p. 4). Furthermore, the dataset was cleaned [5] whilst extracting information related to the patent index, patent number, the title, the abstract, the publication date, the geographic area of registration, the inventor, the assignee, the Derwent class, the manual code and the type of publication. Moreover, we created a variable for the sequence of inventors and the sequence of assignees. So that when looking for the top patent holders, for instance, we can classify according to whether it was the assignee in the first position.

Each index regroups at least one patent number. From the patent number, the Derwent Primar Accession Number, the publication date is extracted. Moreover, the geographical area of registration is obtained by subsetting the two first capital letters of the patent number (for example, for China this corresponds to *CN*). Since one patent can be assigned to several patent holders, Derwent classes and manual codes, a separate dataset was created with split assignees, Derwent classes, and manual codes, allowing a more distinct analysis. Overall, it took around one day for the written code to compile and extract the information from the patents into a data frame.

In a subsequent step, the dataset was further cleaned, taking into account patent applications and not granted patents [6]. Patent applications include changes to the patent content as well as petty patents. Finally, the original database of 238,870 patents was reduced to 228,108 patents, comprising 891,159 patent numbers, of which 69,923 were filed at the SIPO in China. These remaining 228,108 patents were filed at 39 different patent-filing authorities (Appendix: Table 6).

While a part of the analysis covers the whole database of 228,108 patents, we are interested in identifying the therapeutic groups covered by the patents. To this end, a content analysis based on a framing strategy was employed, following the Pharma R&D Annual review 2017 (Lloyd 2017). The database was reduced to 82,672 patents by selecting the patents that are presenting 15 different therapeutic groups: *Alimentary*

[5] The cleaning involved the elimination of certain terms such as ”Detailed description” or ”novelty” that are titles in the patent abstracts.

[6] Since national bureaus tend to have different document types and assigned codes, the codes related to patent applications were identified for each country separately. For the US it was not possible to separate patent applications and grants. Nonetheless, the authors have decided to include the US in the dataset, since the focus of this thesis lies on China, and thus this limitation can be neglected.

Metabolic, *Anti-Cancer*, *Anti-Infective*, *Anti-Parasitic*, *Biotechnology*, *Blood clotting*, *Cardiovascular*, *Dermatological*, *Genitourinary*, *Hormonal*, *Immunological*, *Musculoskeletal*, *Neurological*, *Respiratory* and *Sensory*. The decreased size of the database can be explained by patents not figuring any of these topics. It has to be noted that a patent can belong to more than one therapeutic group since a keyword approach was used; leveraging text mining, each patent abstract was scanned for certain terms. Table 5 in the appendix provides an overview of the 15 therapeutic groups and the keywords used to classify patents in a therapeutic group.

To recapitulate, all computation involving the tokenization of the abstracts is based on Silge and Robinson (2017), using the tidytext library (De Queiroz et al. 2018) and the tidyverse library (Wickham 2017) in R, installed on the Nüance-R platform.

In the following, the data, mined from the patent database, will be described. The data description derived through Data Science techniques in itself thereby constitutes a contribution to the research field due to the quantitative perspective, and since the manual consultation of this large, unique patent dataset would not have attained such a detailed description, not even considering the time it would take.

5.2.2.1 Patents count 1990 to 2017

While 228,108 patents were retrieved in the time span 1990 to 2017, the bulk of the patents was filed between 2004 and 2017 accounting for 179,540 patents. The coverage of Chinese patent publications in the Derwent Innovation Index commenced in 1985. Hence patents filed at the SIPO are included over the whole time span chosen. The following figure (Figure 11) illustrates the distribution of all patent filings over the time span 1990 to 2017 after the data cleaning.

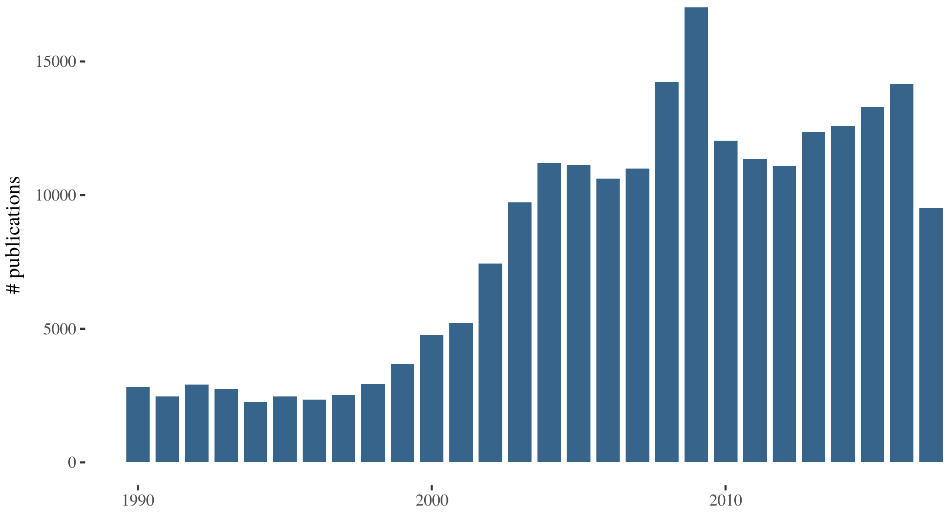


Figure 11: Number of publications (1990-2017)

In 2004 the 10,000 mark was crossed for the first time. Since then numbers have been around 13,000, but with a slight increase in recent years towards the 15,000 mark. The 15,000 mark had been reached before, in 2009. In general, patenting has increased over the years.

When looking at the geographical zones of patent filing, it becomes evident that China surfaces as an innovator in recent years. In the present database, China surpassed the US in terms of patent filing in 2014, being second after the World Intellectual Property Office (WIPO) (Figure 12).

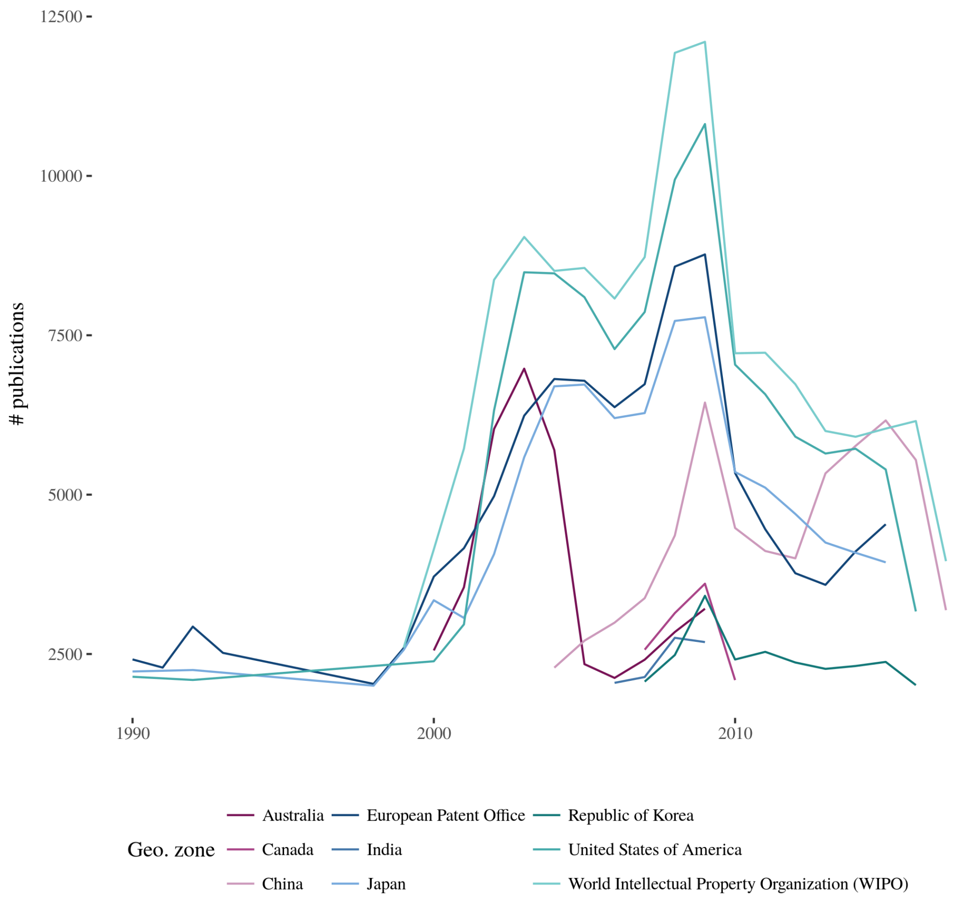


Figure 12: Number of publications per geographical zone of filing (1990-2017); (n>2000)

Since the thesis centers around the pharmaceutical industry in China, it is of importance to look at the number of publications in China seperately (Figure 13). The dataset of Chinese patents covers 69.923 Chinese patents in total, from 1990 to 2017. Since 1990, patent filing has increased rapidly. Nonetheless, there has been a slight decrease from 2009 to 2010, but patenting rose afterward again. In 2015, Chinese patenting reached its maximum, with around 6.000 patents. Since 2015 patenting has decreased. The low number in 2017, however, might be due to the database only covering patents until October 2017. Overall, the Chinese pattern of patenting is similar to the distribution found in the patenting behavior of the whole dataset. The graphic confirms the trend highlighted by experts and journalists (Ovide 2017) that China is emerging as an innovator.

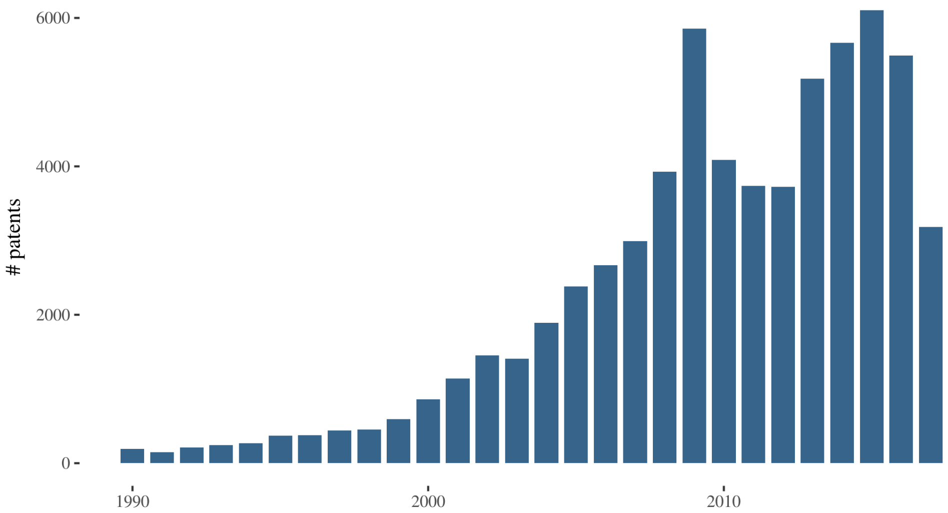


Figure 13: Number of publications in China (1990-2017)

5.2.2.2 Patent Assignees

The following section allows us to have a closer look at the patent assignees of our database. First, we will assess the landscape based on the general patent assignees for the time span 1990 to 2017, provided by the Derwent World Patent Index. Then we will look at the patent assignees of Chinese patents and finally at the patent assignees of patents that had the SIPO as the first geographical zone of filing.

Patent Assignees from 1990 to 2017

The following figure (Figure 14) reveals that the *big pharma* companies figure as assignees with the most patents from 1990 to 2017, although the University of California is present as well. It is important to remark that this description does not regroup all subsidiaries of a company; hence this explains why some entities are represented twice.

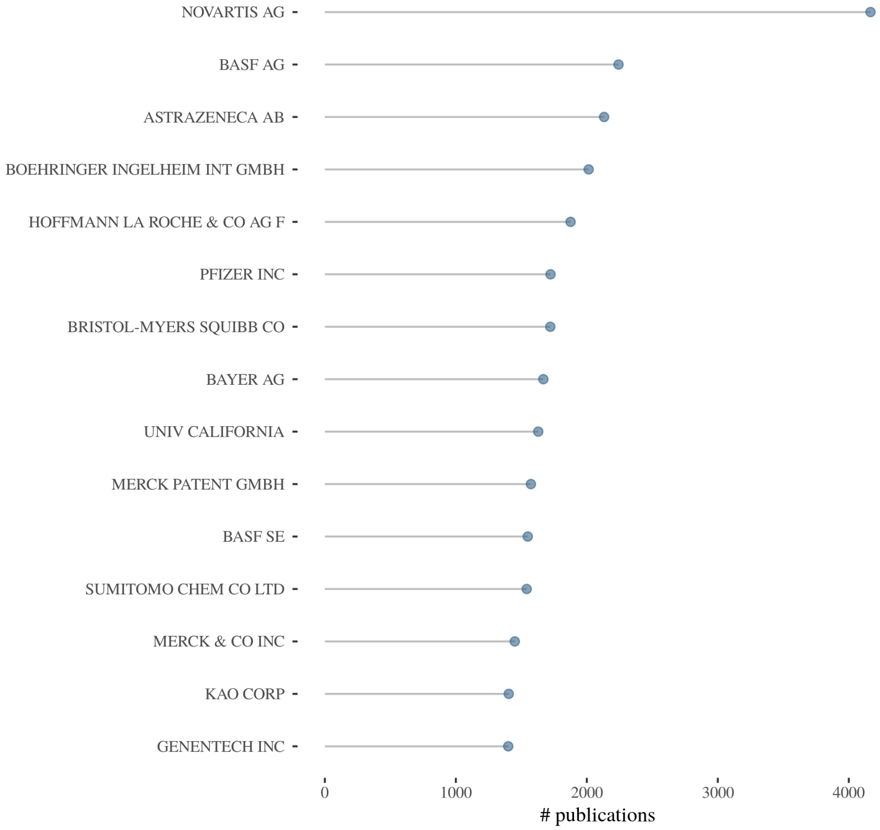


Figure 14: Top 15 patent holders - all geo. zones (1990-2017)

Figure 14 reveals that Novartis is the company with the most patents from 1990 to 2017, with around 4,000 patents. The BASF AG follows it with around 2,000. The big pharma companies are represented in this figure as having more than 1,500 patents. Apart from the typical big pharma companies, we find companies such as BASF, Sumitomo Chem, Kao Corp, Genentech, and the University of California [7]. While most companies are North-American/European, the Kao Corp, and Sumitomo are Japanese.

Patent Assignees in China from 1990 to 2017

Patent holders of Chinese patents are similar to the assignees overall, in the sense that the *big pharma* companies are represented (Figure 15). 4 out of the 15 institutions have changed. The University of California, Sumitomo, the Kao Corp, and the Bayer AG are not present in the top 15 anymore, being replaced by Western pharmaceutical firms: Janssen Pharm NV, DSM IP Assets BV, Schering Corp, and WYETH. While the Novartis AG maintains its top position with around 3,000 patents, the other institutions change their positioning slightly. This finding holds true when splitting the dataset into two time frames, 1990 to 2005 and 2006 to 2017. Thus, the big patent holders have been able to adapt to changes in the

[7] Some companies are represented twice in the figure, for instance, the BASF AG and the BASF SE.

industry as well as economic structure and persist; as Achilladelis and Antonakis (2001) indicated. The top 15 patent holders account for 17,020 patents, thus around 24.3% of the whole dataset.

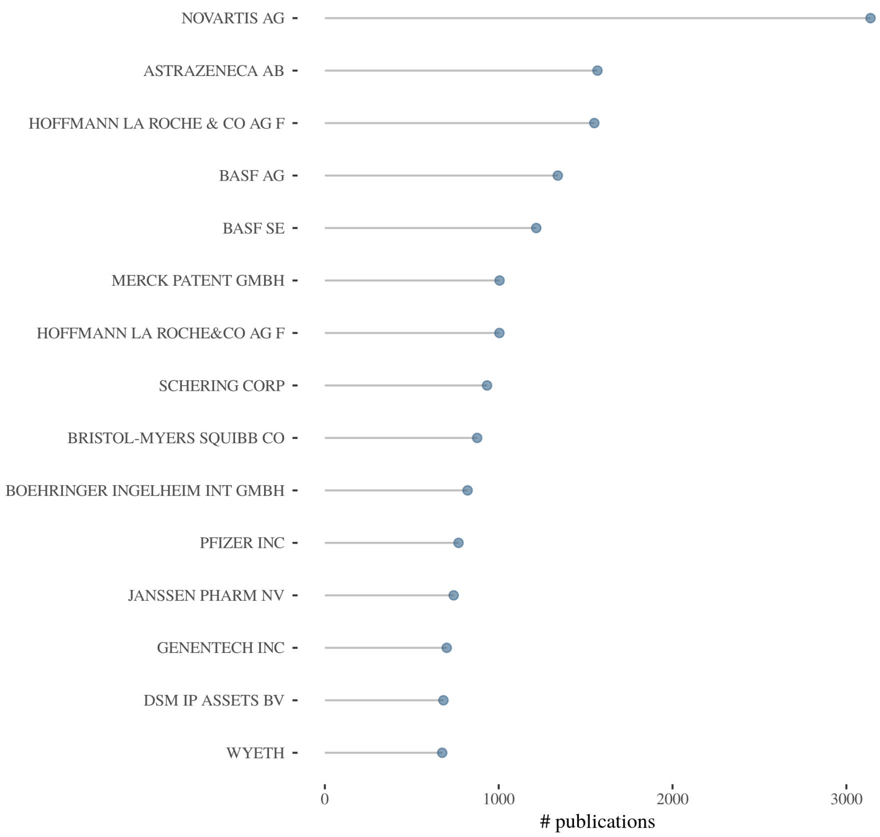


Figure 15: Top 15 patent holders - geo. zone of China (1990-2017)

Due to the global character of the pharmaceutical industry, it deems unsurprising that the top 15 patent assignees in China are big foreign pharmaceutical companies. Nonetheless, it seems reasonable to look into the patent holders of patents filed first in China. This distinction enables us to assess where innovation in China is being sourced from (Figure 16).

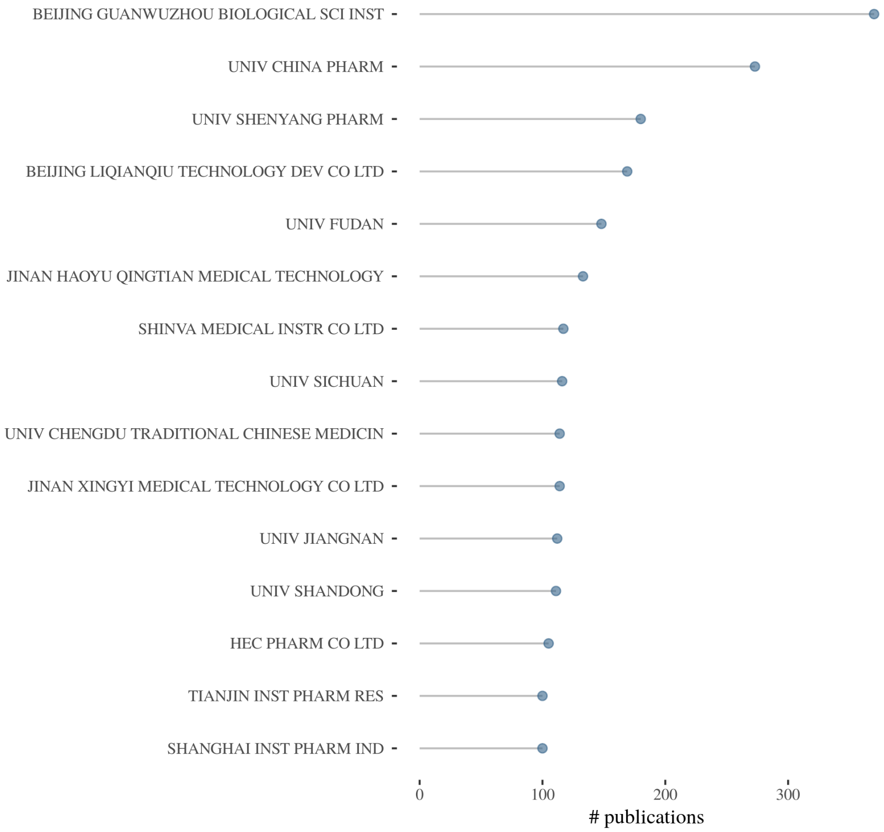


Figure 16: Top 15 patent holders - China 1st geo. zone of filing (1990-2017)

When selecting patents whose first geographical zone of publication is China, it is revealed that the most significant patent holders are Chinese universities or research institutions. The Beijing Guangwuzhou Biological Science Institute figures in top position with around 400 patents. Thus, the patent count of patents filed in China in general and those originating from China differ considerably. Nonetheless, China has become a focus of patenting attention. Big pharmaceutical firms show considerable interest in patenting their innovation in China. Moreover, in the case of patents originating in China, the top 15 patent assignees indicate an interest of Chinese institutions in patenting.

5.2.2.3 Thematic Fields

The Derwent World Patents Index allows for a preliminary analysis of the classification of pharmaceutical patents. While overall categories are indicated, the Derwent class code and the Derwent manual codes allow a more accurate treatment of the present database.

By indication

By simple indication, it is not surprising that *Chemistry*, *Pharmacology Pharmacy* and *Biotechnology Applied Microbiology* are the fields responsible for most of the patents from 1990 to 2017. However, the following figure (Figure 17) reveals that patents related to pharmaceuticals touch upon various other fields as well.

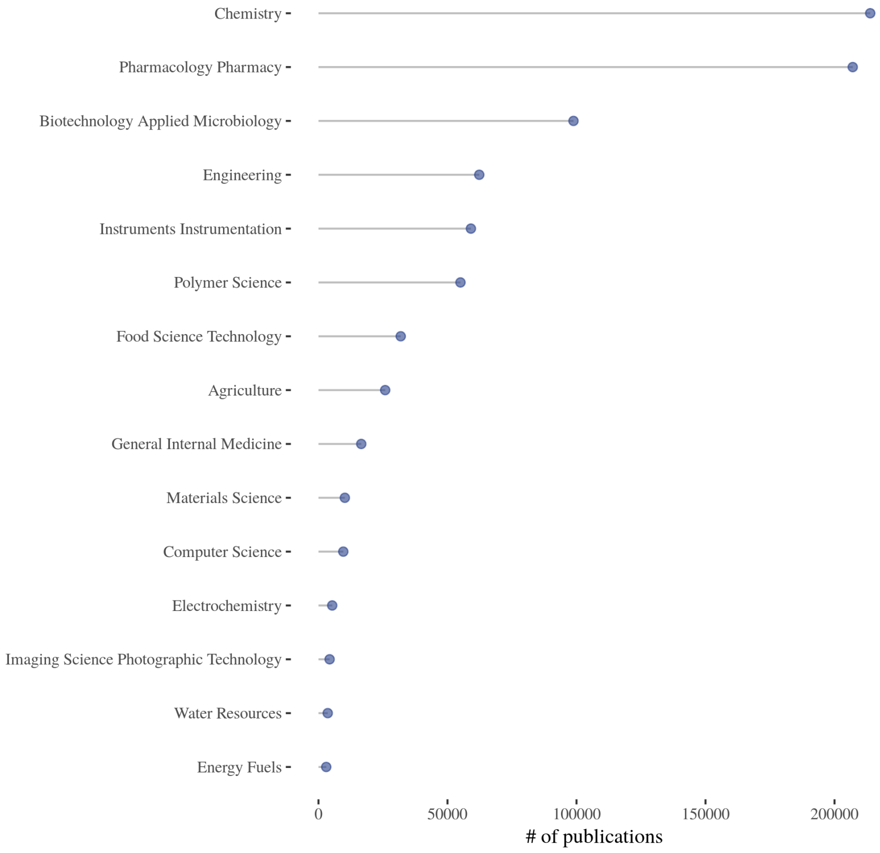


Figure 17: Thematics of pharmaceutical patents

By Derwent Class

The Derwent World Patent Index provides a simple classification into three broad areas: a chemical (A-M), an engineering (P-Q) and an electrical & electronic (S-X) area. The respective letters represent a section of the area. The chemical area contains the pharmaceutical section represented by the letter B. The following figure (Figure 18) reveals the patent count for the top 10 Derwent class codes in China. While five of the class codes are in the *pharmaceuticals section* (B), three are in the *Food, Detergents, Water Treatment and Biotechnology section* (D), and each one in the *Polymers and Plastics* (A), and *Instrumentation, Measuring and Testing section* (S). Hence, while the majority is classified as patents in the *pharmaceuticals section*, a considerable amount is filed in the *Food, Detergents, Water Treatment and Biotechnology section*, also. As explained in the panorama, pharmaceutical firms diversified (Chervenak 2005).

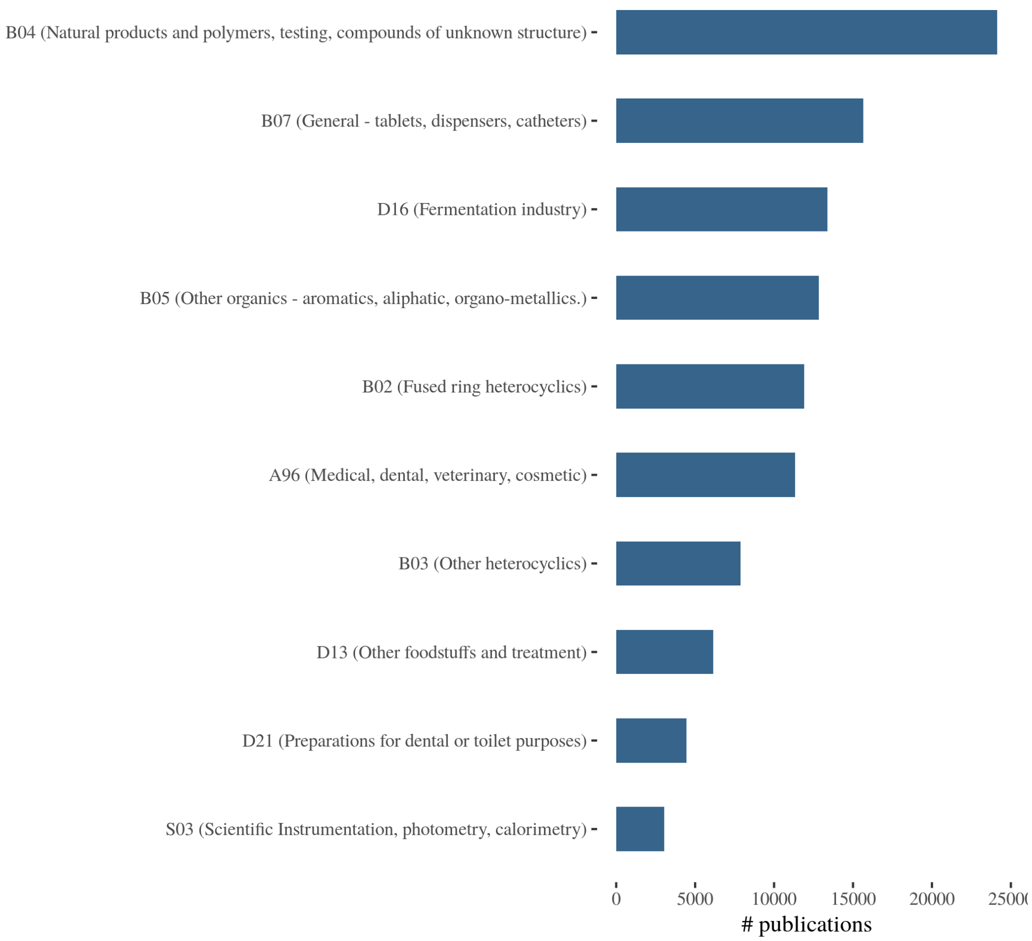


Figure 18: Top 10 Derwent classes of Chinese patents (1990-2017)

The number that follows the letter represents a subsection of the letter section. Hence, the alphanumeric combination allows to identify the more specific thematics that are covered the most by patents from 1990 to 2017. The graphic above reveals that around 25,000 patents are considered patents on natural products and polymers (*B04*), the Derwent class with most of the patents. When looking at the development of publications in the pharmaceutical sections, the class *B04* figures at the top since the year 2000 (Appendix: Figure 40). While this already allows to identify the more specific thematics of the patent database, the Derwent World Patents Index introduced another code, the Derwent manual code that classifies the novelty of patents.

By Derwent manual code

The Derwent manual code distinguishes the same broad sections A to X; each section, however, contains different subsections. A series of alphanumerics follow these. An example of a manual code is A01-A01A1.

The following figure (Figure 19) illustrates the attribution of the manual codes to the patent base of Chinese patents. It is important to mention that a patent can get assigned several codes.

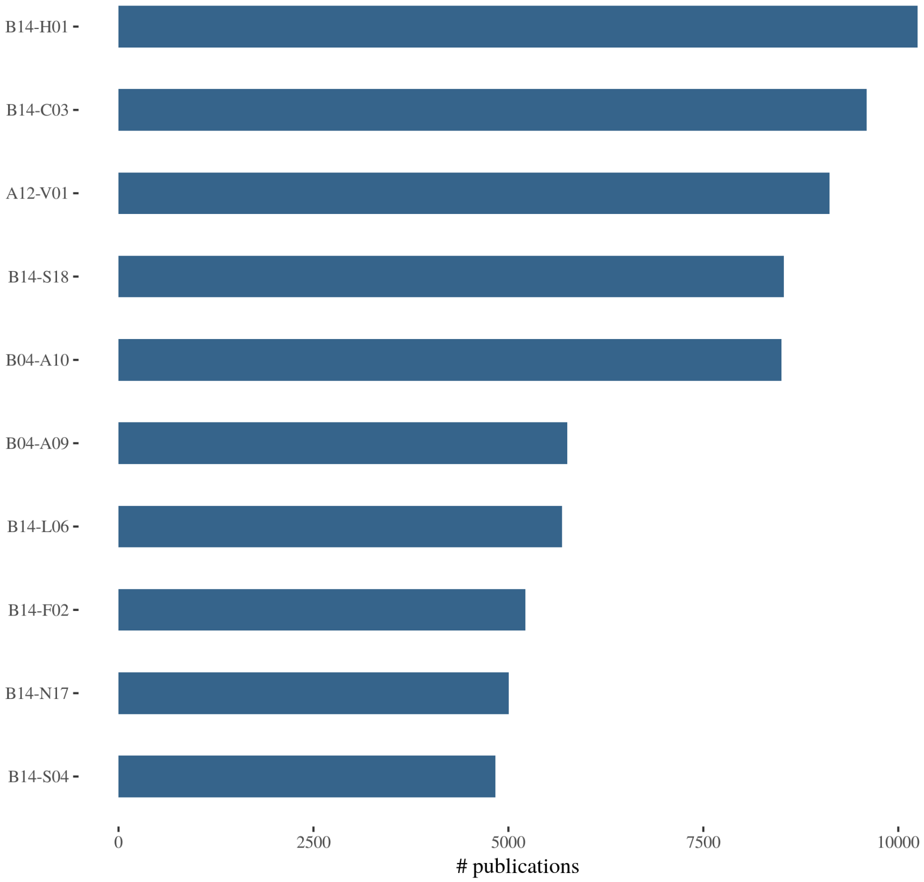


Figure 19: Top 10 manual codes of Chinese patents (1990-2017)

Table 4: Signification of top 10 manual codes in China

Manual Code	**Signification**		
B14-H01	Pharmaceutical activities	Cancer related drugs	Anticancer general and other
B14-C03	Pharmaceutical activities	Anaesthetics and drugs relieving fever, inflammation and pain	Antiinflammatory general
A12-V01	Polymer applications	Medical, dental, cosmetics and veterinary	Medicines, pharmaceuticals
B14-S18	Pharmaceutical activities	Anaesthetics and drugs relieving fever, inflammation and pain	Drug combination
B04-A10	Natural products (or genetically engineered), polymers	Alkaloids, plant extracts	Plant extracts general and other
B04-A09	Natural products (or genetically engineered), polymers	Alkaloids, plant extracts	Plant parts general and other
B14-L06	Pharmaceutical activities	Agonists/mimetics and antagonists/inhibitors not covered elsewhere	Antagonist, inhibitor, antimetabolite general and other
B14-F02	Pharmaceutical activities	Drugs acting on the blood and cardiovascular system	Circulatory active general and other
B14-N17	Pharmaceutical activities	Organs	Skin treatment general and other
B14-S04	Pharmaceutical activities	Miscellaneous activity terms	Diabetes

When focusing solely on the section and subsection of these manual codes, it is revealed that B14 is the section most patents are attributed to (Figure 19 and Table 4). This is not surprising since B14 stands for *Pharmaceutical Activities.* It is followed by B04 in the same section that represents *Natural Products (or Genetically Engineered), Polymers* and by A12 representing *Polymer Applications.* Notice that the manual codes of the Derwent World Patent Index, representing the novel aspect of the patent, integrate genetically engineered products into the pharmaceutical section, while the Derwent class codes do not. This first alphanumeric combination before the hyphen indicates the overall section. It is followed by another alphanumeric combination that represents a specific purpose for treatment in the respective section. Hence the manual code B14-H01 which is represented the most in the patent database is part of the section B14 *Pharmaceutical Activities* and more specifically, -H01 *Anticancer general and other* [8]. The second most recurring manual code is B14-C03, which stands for the treatment purpose *Antiinflammatory general* in the *Pharmaceutical Activities - Anaesthetics and drugs relieving fever, inflammation and pain* section. It is followed by A12-V01, part of A12-V *(Polymer Applications - Medical, dental, cosmetics and veterinary)* and stands for *Medicines, pharmaceuticals.* (For an explanation of all manual codes see Table 4).

When analyzing the manual codes closer, it is noticed that the section B14, although the most often represented is distributed over several subsections; the same goes for B04.

While helpful, the manual codes in their entirety provide such a detailed classification of the novelty that it is difficult to see beyond the details provided. Moreover, they cover a smaller part of the Chinese patents. Hence, in the following solely the first four alphanumeric characters will be taken into account [9] (Figure 20).

[8] B14-H alone stands for all cancer-related drugs.

[9] Patents referring to several manual codes beginning with the same four digit alphanumeric combination will be taken into account more than once, allowing to describe the occurrence of a manual code section.

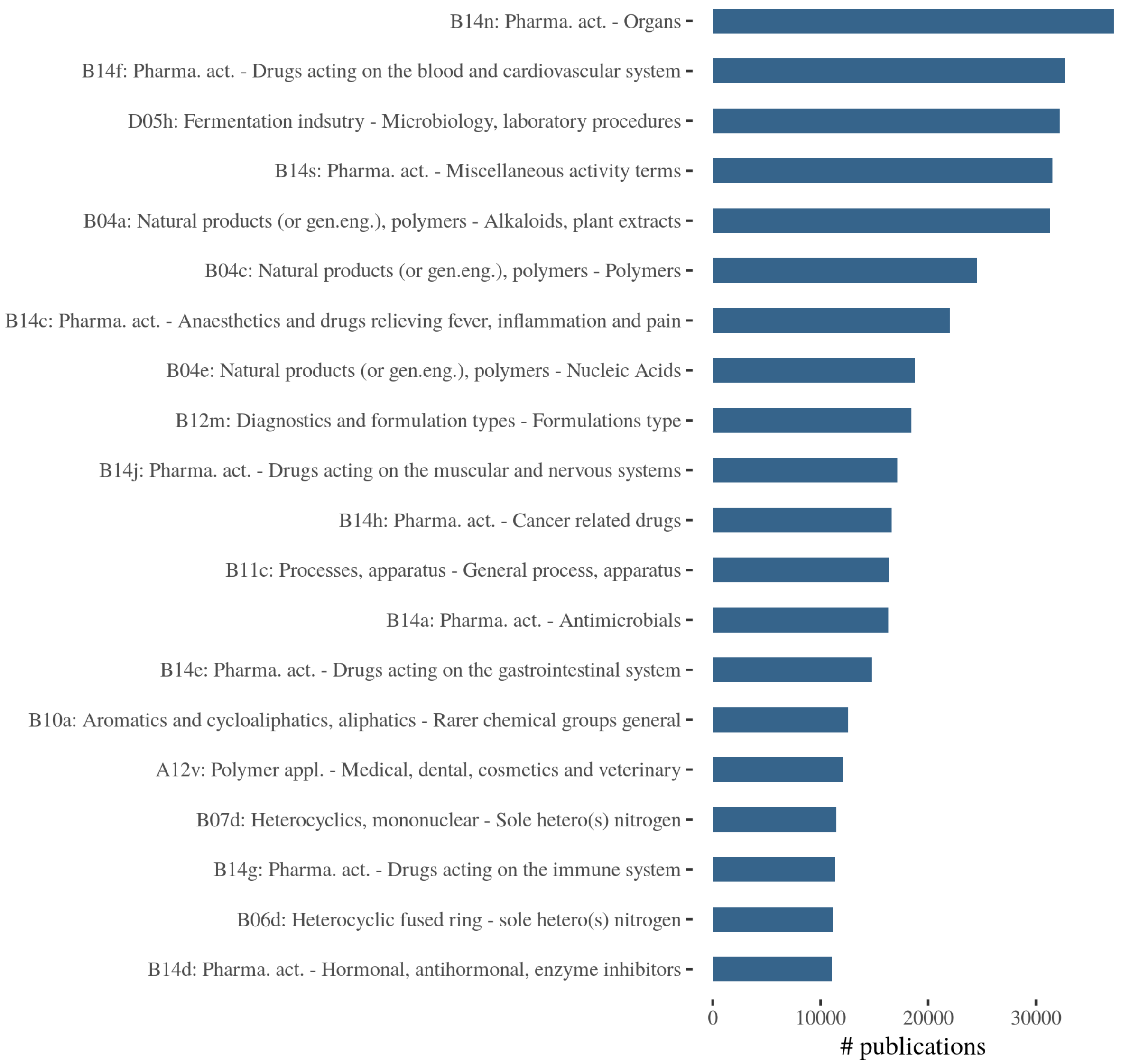


Figure 20: Top 20 manual code sections of Chinese patents (1990-2017)

From Figure 20, it is obvious that the look at the four digit alphanumeric combination representing classification yields a better description when one wants to look at the entire Chinese dataset. It allows a detailed but not too narrow description of the classifications most present in the database. A three digit alphanumeric combination instead of a four digit alphanumeric combination reveals a concentration on B14 _ Pharmaceutical activities_ and B04 *Natural products (or genetically engineered), polymers*, the former occurring approximately 225,000 times and the latter 125,000 times, followed by B10 with around 45,000 occurrences. While a statement for itself, the section B14 *Pharmaceutical activities* is too broad to reveal anything specific.

Hence, the four digit alphanumeric combination is preferred. It reveals that B14-N *Pharmaceutical activities - Organs* occurred around 35,000 times in the patent database, followed by B14-F *Pharmaceutical activities - Drugs acting on the blood and cardiovascular system*, D05-H *Fermentation industry - Microbiology, laboratory procedures*, B14-S *Pharmaceutical activities - Miscellaneous activity terms* and B04-A *Natural products (or genetically engineered), polymers - Alkaloids, plant extracts* (Figure 20). When comparing the occurrences of manual code sections to the number of patents in a manual code section, the top 20 manual code sections are similar except for B07-D and B10-A that are replaced by B04-B and B04-N. Moreover, the number of patents attributed to each are lower. The top 5 in that case are: B14-S (around 20,000 patents), B14-N (around 17,500 patents), B04-C (around 15,000 patents), B14-F (around 14,000 patents) and B14-C (around 13,500 patents) (Appendix: Figure 41).

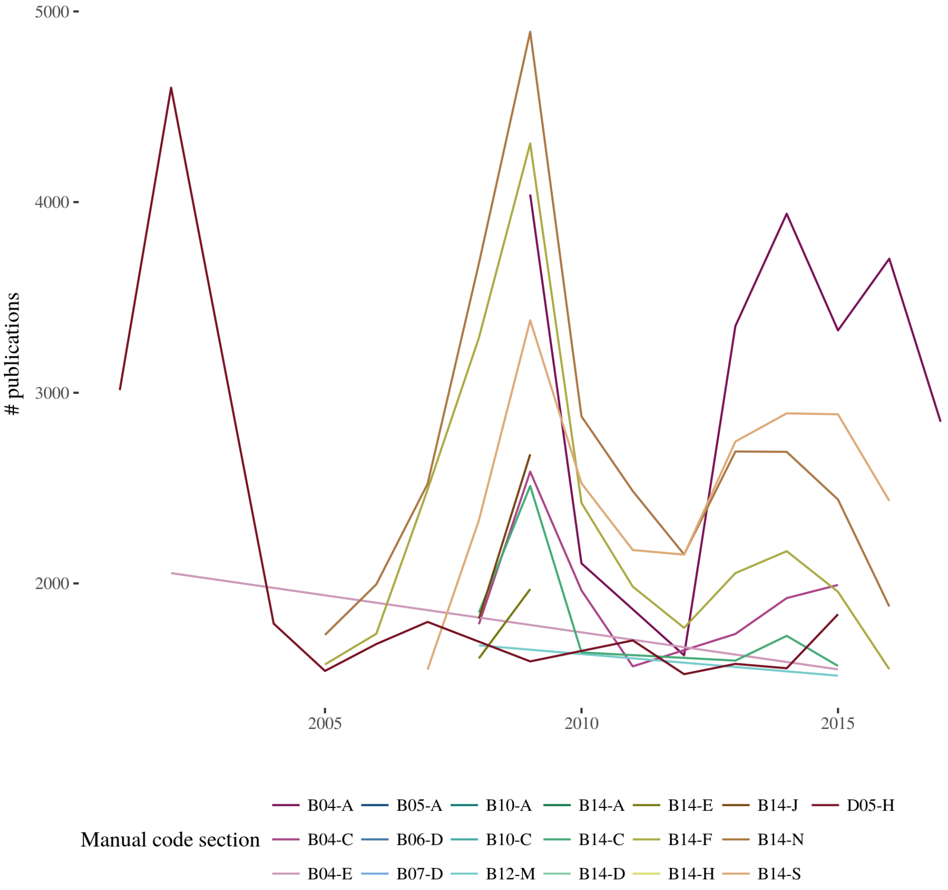


Figure 21: Development of the top manual codes of Chinese patents, n>1500 (1990-2017)

The main area of research has changed over the years (Figure 21). While at the beginning of the 21st century D05-H *Fermentation industry - Microbiology, laboratory procedures* was predominant, it lost its importance in recent years. The same goes for B04-E *Natural products (or genetically engineered), polymers - Polymers* that was second at the beginning of the 21st century. Finally, some manual code sections saw an uprising in

China in 2009, namely B14-N *Pharmaceutical activities - Organs*, B14-F *Pharmaceutical activities - Drugs acting on the blood and cardiovascular system*, B04-A *Natural products (or genetically engineered), polymers - Alkaloids, plant extracts*, and B14-S *Pharmaceutical activities - Miscellaneous activity terms.* Out of these, B04-A is the section with the most occurrences since 2011, followed by B14-S, B14-N, and B14-F. Thus, while Figure 20 revealed that B14-N *Pharmaceutical activities - Organs* had most of the occurrences in Chinese patents, the last graphic highlights that there has been recently a shift towards innovation in *Natural products (or genetically engineered), polymers - Alkaloids, plant extracts* in China. Nonetheless, the top 5 manual code sections from 1990 to 2017 are among the top sections in recent years.

By therapeutic groups

While the thematic fields by indication and by Derwent and manual codes have provided insights into the area of research from a thematic and research field point of view, it does not provide insights into the therapeutic groups covered by Chinese patents. Thus, a content analysis based on a framing strategy has been employed, allowing us to reveal the Chinese patents' focus on therapeutic groups. Each patent has only been taken into account once, despite some having been filed twice in China. Through the unique patent index duplicates could be identified and sorted out.

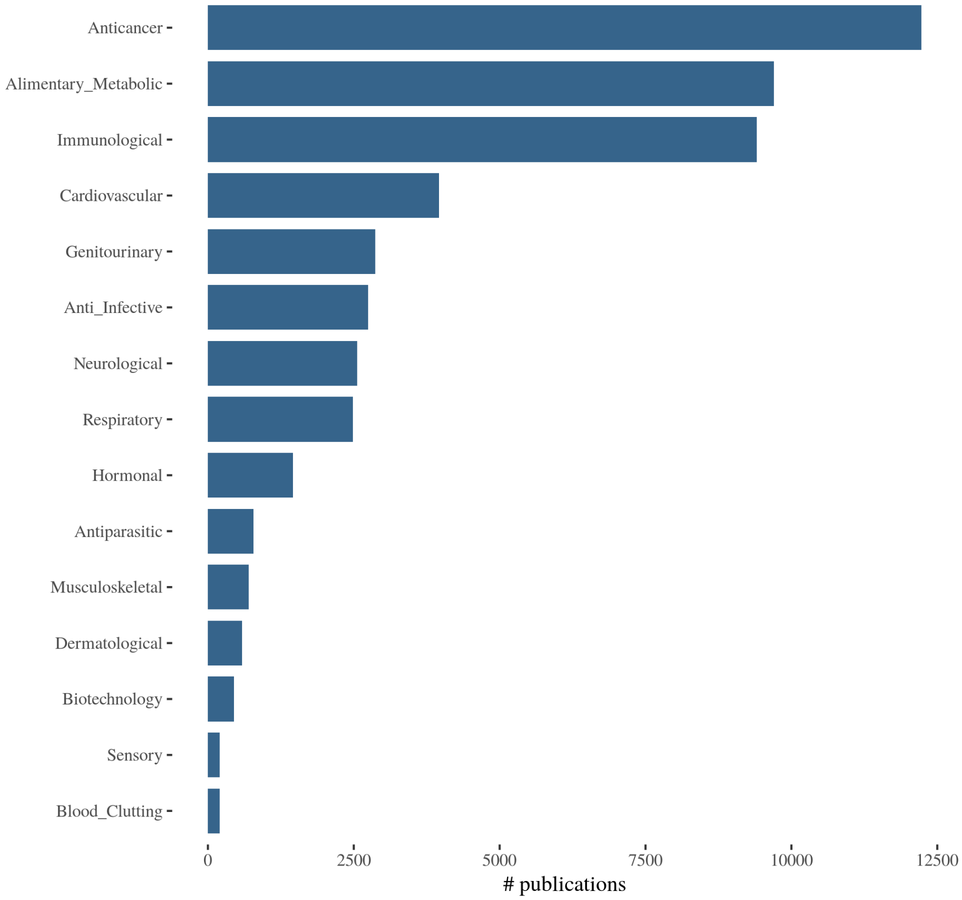


Figure 22: Patent count per therapeutic group in China (1990-2017)

Chinese patents mostly belong to the *Anticancer* category, with around 12,500 patents (Figure 22). However, the therapeutic groups *Alimentary metabolic* [10] and *Immunological* each cover around 10,000 patents also.

[10] The therapeutic group "Alimentary metabolic" comprises diabetes, for instance.

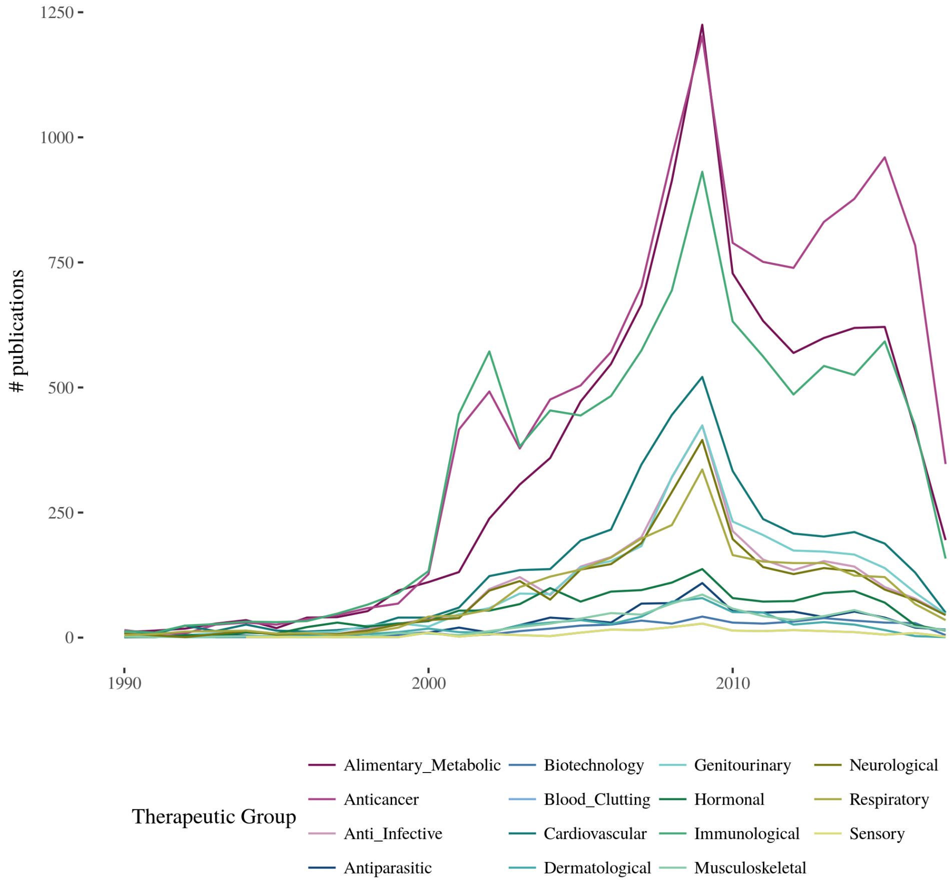


Figure 23: Development of patent count per therapeutic group in China (1990-2017)

These three therapeutic groups *Anticancer*, *Alimentary metabolic* and *Immunological* have each experienced tremendous growth in the 2000s (Figure 23). Other therapeutic groups have increased as well, although not to the same extent. Overall, they follow the distribution of patenting we found in Figure 13.

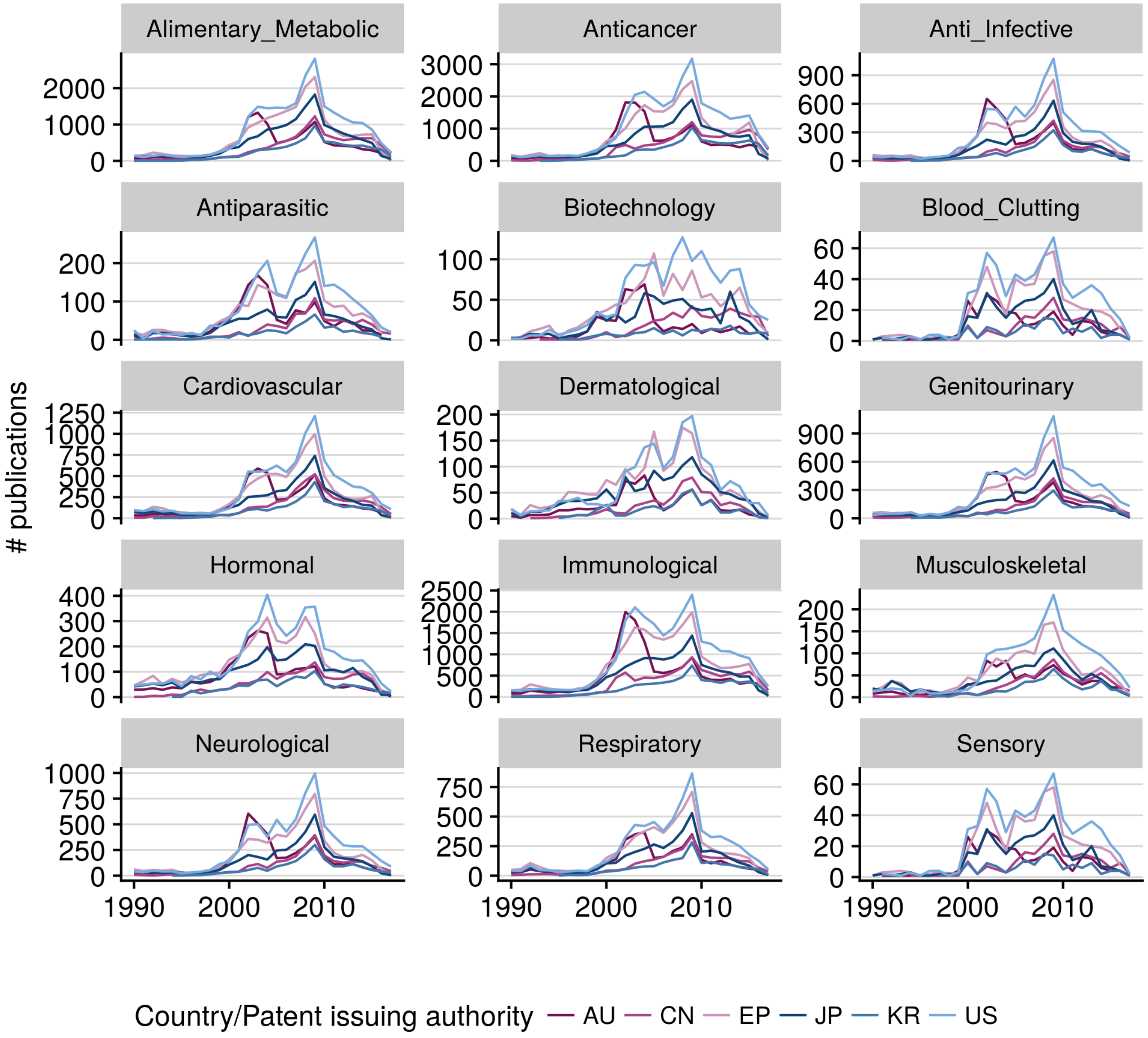


Figure 24: Comparison of patent count per therapeutic group between China/US/Japan/Korea/EPO/Australia (1990-2017)

In the following figure (Figure 24) the development of patenting in the therapeutic groups has been compared between China, the US, Japan, Korea, the European Patent Office and Australia [11]. In the last decade, the amount of Chinese patents per therapeutic group has been getting closer to the other patent-issuing authorities. However, this is due to a decrease in patenting in these countries.

[11] These countries/patent-issuing authorities are preceding China in terms of patent count of our therapeutic group patent dataset. The WIPO figures at place 1, but has not been taken into account since it figures in the top position for each therapeutic group.

In conclusion, while allowing to identify specific therapeutic groups, China seems to have been building up capabilities in, the explanatory power of the results is limited since the dataset on therapeutic groups covers 24,087 of 69,923 Chinese patents in total. In the following, we will focus on a computational classification that provides an unbiased picture of the dataset, since it is not based on the human application of keywords for the classification.

All in all, the Data Science approach that provided these descriptive statistics has provided the first insight into China's emergence in the innovation landscape and its specificities, the yield, however, is limited.

5.3 Methodological Approach

In the following, the chosen methodological approaches for this thesis will be highlighted. Together with text mining, these generate the protocol: the unsupervised (semi-supervised) classification based on a Latent Dirichlet Allocation (LDA) approach, the Jaccard similarity, and a subsequent econometric part.

5.3.1 Step 1: LDA analysis

After having analyzed the structured data of the patents, as well as its patent content leveraging text mining and elaborating on the descriptive statistics, the protocol further foresees an LDA analysis. While the descriptive statistics compared patents filed at the Chinese patent office in part to other patent offices, the LDA focuses solely on Chinese patents, including Chinese patents that are part of patent families. All Chinese patent abstracts will be gathered through an unsupervised learning method with a Latent Dirichlet Allocation (LDA). It can be considered a method of classification. The LDA enables us to identify groups of words and ideas together. These groups of terms allow us to have a deeper understanding of the topics of the pharmaceutical patents as well as enable us to highlight connections between them. This step of the analysis will answer to question (1): *Where lies the focus of innovation in the Chinese pharmaceutical industry? For the whole time span of 1990 to 2017? For 1990 to 2000? For 2001 to 2017?*

We are particularly interested in the LDA analysis since it allows the modeling of topics within documents without having to go through the dataset manually. In case of the present dataset, the 24,087 Chinese patents [12] would have taken approximately 3 years to assign to topics, when assuming that it is undertaken by one researcher and it takes an hour to read one patent. The applicability of this research duration is doubtful, due to the aging of the dataset, among other things. Moreover, it would have been difficult to classify the large dataset manually without pre-defined classifications. There would have been the uncertainty of researcher bias. A computational, Data Science technique, obliterates this specific bias, using statistical methods to derive topics from documents. It renders the topic modeling verifiable for other researchers, and hence the results are more robust. To our knowledge, researchers have not applied topic modeling and an LDA on pharmaceutical patents yet. In general, no Data Science approach has been used to characterize the industry's innovations.

[12] These 24,087 Chinese patents are the ones that figure among patents covering the 15 therapeutic groups. This subset was chosen, since the results will be compared to the therapeutic groups.

An LDA is “a generative probabilistic model of a corpus. The basic idea is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words” (Blei, Ng, and Jordan 2003: p. 996). Adopting a bag-of-words approach, a word can thereby be allocated to several topics, since the bag-of-words creates a topic. Moreover, patents can be associated with several topics. Blei, Ng, and Jordan (2003) explain that the LDA performs better than simpler latent variable models for texts, such as the unigram model, a mixture of unigrams and the pLSI model.

Before applying the LDA analysis, further data cleaning is required. While the data has been cleaned before, when it was decided to focus solely on patent applications, and thus the number of documents was reduced, the LDA analysis necessitates further cleaning. So-called stop words were excluded by applying a filter to the dataset. Stop words thereby refer to pronoun forms (i.e. “I”, “yours”), verb forms (i.e. “am”, “has”, “did”), auxiliaries (i.e. “should”, “will”), articles (i.e. “the”), verbs or auxiliaries with negations (i.e. “should not”), compound forms (i.e. “it’s”) and other words such as here, by, of, all, other, etc. The stop word list *snowball* in R has been used. Apart from these stop words, some words were eliminated by the researchers due to their occurrence throughout all topics. If not eliminated they might skew the results. These are pharmaceutical, compris*, composition, e.g., as well as all numbers. Moreover, the dataset was filtered in order to keep terms that are found at least 50 times. Then, by using a VEM algorithm (variational expected maximization) (Blei, Ng, and Jordan 2003), from the topicmodels library in R by Grün and Hornik (2011), we segmented the dataset into five different clusters and provided the most important terms associated to each topic/category of patents. The segmentation has been undertaken with several amounts of clusters. Finally, the segmentation in five clusters deemed the most appropriate since the overlap was minimized and it yielded a clear distinction between topics.

These groups of words enable us to describe the automatically derived classifications in the dataset. Moreover, we can compare each cluster of words to the fifteen therapeutic groups that were built to measure the fit of the therapeutic groups. it has to be noted that a cluster of words might be associated with more than one specific topic and that the words describing the classifications might reveal connections between the automatically derived classifications. While this provides first indices towards innovation dynamics, a second method will be employed that enables us to paint a more thorough picture of the innovation dynamics.

5.3.2 Step 2: Jaccard similarity coefficient

The text-as-data approach and more specifically the Jaccard similarity coefficient, used in this thesis allowes to measure the text similarity of the abstracts of the patents with each other and between whole patent corpora. While text analyses have been hand-coded in the past, an automated text analysis has been gaining ground in research. Despite there being advantages to human coding, such as meaningful variation in language, it is an expensive and time-consuming activity. The analysis of 228,108 patents would not be possible in the scope of a thesis if hand-coding had to be used, Data Science techniques, however, allow to do so. Moreover, hand-coding risks a coder bias. Finally, patents due to their standardized nature of scientific language represent an excellent base for text-as-data approach, and since the present patent database is in English despite the patents being filed at different patent offices around the world, it is possible to compare all abstracts to each other.

This thesis will measure the similarity between patent thematics leveraging the Jaccard index. We refrained from using the patent thematics identified in the previous step, since we prefered the measurment of similarities across more than the five topics identified and across a more systematic classification. Thus, patent thematics are represented by the Derwent manual code sections. Since the descriptive statistics revealed the focus on a few classes, *B04, and B14*, the subcategories will be used. Representing a proof of concept, the analysis will be limited to the top 20 "4 digits" manual code classifications. This approach will enable us to measure the dynamics between the most used classifications of pharmaceutical patents. It constitutes a completely new approach to innovation dynamics and answers to question (2): *What kind of dynamics do we find between patent thematics/classes? More precisely: What is the relationship of patent classes across time?*

From an operational perspective, the dataset of Chinese patents is, therefore, subset into 537 RData files on the 20 classifications from 1990 to 2017. Each manual code section and each year build a separate file, allowing us to compare the sections across time with each other. In a next step, the abstracts of the patents of each file will be pasted together, creating one single text. These texts will be each split in 5-character gram components, following Alschner and Skougarevskiy (2016). The word "Genetic engineering" would hence be split into: "geneti", "enetic", "netic_", "etic_e", "tic_en", "ic_eng", etc. In our case, the computational segmentation of the texts into 5-character gram components took around eight days. The advantage of this approach is that it retains the word order in contrast to bag-of-words approaches that measure the word occurrence. When analyzing patents, it is beneficial to take the word order into account due to the scientific nature of the abstracts. Pharmaceutical components might occur in different contexts. Thus a bag-of-words approach would skew the results.

After having split the text into 5-character gram components, the text similarity will be measured. To this end, the Jaccard similarity coefficient will be used. It allows to compare the set of common items of two entities, text A_i and text A_j, to the set of unique items of these entities. In our case, this will be the aggregated patent sets per manual code section per year. This measurement of the similarity between the texts split in 5-character grams took another four days.

$$s_{ij} = \frac{|A_i \cap A_j|}{|A_i \cup A_j|}$$

The Jaccard similarity coefficient which can be calculated by using the formula above goes back to Paul Jaccard, a botany professor in Switzerland, who developed the similarity measure to analyze the distribution of flower species in different areas in the Alps and the Jura (Jaccard 1902). In international business, the Jaccard similarity has been used to measure text similarity between trade agreements. Alschner, Seiermann, and Skougarevskiy (2017) used it to analyze the impact of the TPP on trade between member countries. They constructed a gravity dataset and run regressions integrating the text similarity as variable.

While the benefits of textual similarity measures have been highlighted, the approach features limitations. Textual similarity implies a relational concept, hence "[…]requires comparison and context to become meaningful." (Alschner, Seiermann, and Skougarevskiy 2017: p. 6). Moreover, "[…] textual variation does not always imply a variation of meaning." (Alschner, Seiermann, and Skougarevskiy 2017: p. 6). Nonetheless, it is a promising approach for the comparison of patents and the identification of innovation patterns.

Finally, since the context is important when the similarities of manual code sections across time and classes are of interest, a data frame is created with new variables. The data frame included the names of the files that we compared, the manual code sections, the year of each file, an ID for the files that we compared as well as every single file and the Jaccard similarity. Moreover, the patent count for each file was added. The manual code section A12v of the year 1990, for instance, encompasses eight patents. A separate variable for the sum of the patent count was created. Furthermore, we added a variable time lag. Files that were compared with itself were excluded from the data frame since their similarity of 100% would skew the analysis. Since the code developed for the data analysis yielded the Jaccard similarity measurement in a matrix format (537ˆ2), the outcome included duplicate data. Thus, we created two separate data frames, one including duplicates and one excluding them, facilitating the further analysis, for instance filtering. Finally, descriptive statistics will enable to explore the Jaccard similarity in the context of these variables.

5.3.3 Step 3: Econometrics

After an exhaustive analysis of the Jaccard similarity between patent classes, an econometric analysis looks at the impact of the Jaccard similarity through time on the patent count. It answers to question (3): *What is the relationship between the similarity of patent classes and the patent count when taking into account the time lag?* The results let on the innovation cycle of the pharmaceutical industry in China, meaning the time it takes for innovations benefitting from knolwedge flows across research fields to yield innovations. Thus, we introduce a new explanatory variable for the patent count. While not exhaustive, the analysis is considered a proof of concept of the usefulness of this new variable.

The dataset used for this econometric analysis was created in step 2 and constitutes panel data, of Jaccard similarities per manual code sections and years and associated with their respective patent count. Thereby the panel data allows to "distinguish inter-individual differences from intra-individual differences" (Baltagi 2001: p.349). For the present analysis, it is of importance that a panel data approach assesses individual outcomes, in this case, mirrored by the manual code sections, and thus does not claim to be holistic. It is a simplification of the relationship among variables. The analysis is of experimental character.

For the further analysis, the dataset is wrangled as to compose the aggregated mean Jaccard similarity per section, per year, across sections. Each time period originally exhibits N = 20 x 19 cross-sectional observations. The data covered the similarities from 1990 to 2017, yielding a total sample of n = 380 x 28. For the purpose of this analysis, the aggregated mean Jaccard similarity per year and section was of interest, as well as the aggregated sum of the patent count per section per year. Since the panel data was unbalanced due to some sections not exhibiting patents in some years, we reduced the dataset to the term of 1994 to 2017. Thus, the data constituted a balanced panel dataset. This step was undertaken since the further analysis would have faced constraints otherwise (Croissant and Millo 2007). The final dataset used for further analysis encompasses 480 observations, n = 20 x 24.

In the following, a panel linear model will be employed measuring the impact of the lagged Jaccard similarity on the patent count. This linear model is based on Croissant (2017)'s plm package. A panel linear model with fixed effect estimation was ultimately chosen, since the impact of variables that vary over time is of interest (Torres-Reyna 2007), and while a poisson regression was considered, since the dependent variable consists of count data, the model's assumptions were not respected and thus the model was refuted ("Poisson Regression | R Data Analysis Examples" 2018). The negative binomial regression was not adequate either.

The linear panel model allows for different estimation methods. Following Croissant and Millo (2007), several statistic tests were applied, to determine the best estimation method:

(1) a pool test testing the significant presence of homogeneous coefficients for each individual;

(2) Lagrange Multiplier tests testing the significant presence of random effects (endogenous or not);

(3) a Hausman test characterizing the endogeneity or not of random effects.

The pooled OLS estimation could be refuted through an F-Test, which p-value was below 0.05. The null hypothesis H_0 (*there are homogeneous coefficients*) was refuted. Concerning the Lagrange Multiplier test, the p-value was below 0.05. Thus the null hypothesis H_0 (*the variance of specific individual errors is null*) was refuted. In the last step, the Hausman test has been undertaken rejecting the Null hypothesis H_0 (*the random effects of individuals on errors are exogeneous*), hence that a random effect model would be adequate. In conclusion, the use of a fixed effects model was supported from several sides. Random effects, as well as a pooled OLS estimator could be rejected due to the statistic tests performed.

Having applied the Lagrange Multiplier test on the individual as well as time variable, the necessity of a twoways fixed effects estimation was revealed. The constructed model allows to explore the relationship between the lagged Jaccard similarity and the patent count. Fixing the effects for "individual" and "time", we control for something within the individual and the time that may impact the independent or dependent variable. Moreover, it has the advantage of controlling omitted variable bias (Dranove 2003).

The final fixed effects model, also called unobserved effects model (Croissant and Millo 2008), is of form:

$$y_{it} = \alpha + \beta^T x_{it} + \mu_i + \lambda_t + \epsilon_{it}$$

i = 1, … n is the individual (manual code section) and index t = 1, … T is the time index. μ_i represents the individual error component and λ_t the time error component. Since this model fixes both time and individual, the error u_{it} is composed of both μ_i, λ_t and a third component, ϵ_{it}, the idiosyncratic error. Both individual and time effects are specific and do not change over time.

Five different models were developed identifying the effect of the Jaccard similarity on the patent count. More specifically, the Jaccard similarity was lagged for year 1, 2, etc. up to 13. While model 1 contained the lagged Jaccard similarity from year 1 to year 9, the second model contained year 1 to 10, etc. up to model 5 from year 1 to year 13. The lagged Jaccard similarity of the five different models was chosen in concordance with the information from the book "Innovation and Marketing in the pharmaceutical industry" from Ding, Eliashberg, and Stremersch (2014). They explained that the R&D process for new drugs from idea to clinical trials takes between 9 to 13 years. Often patenting is delayed as long as possible so that the commercialized drug is protected longer (Ding, Eliashberg, and Stremersch 2014). Therefore, we assume that the time from the idea to clinical trials represents the time from the idea to patenting. The 9 to 13 years were chosen since the lagged Jaccard similarity is assumed to affect the patent count in subsequent years in this time frame.

In a final step, the models were tested for serial correlation and cross-sectional dependence. The Breusch-Godfrey test also called LM test for serial correlation, tests for autocorrelation in the errors in a regression model. Thereby, it tests temporal dependence. The null hypothesis H_0 (*there is no serial correlation*) was rejected with a p-value below 0.05. Furthermore, the Pesaran test was employed, rejecting cross-sectional dependence. It did not reject the null hypothesis H_0 (*there is no cross-sectional dependence*); the p-value was above 0.05. In consequence, the standard errors have to be adjusted. We obtained standard errors consistent with heteroskedasticity and autocorrelation (HAC) by using the Allerano method (Croissant and Millo 2008). Since there are no cross-sectional dependence issues, the method does not have to take them into account. After having adjusted the models with the Allerano method, these models will be compared in their yield of significant variables. Since these models focus on the impact of an individual year's similarity measurement, in a second step, a model was build that separates the analysis in what we call periods of maturity. The impact of the average similarity from 1 to 5 years ago and from 6 to 10 years ago will be assessed on the patent count at a given time. Thus, we split the analysis, assessing whether similarities in the near past affect subsequent patenting and whether similarities in the distant past affect subsequent patenting. When averaging the similarities in periods of 1 to 5 years ago and 6 to 10 years ago this allows to better assess the impact of the periods of maturity of existing patents. Moreover, this allows for a robustness check of the results from the previous models. The explanatory power of the analysis might be higher, when regrouping similarities in periods of maturity than when looking at single years. With regards to the operationalization, we calculated a moving average using the dyplr and zoo package in R. Then the same tests as for the other five models were performed. The model was as the other models corrected for heteroskedasticity and autocorrelation using the Allerano method. Finally, this part of the analysis yielded a model with two independent variables: the similarity from the near past (1-5 years ago) and the similarity from the distant past (6-10 years ago). The results of the latter model will be compared to the results of the fifth models with thirdteen independent variables, one individually lagged similarity for each year. The fifth model was chosen for the comparison since it had the highest R squared.

In conclusion, having built models, respecting the underlying factors, we will be able to assess the impact of similarities between patent thematics on subsequent patent counts. More importantly, we will be able to identify the temporal aspect of it. Thus, this analysis enables us to capture the, what we call, innovation cycle of the pharmaceutical industry through a patent proxy. It constitutes a proof of concept, revealing the explanatory power of the variable similarity.

6 Analyses and results

The following chapter concentrates on the analysis of the 69,923 Chinese patents derived from the Derwent World Patent Index in 2017. While the descriptive statistics have allowed to confirm China's emergence as an innovator, with patents equally being filed by foreign but also domestic entities, the subsequent analysis focuses on strengthening the knowledge gained through the data description, checking its validity through a computational approach and capturing the dynamics of innovation. Leveraging statistical and computational methods, Data Science techniques have enabled us to analyze an enormous dataset comprehensively and plausibly.

6.1 Step 1: Topics modeled from Chinese pharmaceutical patents, an LDA analysis

In what follows, the results of the LDA analysis will be presented, and thereafter interpreted, allowing us to answer the research question (1): *Where lies the focus of innovation in the Chinese pharmaceutical industry? For the whole time span of 1990 to 2017? For 1990 to 2000? For 2001 to 2017?* The LDA has allowed to classify pharmaceutical patents for the first time without relying on the manual classification process of experts or patent holders that classify their innovation themselves. The Data Science approach while having been used to capture technological gaps in high-tech industries (Jun and Sung Park 2013) has not been employed to the pharmaceutical industry, yet.

After having read all 228,108 patents, of which 69,923 Chinese patents, the computational methods were leveraged for an unsupervised automatic classification. The LDA analysis was undertaken on a sample of Chinese patents, namely a subset of the Chinese patent dataset covering the fifteen therapeutic groups; this yields 24,087 patents. Patents were only taken into account once, despite figuring in more than one therapeutic group. The topics were modeled for three different time spans, 1990 to 2017, 1990 to 2000 and 2001 to 2017. The distinction of time spans enabled the researcher to compare the topics modeled from the data across time. Finally, the results were compared to the data description and cross-checked with the therapeutic groups derived through the framing strategy. While applied through computational methods, the framing strategy is close to a manually conducted analysis. Therefore the comparison with the computationally derived topics of the dataset is particularly interesting.

Topic modeling for Chinese patents (1990-2017)

After a data cleaning process described in chapter 5, the dataset of Chinese patents was classified automatically through the LDA analysis, whose results are shown in the following figure (Figure 25).

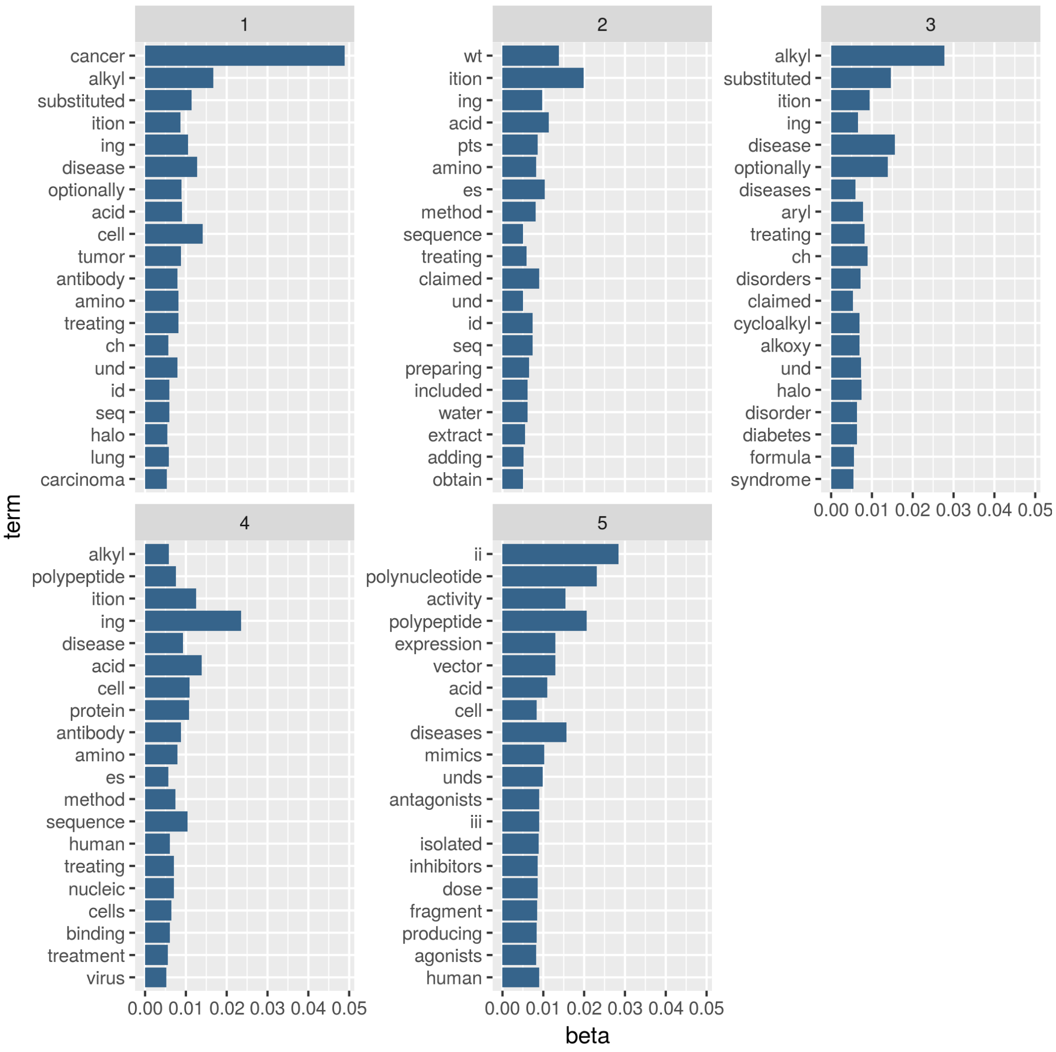


Figure 25: Results LDA analysis - China 1990-2017

The LDA analysis has formed five topics out of the Chinese patent dataset. Each topic is being described by the twenty words that occurred the most. By means of these words, the general thematic of each classification can be identified. Topic 1 is being described with words such as cancer, carcinoma, tumor, lung, and treating. These words indicate that one of the thematics of the Chinese patents is on cancer treatments, and lung cancer can be assumed to be a specific cancer being researched on. Words such as amino, acid, sequence, and method form Topic 2. It is noticeable that this is the topic described the most by verbs. These verbs include extract, adding, obtain, claimed, treating. From the Oxford Dictionary of Biochemistry and Molecular Biology (Cammack 2006a), we know that aminos are the chemical group -NH_2 in an organic molecule. Other words of the top 20 words include acid and sequence hence it might be possible that the topic is on amino acid sequences, also. A string of amino acids in a particular order, called amino acid sequences, constitute a protein molecule. There is, however, another topic modeled with words such as protein, hence the topic 2 is likely on aminos or amino acids and their treatment, i.e., extraction, preparation, etc. Topic 3 was modeled on words such as disease, diseases, disorder, disorders, syndrome, diabetes. These words indicate that the pharmaceutical patents focus in large part on treating diseases and disorders. Moreover, diabetes seems to be, next to cancer, a disease research has been focusing on in China. As mentioned, the fourth topic is related to the second topic on aminos. It centers around polypeptide, protein, amino, disease, antibody, etc. Polypeptides comprise "about 10–20 amino-acid residues connected by peptide linkages" (Cammack 2006c: para. 1). While sometimes used as a substitute for the word protein, we can assume that it is not the case in the scientific environment. After consultation of scientific literature, the connection between antibody, polypeptides, protein, and aminos is disclosed. The Biology Project (University of Arizona 2000) explains that "Antibodies are immune system-related proteins called immunoglobulins. Each antibody consists of four polypeptides– two heavy chains and two light chains joined to form a"Y" shaped molecule. The amino acid sequence in the tips of the "Y" varies greatly among different antibodies" (University of Arizona 2000). Thus, it can be assumed that topic 4 focuses on antibodies which mostly stands for immunological functions. The last topic that has been modeled features polynucleotide, antagonists, inhibitors, fragment, and agonists among others. Polynucleotides thereby refer to a specific form of one-stranded homo- or heteropolymer (Cammack 2006b). While the words antagonist and inhibitor refer to slowing or blocking a specific chemical reaction, the word agonist implies the cause of action. In conclusion, the LDA analysis has roughly pictured the topics Chinese patents focus on: cancer treatments, aminos, and their functioning, diseases (i.e., diabetes), antibodies/immunology, and finally polynucleotides, the specific form of a homo- or heteropolymer.

Topic modeling for Chinese patents (1990-2000)

In view of Chinese patenting increasing rapidly from the year 2000 on (Figure 13), the topic modeling was repeated for the period 1990 to 2000 and 2001 to 2017 separately. The results will highlight whether there are differences in topics modeled between these two periods and the whole time span.

In comparison to the five topics modeled on account of the dataset from 1990 to 2017, the topics for the time span 1990 to 2000 vary slightly (Figure 26). In general, the topics on cancer (topic 3), diabetes (topic 2) and antibodies/immunology (topic 4) are recurring, despite differences in the composition of each topic.

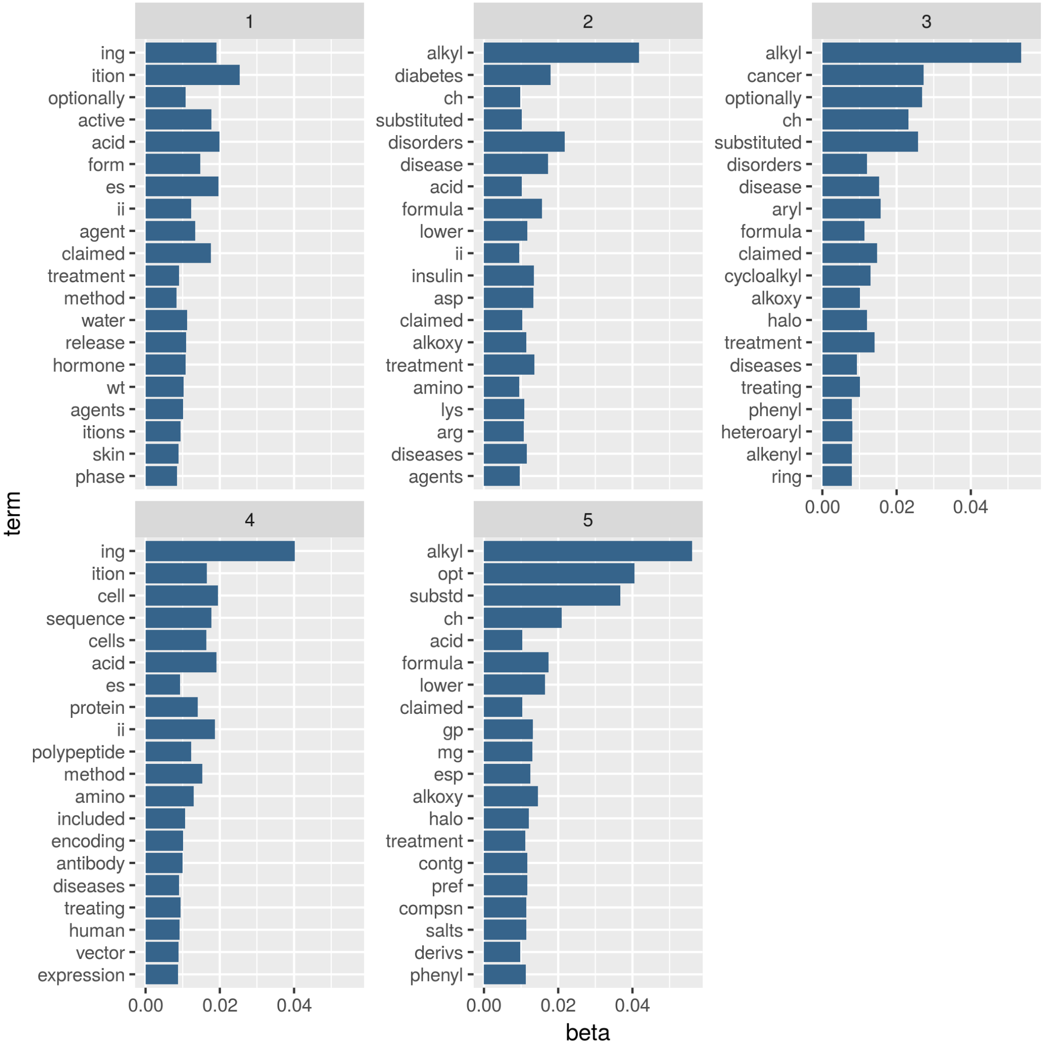


Figure 26: Results LDA analysis - China 1990-2000

The words for the diabetes topic include diabetes, disorder, disease, alkoxy, among others. In case of the dataset for 1990 to 2000, insulin is mentioned on top, emphasizing the topic's focus on the specific disease diabetes. With regards to the topic of cancer, both datasets describe the topic with the words: cancer, substituted, disease, treating, among others. While the first LDA analysis on the whole dataset hinted at a specific interest in lung cancer, the topic cancer modeled for 1990 to 2000 does not hint at any specific type of cancer. The topics on antibodies/immunology are associated with the words: antibody, amino, polypeptide, protein, human, cells, etc. Both match very well, despite the topic for 1990 to 2017 featuring the word virus, which highlights the specific focus of treatment on viruses. The remaining two topics (1 and 5) of the dataset from 1990 to 2000 differ with the topics for the whole dataset. Topic 1 might thereby refer to hormone treatments, considering that the topic groups the words hormone, treatment, active, agents, method, release. In contrast, we cannot identify a definite theme for topic 5 that groups words such as formula, alkoxy, treatment, salts, and phenyl.

Topic modeling for Chinese patents (2001-2017)

Since the year 2000 on, the patenting in China has grown rapidly. Hence topics of the patents might differ from the ones of patents filed between 1990 and 2000. The majority of the whole dataset has been filed in this time frame. Therefore it is not surprising that the topics of the 2001 to 2017 time frame match the topics of the 1990 to 2017 time frame better. The following figure (Figure 27) illustrates the five topics modeled from the 2001 to 2017 dataset.

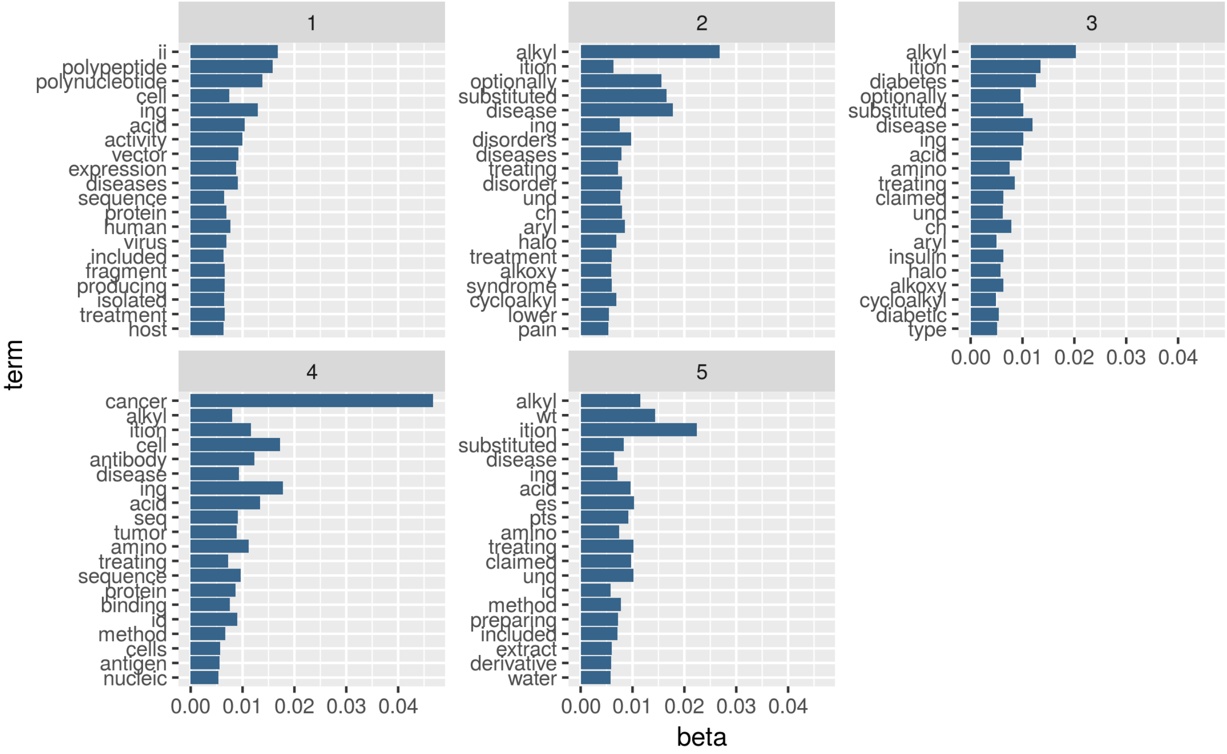


Figure 27: Results LDA analysis - China 2001-2017

Topic 1 centers around polynucleotide, polypeptide and therefore matches the same topic 5 of the 1990 to 2017 dataset. Both have several words in common. The 1990 to 2000 dataset, however, does not reveal this topic. Therefore we can assume that polynucleotides and thus a kind of homo- or heteropolymers have seen increasing importance across time. The second topic modeled is very similar to the whole dataset's topic 3 on diseases and disorders. The analysis reveals words, such as disease, disorder, pain, syndrome. While similar to the topic 3 on diseases/ diabetes in the whole dataset, the topic differs since the whole time frame includes diabetes in this topic, while the topic of the time frame 2001 to 2017 does not. Instead of diabetes, the term pain figures among the descriptors. This might indicate a concentration on pain medication. This topic on diseases and disorders does not fit any topic of the 1990 to 2000 time frame.

While not figuring in topic 2 on diseases and disorders in general, diabetes is a topic on its own, in the 2001 to 2017 dataset. Thus, this topic figures in all three LDA analyses. Diabetes has been a topic in Chinese patents since 1990 and continues to be an important topic. Both the diabetes topics in the datasets from

1990 to 2000 and 2001 to 2017 include not only the word diabetes but also insulin, indicating that the topic is fairly certain focused on diabetes. Apart from diabetes, cancer has been a topic disclosed by all three time frames. The words this topic is described by, are particularly close to the words of the same topic in the time frame 1990 to 2017. In general, this topic is associated with the terms cancer, tumor, protein, antigen, antibody. Due to the words cancer and tumor, it is reasonable to say that the topic is on cancer. Finally, the last topic is close to topic 2 (aminos) in the 1990 to 2017 time frame dataset, while not matching any in the 1990 to 2000 dataset. Common words of both topics are: amino, method, water, claimed, treating, included, extract, etc. Hence this topic focuses on aminos and their functioning, as has been reasoned in topic 2 of the 1990 to 2017 dataset.

Overall, the topics of the 2001 to 2017 time frame match the first LDA analysis better, with the recurring topics on cancer, diabetes, general diseases, aminos, and polynucleotides (specific homo- or heteropolymers). The topic antibodies/immunology of the 1990 to 2017 time frame dataset is not exposed in the 2001 to 2017 time frame dataset. The topics from the 1990 to 2000 time frame dataset only match the topics diabetes, cancer and antibody/immunology of the 1990 to 2017 topic modeling results. While it is reasonable to interpret that the topics of diabetes and cancer have been topics throughout the whole time span, some topics, such as polynucleotides and the amino's functioning seem to be fairly new. Antibodies and immunology might have been a separate topic from 1990 to 2000, but are not associated with a proper topic in current times.

Contextualising the LDA results

Having gained more insights into the topics dominant in the dataset of Chinese patents the results have to be compared to the previously presented data description. It is of interest to assess the match between the computational-derived topics as well the dominant topics derived from the content analysis, allowing us to confirm or question results. Before this comparison, it is important to mention whether the topics China's pharmaceutical industry focuses on in its pharmaceutical innovation according to the literature match the topics modeled from the data partially. The literature (e.g., Ni et al. 2017; Zhang and Zhou 2017; Shen 2010) hinted that liver disease, gastric cancer, cardiovascular disease, diabetes, and biotech are therapeutic areas or topics China's pharmaceutical industry focuses on. In general, cancer and diabetes can be found in the topics modeled from the data. Chervenak (2005)'s assessment of the biotechnology sector focusing on gene therapy, antibodies, and TCM modernization might in part be reflected since antibodies/immunology were one of the topics modeled in the 1990 to 2017 data frame. Thus, hypothesis 1.1 is confirmed. Some biotech topics, and the topics of cancer as well as diabetes have emerged as topics in the time frame 1990 to 2017. The latter two also in the time frame from 2001 to 2017. Diseases have been mentioned, but liver or cardiovascular diseases do not constitute a separate topic. Despite not constituting a separate topic in the dataset, we consider hypothesis 1.1 confirmed since generally diseases constitute a topic. Hypothesis 1.2 is confirmed as well since the five topics for the time span 1990 to 2000 are not the same as the five topics of the whole time span or the one from 2001 to 2017. Nonetheless, two of the topics from earlier innovations are congruent to the topics of later innovations (2001-2017). Thus there are differences, but these are not large. The topics of cancer and diabetes are present in both time frames from 1990 to 2000 and 2001 to 2017.

With regards to the comparison between the data description and the topics modeled, the results will be compared to the thematic fields derived from the Derwent classes and the description of manual code sections dominant in the dataset, in a first step. Then, the five topics identified through the LDA analysis will be matched with the therapeutic groups that were applied as framing on the dataset (Consult Figure 22 for a list of the therapeutic groups classified by patent count).

In general, we can find the topic of polypeptides/homo- or heteropolymer reflected in the Derwent classes B04 and C03 on polymers. B04 *Natural products and polymers, testing, compounds of unknown structure* was the Derwent class most patents were categorized into. With regards to the other topics a comparison with the universal Derwent classes is of difficulty, hence the comparison with the manual code sections is preferable. These are more detailed and therefore allow to match the codes with the topics better.

To recall, according to the manual code sections, the Chinese patenting concentrated roughly on pharmaceutical activities as well as natural products (or genetically engineered), polymers [13]. In general, we can assume that the five topics modeled can be classified under these two categories. The topics modeled on cancer, immunology, polynucleotide and generally diseases can be found in the most used manual code sections more explicitly. The topic on diseases and diabetes among others relates, for instance, to the manual code section B14-F *Pharmaceutical activities - Drugs acting on the blood and cardiovascular system*, since diabetes affects the blood system. Cancer relates to B14-H *Pharmaceutical activities - Cancer related drugs* and immunology to B14-G *Pharmaceutical activities - Drugs acting on the immune system.* Finally, as a homo- or heteropolymer, the topic on polynucleotide corresponds to the categories B04-C *Natural products (or genetically engineered), polymers - Polymers*, as well as B04-E *Natural products (or genetically engineered), polymers - Nucleic acids*, since polynucleotides are part of a nucleic acid molecule (Cammack 2006b). In conclusion, while not claiming this comparison to be exhaustive, it shows that the five topics modeled can be attributed to manual code sections that figure among the top 20 manual code sections. Hence, the computational topic modeling reflects the experts' classifications – even if less detailed since we have five topics – and can be considered successful.

When comparing the five topics identified through the LDA analysis (1990 to 2017) to the therapeutic groups, it is revealed that the top three therapeutic groups, *Anticancer*, *Alimentary Metabolic* and *Immunological*, according to the content analysis are part of the five topics identified through the LDA, also. Thus, we assume that there is a correspondence between the computational topic modeling and the applied framed topics. For a more precise comparison of the topics, it has been analyzed how the five topics overlap with the therapeutic groups. The following figure (Figure 28) visualizes the correspondence in terms of the number of publications where it was the highest percentage of topics. Thus, it allows to assess whether a topic is dominantly assigned to one therapeutic group or several groups. In the same time, it allows to assess the compilation of the therapeutic groups. To our knowledge, this is an entirely new approach to characterize innovation dynamics in an industry. It allows to disclose linkages between therapeutic groups, since the words contained in each five topics modeled are compared to all therapeutic groups. It has to be mentioned that the explanatory power of the analysis is limited since duplicate patents belonging to two therapeutic groups were not taken into account. This is a limitation arising from the underlying algorithmic calculations, which could be expected due to the novelty of the approach. Further research is needed to improve the

[13] Consult Table 7 in the appendix for a description and list of the top 20 manual code section.

algorithms developed for this part of the analysis. Nonetheless, the results are of interest, despite having to treat them with caution.

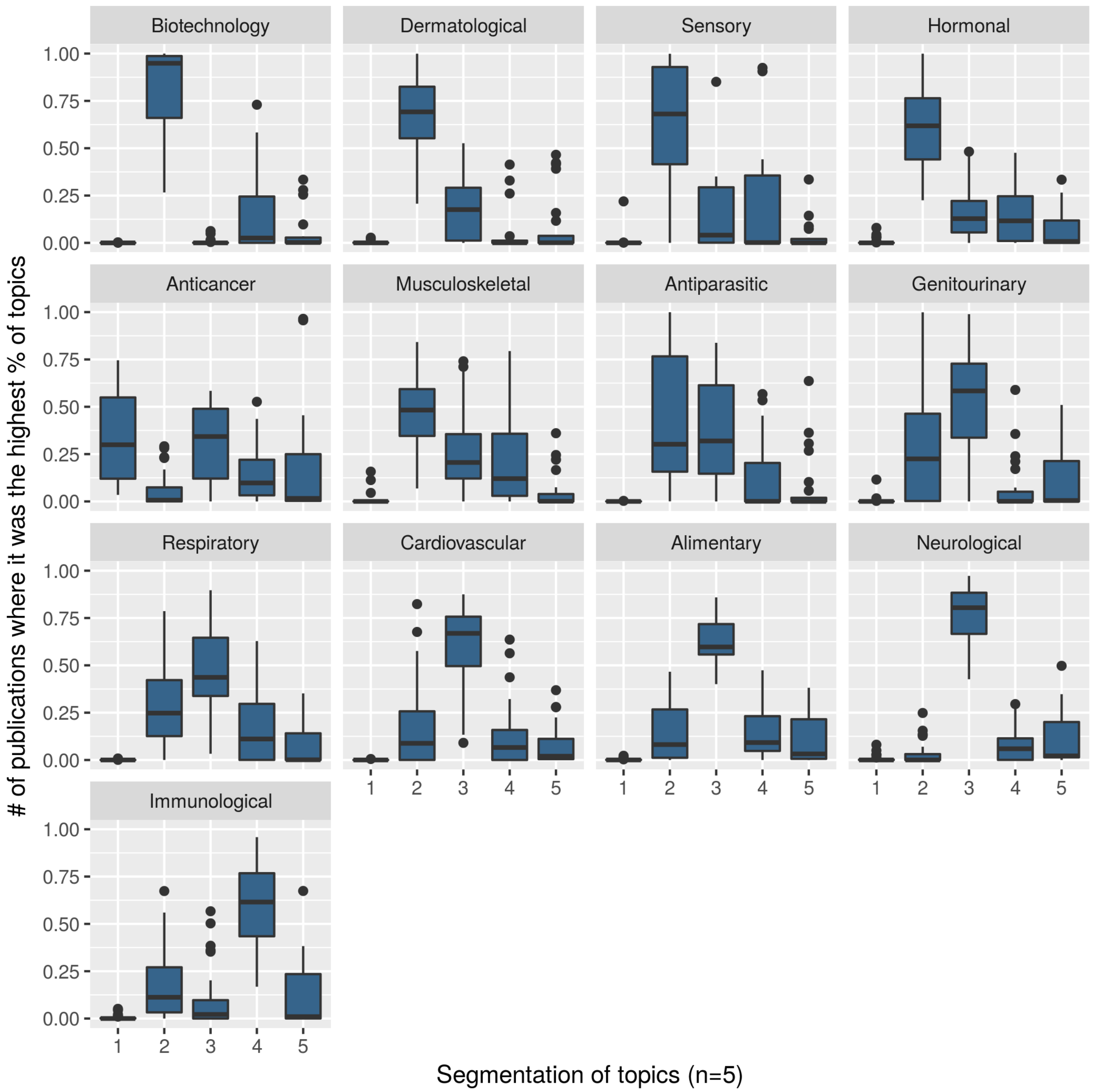


Figure 28: Comparison LDA results to therapeutic groups

The *biotechnology* group is being described by the second topic that was summarized as being on aminos and their functioning. The box plot indicates that from the lower quartile to the upper quartile for 50% of the patents the topic 2 was between around 70% to 98% the most important topic. For another 25%, the topic has an importance of between 98% and 100%. The median is at around 95%, which indicates that topic 2 (amino) is vital for patents in the biotechnology group. Remarkably, the lower whisker indicates that 25%

of the patents are however distributed between around 25% and 70%. While topic 1 (cancer), 3 (diseases) and 5 (polynucleotide) can be neglected due to their poor importance, topic 4 (antibodies/immunology), is slightly important as well. Topic 2 (amino) equally figures as the most important topic for the *dermatological*, *sensory* and *hormonal* group, where the median lies above the 50% threshold. The box areas for topic 2 of the *dermatological* group exposes that for 50% of the patents the topic matches them between around 57% and 80%. The upper whisker, however, indicates that for 25% of the patents the topic matches between 80% and 100% of the content. The lower whisker, in contrast, exposes that for the remaining 25% of patents the topic 2 (amino) only figures in between 25% to 57% of the abstract. With an equal median of 70%, 50% of the patents of the *sensory* group accord, however, a more significant range of importance to aminos. The box area spans from the importance of 40% to 90%. The upper whisker indicates that the topic aminos has very high importance for 25% of the patents. Nonetheless, the topic does not match all the patents since the bottom whisker indicates that for 25% the topic matches from 0% to 40%. As mentioned, topic 2 (amino) is equally important for the *hormonal group*. The median lies at around 62% and 50% of the patents match the topic amino in a range from 42% to 75%. For 25% of the patents, the topic has an importance of between 75% and 100%. While these three therapeutic groups and the *biotechnology* group have a median match of above 50%, other therapeutic groups' patents match the topic 2 (amino) as well. For instance *antiparasitic* and *genitourinary* where the upper whisker ends at a 100% match. Albeit a low match, it can indicate some form of weak link between therapeutic groups as well. In conclusion, aminos (topic 2) cannot be solely attributed to one therapeutic group, this is not surprising since aminos can be considered a component applicable to several functions. These connections between therapeutic groups might indicate common fields in research.

Similar to the therapeutic group *biotechnology*, the therapeutic group *neurological* matches mostly one topic, topic number 3, on diseases and diabetes. It is understandable that a topic such as neurological matches patents that are about diseases, disorders or syndromes. Its median lies at around 80%. Hence, 50% of the patents match the topic diseases between 80% to 95%. Topic 3 (diseases) is essential in the *genitourinary*, *respiratory*, *cardiovascular* and *alimentary metabolic* therapeutic groups, also. In particular, the match between topic 3 on diseases and the *alimentary metabolic* group is without surprise considering that both cover diabetes, among others. Finally, topic 3, on diseases and disorders matches a lot of different therapeutic groups to some degree, which is understandable due to its generic nature.

Topic 4, on antibodies and immunology, despite being present as a topic in some therapeutic classes, is the most important topic for the patents of the *immunological* therapeutic group. The median of the box plot for topic 4 lies around 62%, with 50% of the patents exhibiting a match between abstracts and topic of between 62% and 95%. The upper whisker indicates that 25% of the patents match with the topic (antibodies/immunology) from 75% to 95%. The bottom whisker, however, reveals that 25% of the patents only match between 15% and 44%. The range is therefore vast with the minimum of 15% and the maximum of 95%.

To the researcher's surprise, topic 5, on polynucleotides and roughly polymers matches the therapeutic groups little. The median figures for each therapeutic group around 0%. While the topic does not figure among the therapeutic groups of *biotech*, *dermatological*, *sensory*, *musculoskeletal* and *antiparasitic*, except for some outliers, it is covered to some extent by the therapeutic groups *hormonal*, *anticancer*, *genitourinary*, *respiratory*, *cardiovascular*, *alimentary metabolic*, *neurological* and *immunological*. This finding might indicate that the classification in therapeutic groups is not the aptest, considering that polymers are one of the most represented categories in pharmaceutical innovation in China (Figure 20).

Finally, topic 1 on cancer, only matches the therapeutic group *anticancer*, except for some outliers. While generally not surprising, the match between patents in the *anticancer* therapeutic group and the topic cancer modeled through computational methods only matches to some degree. With a median of around 30% and 25% of the patents with a match between 6% and 12%, the topic of cancer does not match the therapeutic group much. The upper whisker indicates that 25% of the patents matched the topic by 55% to 75%. Apart from topic 1, the *anticancer* therapeutic group can be classified into topic 3, diseases, to some extent, also. Topic 3's median is around 35%, higher than the median of the topic 1 (cancer). The match between topic 3 (diseases) and the *anticancer* therapeutic group is reasonable since cancer is a disease. Hence the words of topic 3, which center around diseases and disorders, correspond with the therapeutic group on cancer treatments.

It has been shown that the topics, modeled through computational methods, classify the Chinese patent dataset of the 15 therapeutic groups to some extent in therapeutic groups, but more so classify them according to groups of words, which explains why aminos or polynucleotides get assigned topics. An explicit match between a therapeutic group and a topic that was modeled has occurred seldom. Thus, therapeutic groups cover several topics. In consequence, we assume that there are synergies between different therapeutic groups and topics, as has been indicated earlier by words the modeled topics have in common. Moreover, this comparison of therapeutic groups and topics modeled has hinted that indeed cardiovascular diseases play a role in Chinese patents. This might be the case, since topic number 3, on diseases and diabetes, did reasonably well match patents of the cardiovascular therapeutic group. Nonetheless, this would have to be further analyzed before making a valid statement.

In conclusion, not only have the LDA analyses revealed that China has focused on some therapeutic regions, such as cancer and diabetes since 1990, but it has also allowed to identify polynucleotides, a polymer, as a new area of research. Overall, the LDA analysis of the data corresponds to the data description through the human-made classification to some extent as well as to the little indication in the literature. Finally, both hypothesis could be confirmed. In particular, it has been shown that there is a difference between topics modeled from older patents (1990 to 2000) and newer patents (2001 to 2017).

6.2 Step 2: Synergies across patent thematics of Chinese pharmaceutical patents, a Jaccard similarity perspective

The following chapter provides descriptive statistics with the purpose of giving insights into the similarity between manual code sections across time and classes. While the LDA analysis and the data description have described the general topics of the Chinese dataset, the following analysis will enable us to picture the dynamics better. The similarity measurement identifies connections between topics through time. Furthermore, we will be able to assess the relation between the time lag of sections and their similarity. Thus, it will enable us to answer the question (2): *What kind of dynamics do we find between patent thematics/classes? More precisely: What is the relationship of patent classes across time?*

First, the following figure (Figure 29) illustrates the measurement of the Jaccard similarity. On both axes, the 537 files of the dataset are represented. Each file thereby contains several patents. The figure depicts the vast amount of data that Data Science techniques allowed to analyze. Each color of the heat map illustrates the grade of similarity between two files. The darker the color, the higher the similarity between two files.

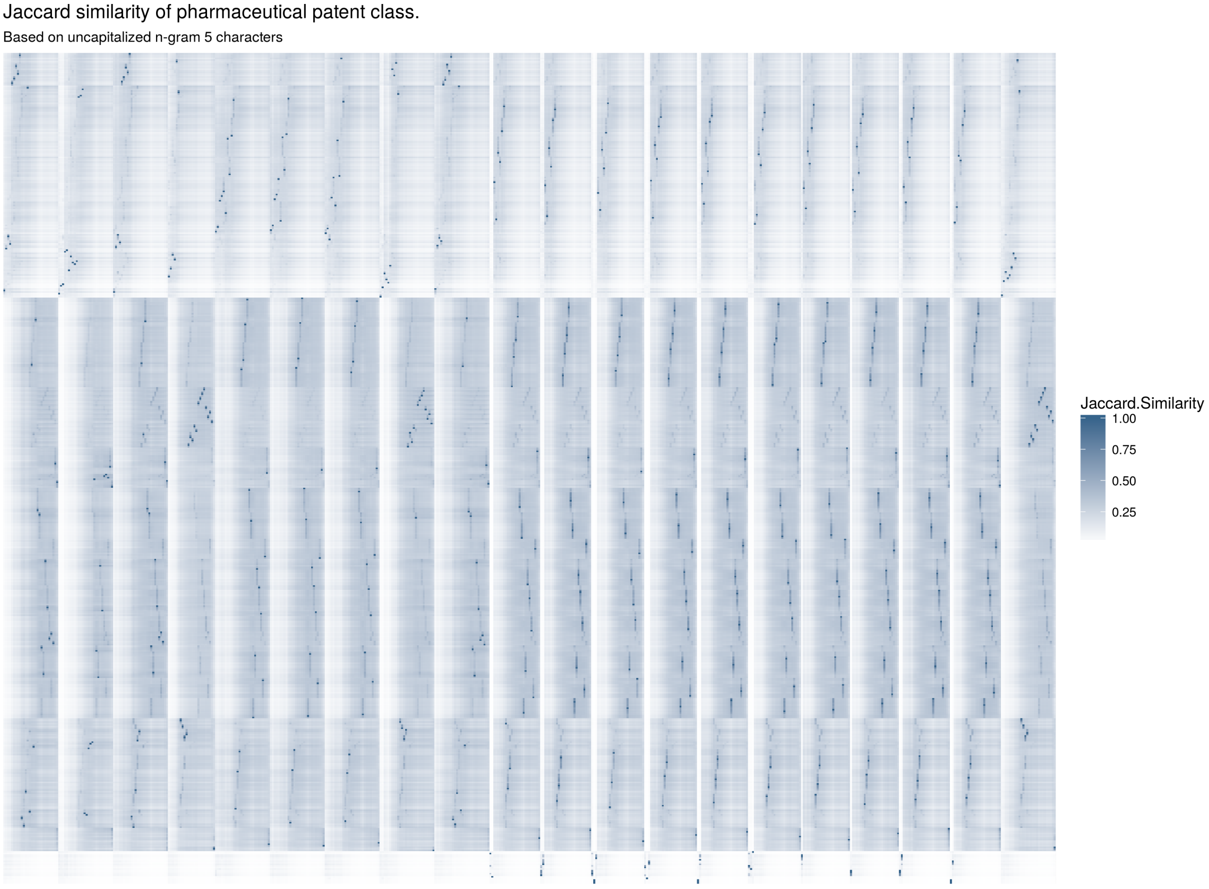
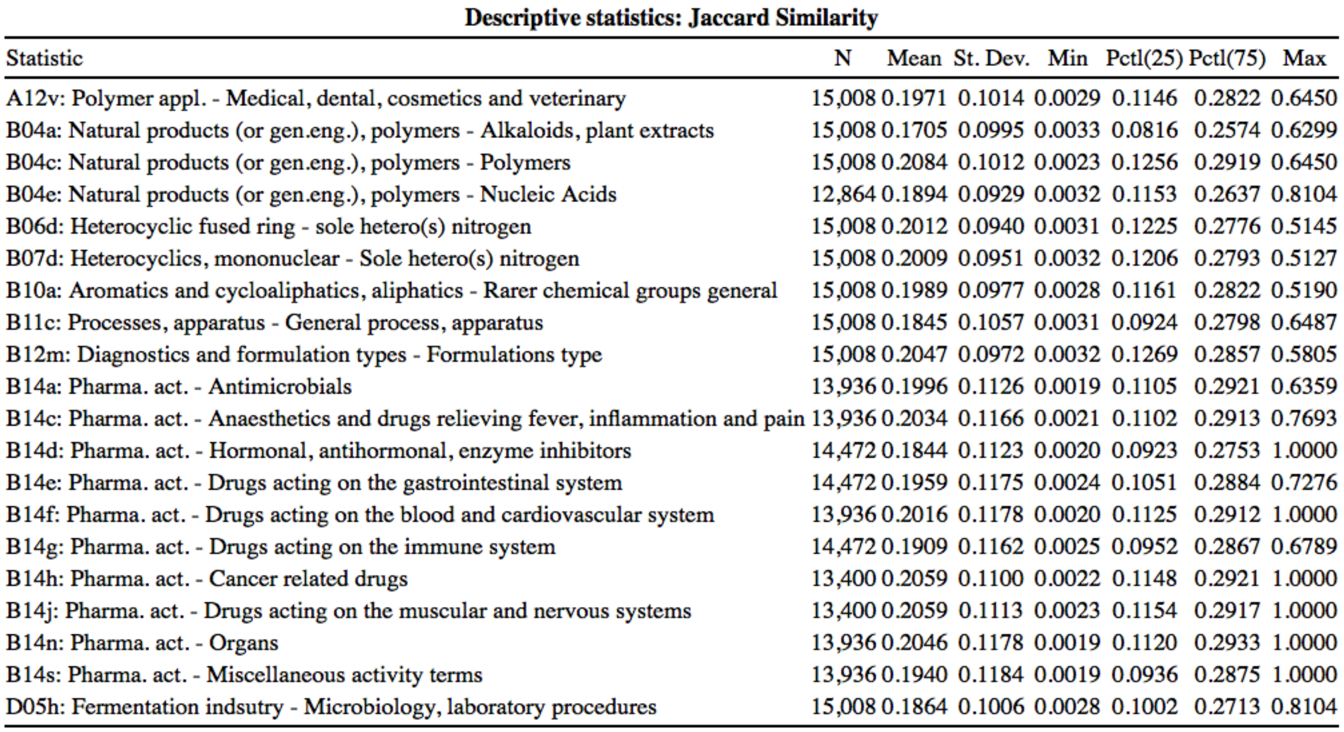


Figure 29: Heatmap of Jaccard similarity results

While giving an overview of all the measurements, the number of files is too large for the graphic to enable a comprehensive interpretation of the classes and years that have a high similarity. Thus, the descriptive statistics might provide a better overview (Figure 30).

It is noticeable that – no matter the patent class – both the means of the Jaccard similarity and the standard deviations are close to each other. While the mean Jaccard similarity per class is close to 0.19, the standard deviations are around 0.1. The average textual similarity between patent thematics seems to be high. A match of 19% in a patent abstract represents a match of almost one fifth. In view of patents having to encompass a novelty, we consider a match of 19% high. The minimum values are equally close to each other, with around 0.002. Thus, each class has years where the similarity is considerably low. While overall, the minimum values are close to each other, the maximum values are more dispersed, ranging from 0.5127 to 1.00. Considering, that for each manual code section each year is compared to all sections and years, the similarities of 100% might be due to a comparison of different manual code classes in the same year, containing the same patents.



Descriptive statistics: Jaccard Similarity

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
A12v: Polymer appl. - Medical, dental, cosmetics and veterinary	15,008	0.1971	0.1014	0.0029	0.1146	0.2822	0.6450
B04a: Natural products (or gen.eng.), polymers - Alkaloids, plant extracts	15,008	0.1705	0.0995	0.0033	0.0816	0.2574	0.6299
B04c: Natural products (or gen.eng.), polymers - Polymers	15,008	0.2084	0.1012	0.0023	0.1256	0.2919	0.6450
B04e: Natural products (or gen.eng.), polymers - Nucleic Acids	12,864	0.1894	0.0929	0.0032	0.1153	0.2637	0.8104
B06d: Heterocyclic fused ring - sole hetero(s) nitrogen	15,008	0.2012	0.0940	0.0031	0.1225	0.2776	0.5145
B07d: Heterocyclics, mononuclear - Sole hetero(s) nitrogen	15,008	0.2009	0.0951	0.0032	0.1206	0.2793	0.5127
B10a: Aromatics and cycloaliphatics, aliphatics - Rarer chemical groups general	15,008	0.1989	0.0977	0.0028	0.1161	0.2822	0.5190
B11c: Processes, apparatus - General process, apparatus	15,008	0.1845	0.1057	0.0031	0.0924	0.2798	0.6487
B12m: Diagnostics and formulation types - Formulations type	15,008	0.2047	0.0972	0.0032	0.1269	0.2857	0.5805
B14a: Pharma. act. - Antimicrobials	13,936	0.1996	0.1126	0.0019	0.1105	0.2921	0.6359
B14c: Pharma. act. - Anaesthetics and drugs relieving fever, inflammation and pain	13,936	0.2034	0.1166	0.0021	0.1102	0.2913	0.7693
B14d: Pharma. act. - Hormonal, antihormonal, enzyme inhibitors	14,472	0.1844	0.1123	0.0020	0.0923	0.2753	1.0000
B14e: Pharma. act. - Drugs acting on the gastrointestinal system	14,472	0.1959	0.1175	0.0024	0.1051	0.2884	0.7276
B14f: Pharma. act. - Drugs acting on the blood and cardiovascular system	13,936	0.2016	0.1178	0.0020	0.1125	0.2912	1.0000
B14g: Pharma. act. - Drugs acting on the immune system	14,472	0.1909	0.1162	0.0025	0.0952	0.2867	0.6789
B14h: Pharma. act. - Cancer related drugs	13,400	0.2059	0.1100	0.0022	0.1148	0.2921	1.0000
B14j: Pharma. act. - Drugs acting on the muscular and nervous systems	13,400	0.2059	0.1113	0.0023	0.1154	0.2917	1.0000
B14n: Pharma. act. - Organs	13,936	0.2046	0.1178	0.0019	0.1120	0.2933	1.0000
B14s: Pharma. act. - Miscellaneous activity terms	13,936	0.1940	0.1184	0.0019	0.0936	0.2875	1.0000
D05h: Fermentation indsutry - Microbiology, laboratory procedures	15,008	0.1864	0.1006	0.0028	0.1002	0.2713	0.8104

Figure 30: Descriptive statistics for the Jaccard similarity

Hence, we will contrast the distribution of the similarity for measurements across or within classes. A graphical representation of box plots (Appendix: Figure 42) indicates that indeed the high Jaccard similarities might be explained by comparing two different manual code classes to each other. Apart from the upper whisker and the outliers of a Jaccard similarity of one, the distributions are similar. Both, the median and mean of the box plots are similar. It supports the hypothesis that the same patents might be included in two classes in the same year and hence explain the high Jaccard similarity. A time component might play a role with regards to the similarity across patent thematics.

Therefore, since it is of interest to know whether a time component can explain the high similarities, the time lag was plotted against the Jaccard similarities. It is revealed that indeed classes compared in the same year explain the Jaccard similarities of one (Figure 31). This is also the time lag with the highest dispersion of Jaccard similarities. The appendix (Appendix: Table 8 and 9) provides a detailed description of these classes with a Jaccard similarity of one and of the classes with the highest similarities after one across time.

More importantly, it is shown that with increasing time lag, the mean and median Jaccard similarity decreases (Figure 31). Hence, the data indicates that patents are the less similar to each other with increasing time. Thus, subsequent innovation relies on previous innovation, but the less ans less with passing time. While the mean similarity lies at approximately 37% when there is a time lag of 0, it drops under 25% at a time lag of 5 years. After roughly 23 years, the subsequent innovation does not rely on previous innovation anymore. This finding is indicated by the slope of the mean Jaccard similarity that remains steady. At this point, the mean similarity for this time span reaches around 3%. As the literature has indicated, there is a trend towards an increased pace of innovating and incremental innovation in the pharmaceutical industry (Albrecht et al. 2016). Overall, we find that the similarity between patent thematics seems to follow a certain pattern across time. So far, the assessment of the similarity across patent thematics and across time has already allowed us to make some conclusions no researchers had been able to quantify.

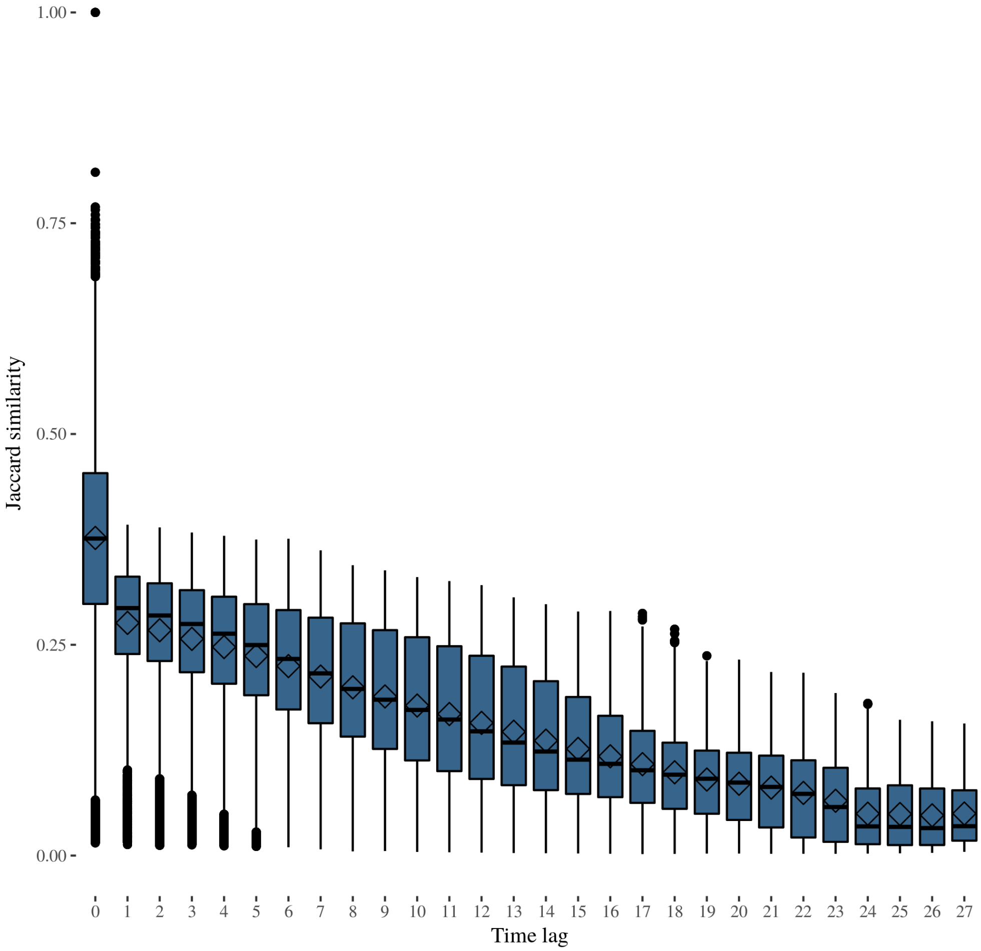


Figure 31: Distribution of Jaccard similarity of all classes with time lag

Having established that an increasing time lag has a negative relationship with the Jaccard similarity, we introduce the patent count as a variable. The following graphics will allow us to assess whether there is an indication that the new variable we propose, the similarity, has an explanatory power with regard to the patent count. The graphs will look at each manual code class separately.

Figure 32 confirms that overall there is a negative relationship between the time lag and the Jaccard similarity. This negative relationship holds true for each manual code class. Excluding a time lag of 0 and above 23 years, the Jaccard similarities across manual code classes are close. With regards to the patent count the graphic reveals that with an increasing time lag and a decreasing Jaccard similarity, the sum of the patent count for each class decreases. Synergies across classes and time might explain the increasing patent count with an increasing Jaccard similarity. Leveraging the knowledge from other domains as well as the specific knowledge of the domain of the researchers might create these synergies that are reflected in a high patent count. With increasing time lag, the researchers might not leverage knowledge from other classes anymore, and thus no synergies are being created that might account for the higher patent count. This hypothesis derived from the data is in line with the literature on organizational economics and R&D organization, that promotes an integrated approach, where different R&D departments working on different things should work together, adopting an integrated network model inside firms and across borders (Gassmann and von Zedtwitz 1999). In the same vein, open innovation models highlight the importance of integrating different actors in the innovation process, thus leveraging different knowledge and knowledge flows (Chesbrough 2006). On the same lines, Mudambi (2002) classifies these knowledge flows as collaborative knowledge flows between subsidiaries. While the existing data does not allow to take into account whether an open innovation model or an integrated R&D organization approach has been chosen, it allows nonetheless to assume that creating synergies between classes, reflected by a higher Jaccard similarity, is beneficial for the innovation output in terms of the patent count.

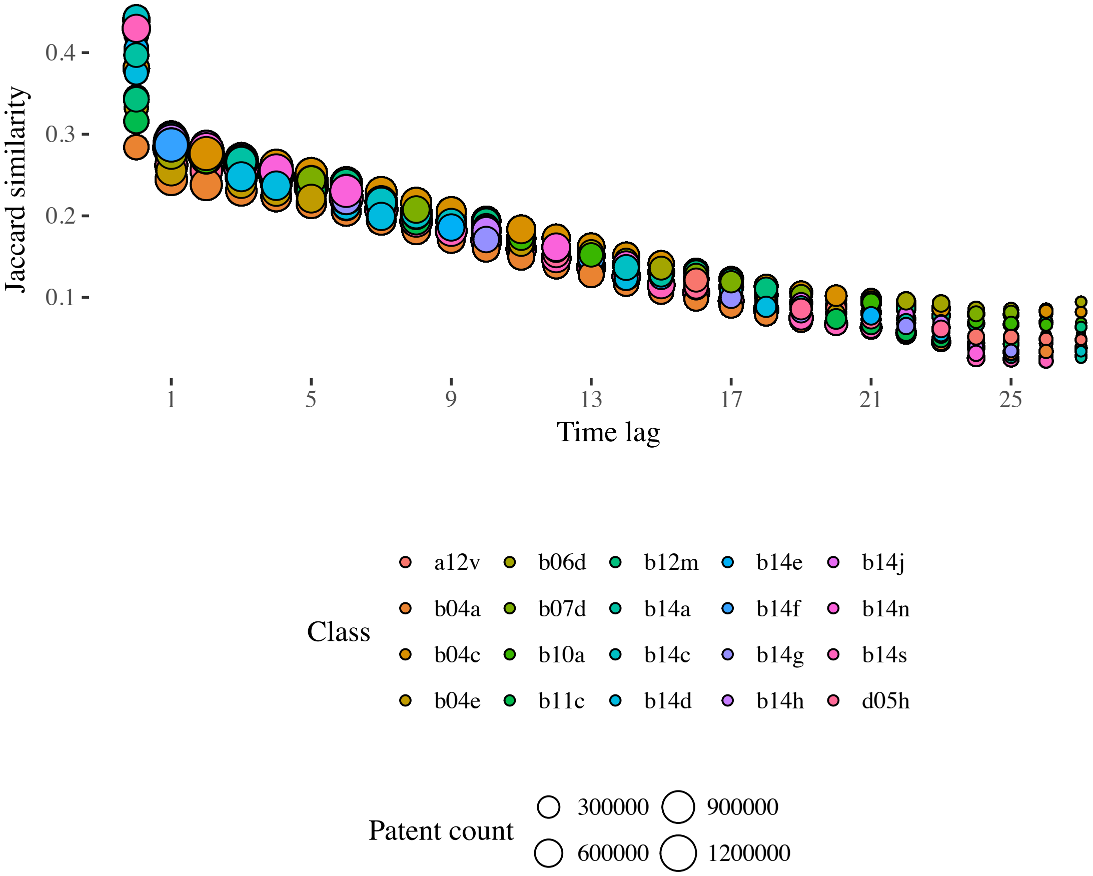


Figure 32: Average Jaccard similarity per time lag per class

In addition to looking at the relationship between the time lag and the Jaccard similarity, the relationship between the year of publication and the Jaccard similarity will be assessed as well. Figure 33 reveals that the average Jaccard similarity per class per year has been increasing from 1990 to 2000, and remained steady from 2000 to 2017. Moreover, the dispersion of the average Jaccard similarity per year per class has decreased considerably since 1990. Meaning, that across all classes in a given year the dynamics are similar. Moreover, the patent count has overall grown over time. This growth is not surprising since Figure 13 on the number of publications in China has shown an increase in patent count over the years. Finally, the classes B14-S and B14-N can be identified as the classes with the highest amount of patent count since 2008. This is in concordance with the findings from Figure 41, that identified B14-N and B14-S as the two manual code sections with the highest patent count from 1990 to 2017.

In conclusion, the developed graphical representations of the Jaccard similarity measurements have allowed to describe the relationship between patent classes across time. Our hypothesis (H2) could be confirmed. The similarity measurement enabled us to characterize dynamics across classes and indicated that time plays a role when capturing similarities across patent thematics. Indeed, with an increasing time lag, the average Jaccard similarities of each class decreased, indicating that innovations rely less and less on previous innovations. Furthermore, the analysis suggested that in recent years, the dispersion of the average Jaccard similarities is getting smaller across different patent sections, meaning that all classes collaborate to a similar extent with other classes. This collaboration might reflect the trend towards integrated R&D models and integrated models of innovation in general, where firms try to integrate their departments (Gassmann and von Zedtwitz 1999; Rothwell 1993). Finally, the graphs indicated that similarity might play a role with regards to the patent count, which is why we will look at the relationship from an econometric perspective.

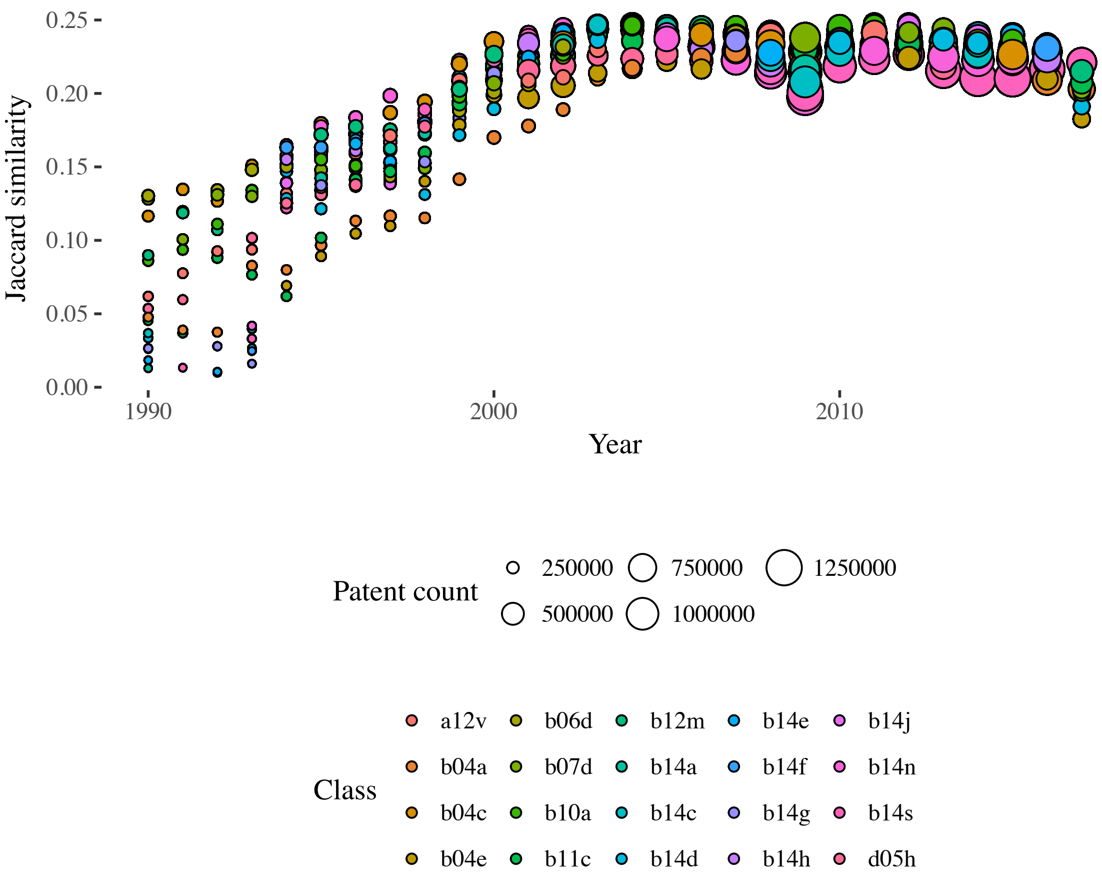


Figure 33: Average Jaccard similarity per year per class

6.3 Step 3: Impacts of patent thematic synergies on the patent count, an econometric perspective

In the following, we will present the results of the econometric analysis that aims to explore the impact a patent thematic category's Jaccard similarity has on the patent count in subsequent years. More precisely we answer the question (3): *What is the relationship between the similarity of patent classes and the patent count when taking into account the time lag?* Due to several reasons, as explained in chapter 5, a panel linear model with fixed effect estimation was chosen fixing the time and individual effect, and leveraging the plm package of Croissant (2017). In contrast, to chapter 6.2 (the analysis of the Jaccard similarity) the following analysis focuses on the Jaccard similarity across classes in a given year, trying to explain how synergies between different patent thematics in a year might explain the patent count in subsequent years. Thus, it looks at knowledge sourcing and how its diversification might impact the patenting behavior of firms. It reveals a new explanatory variable when looking at patenting behavior.

The dependent variable

Since this particular subset and the dependent variable have received less attention in chapter 6.2, we will briefly touch upon it[14]. Figure 34 plots the different patent counts per class. The graphic illustrates that there is heterogeneity across entities. The line illustrates the mean patent count per manual code section. While close to each other, the data cannot be considered homogeneous.

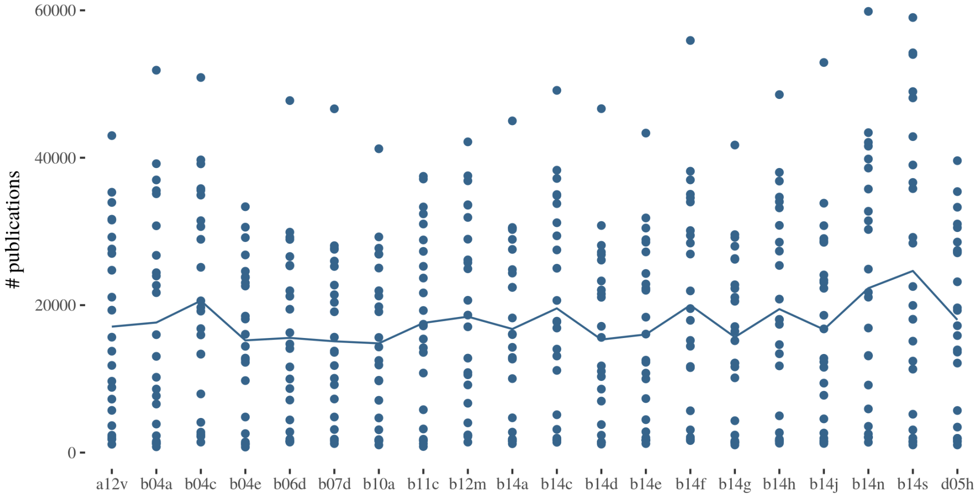


Figure 34: Heterogeneity of patent count across manual code sections

Not only the patent count across classes but also across years is heterogeneous (Figure 35). While following a similar pattern, the data cannot be considered homogeneous. The graphic illustrates that the patent count

[14] The measurement of the independent variable has been elaborated in chapter 6.2.

has been increasing rapidly until 2009, where it experienced a decline for the first time. While recovering in 2012, and rising again until 2015, it has been decreasing since.

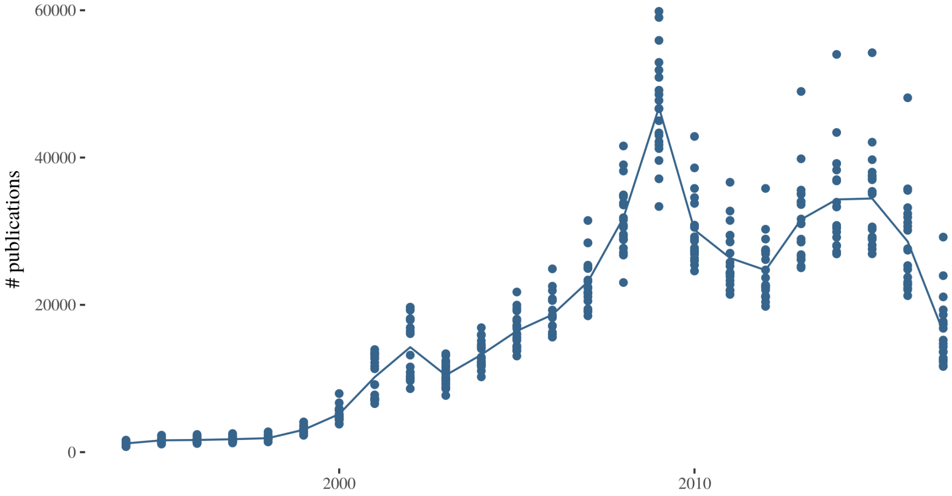


Figure 35: Heterogeneity of patent count across years

The independent variables

As explained in chapter 5, the aim of this analysis is the exploration of whether a lagged Jaccard similarity has an impact on the patent count. Does a high Jaccard similarity across patent thematics, measured by two manual code sections, explain the patenting activity in subsequent years? The character of this analysis is purely experimental and hence does not take into account other underlying variables. Due to the model's design, the omitted variable bias is, however, limited (Wooldridge 2000). To recapitulate, the independent variables are the lagged Jaccard similarities from 1 to 9 years (Model 1), 1 to 10 years (Model 2), 1 to 11 years (Model 3), 1 to 12 years (Model 4) and 1 to 13 years (Model 5). Five models have been fitted since it takes 9 to 13 years from the idea to the patenting of a new drug (Ding, Eliashberg, and Stremersch 2014: p. 29).

The econometric models

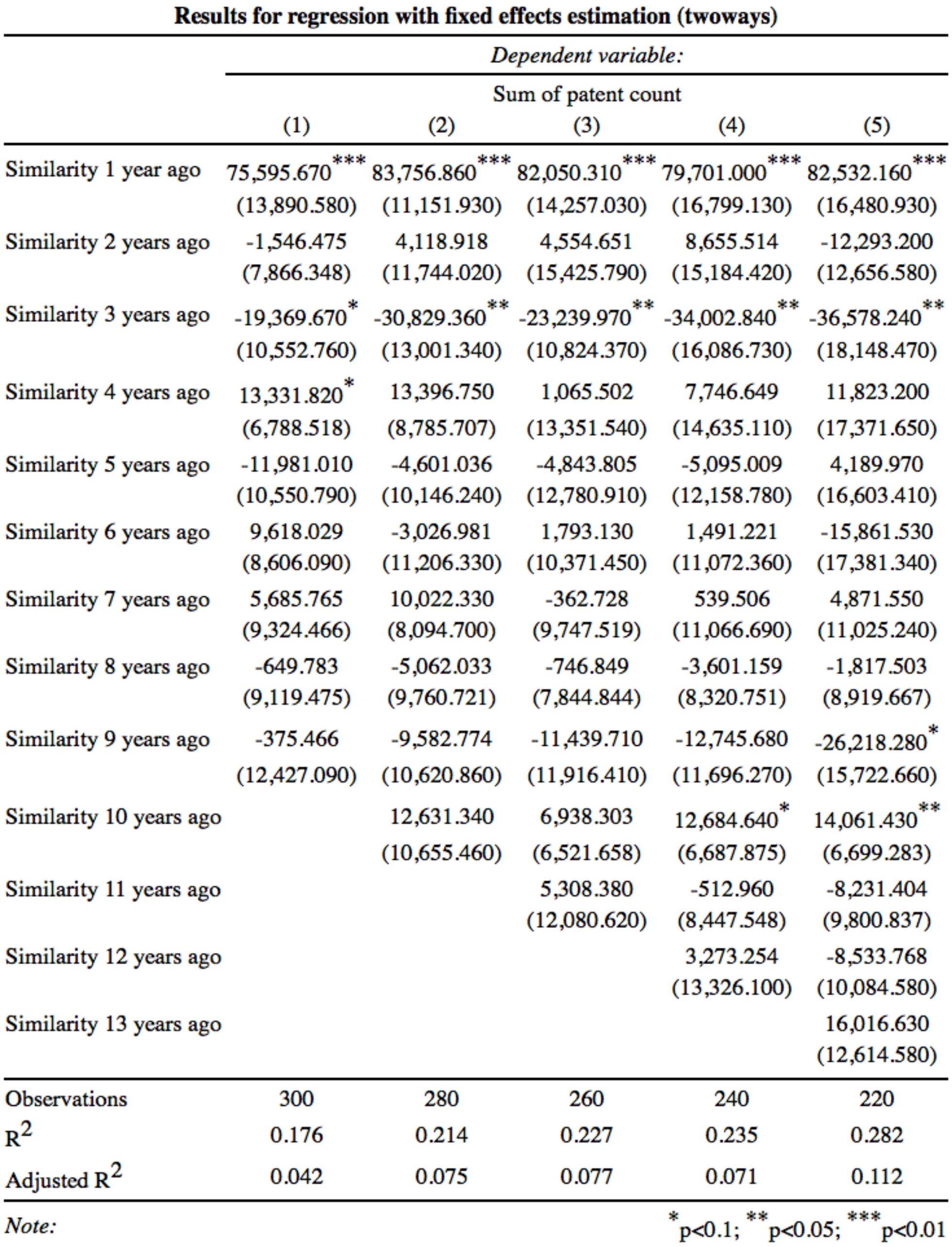
In the following, the results of the five econometric tests, differing in the lagged Jaccard similarities will be presented in a global and exhaustive manner. As explained in the methodology section, the panel linear models with twoway fixed effect estimation were applied on a sample of the Chinese patent data. The data was limited to measurements of similarities between different patent sections of the same year, after 1993. This limitation yielded a balanced panel, enabling us to measure the impact of synergies across sections of the same year on the patent count in subsequent years. Since the year itself is not of interest, but the time lag after which a Jaccard similarity impacts the patent count, the five econometric models contained the lagged Jaccard similarity, lagged by 1-year steps.

Overall, all five models have an R squared that is fairly low, ranging from 17.6% to 28.2% (Figure 36). It increases with the introduction of additional lagged similarities. Despite a generally low R squared, this is not of concern, since the analysis is experimental and does not intend to give a holistic picture. When only taking into account one variable, the similarity across sections, it cannot be expected to capture all dynamics around the patent count. Nonetheless, it is sufficient for this experimental setting.

When comparing the five models, model (1) with lagged similarities from 1 year to 9 years, yields three significant variables: the lagged Jaccard similarity 1 year ago (significant at the 1% significance level), the lagged similarity 3 years ago (significant at 10% significance level) and the lagged similarity 4 years ago (significant at the 10% significance level). Due to the positive coefficient of the first significant variable, model (1) indicates that the similarity across sections from a year ago explains a higher patent count in the present year. The same variable is significant at the 1% level throughout all five models, with a positive coefficient. Model (1) equally indicates that the similarity of 3 years ago explains a lower patent count in the present year, since the coefficient is negative. This does not mean that there are no patents filed, but merely that there are less filed. The same variable is significant throughout all five models, with a negative coefficient but at different significance levels. The similarity of 4 years ago, with a positive coefficient, is only significant in model (1).

Model (2), with lagged similarities from 1 year to 10 years ago, as well as model (3), with lagged similarities from 1 year to 11 years ago, both yield two significant variables: the lagged similarity 1 year ago (significant at the 1% significance level) and 3 years ago (significant at the 5% significance level). While the relationship with the patent count is positive in case of the similarity of a year ago, it is negative in case of the similarity of 3 years ago.

In contrast, to Model (2) and (3), Model (4) adds a third significant variable: the lagged similarity 10 years ago (at the 10% significance level). According to model (4) with lagged similarities from 1 year to 12 years ago, the similarity of 10 years ago explains a higher patent count in the present year. Finally, Model (5), encompassing similarities across manual code sections from 1 year ago to 13 years ago, reveals the same three statistically significant variables as model (4): the similarity a year ago (p-value<0.01), the similarity 3 years ago (p-value <0.05), and the similarity 10 years ago (p-value <0.05). In addition, the model reveals the similarity of 9 years ago as statistically significant at the 1% level (p-value <0.01). The coefficient is negative indicating that the similarity across sections of 9 years ago impacts the patent count negatively. This last model features the highest R squared, indicating that the independent variables explain 28.2% of the variability in the dependent variable.



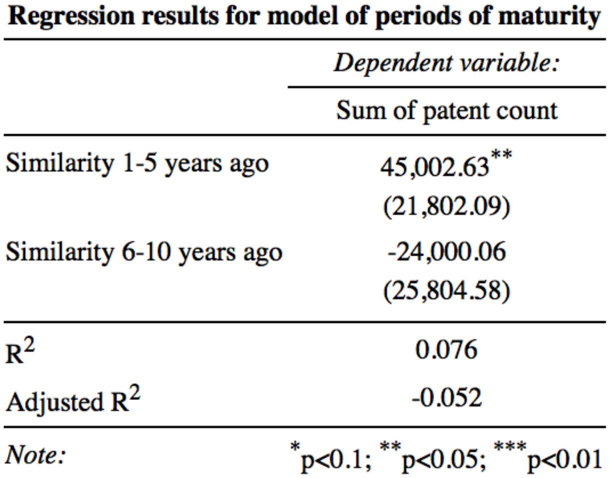
Results for regression with fixed effects estimation (twoways)

	Dependent variable:				
	Sum of patent count				
	(1)	(2)	(3)	(4)	(5)
Similarity 1 year ago	75,595.670***	83,756.860***	82,050.310***	79,701.000***	82,532.160***
	(13,890.580)	(11,151.930)	(14,257.030)	(16,799.130)	(16,480.930)
Similarity 2 years ago	-1,546.475	4,118.918	4,554.651	8,655.514	-12,293.200
	(7,866.348)	(11,744.020)	(15,425.790)	(15,184.420)	(12,656.580)
Similarity 3 years ago	-19,369.670*	-30,829.360**	-23,239.970**	-34,002.840**	-36,578.240**
	(10,552.760)	(13,001.340)	(10,824.370)	(16,086.730)	(18,148.470)
Similarity 4 years ago	13,331.820*	13,396.750	1,065.502	7,746.649	11,823.200
	(6,788.518)	(8,785.707)	(13,351.540)	(14,635.110)	(17,371.650)
Similarity 5 years ago	-11,981.010	-4,601.036	-4,843.805	-5,095.009	4,189.970
	(10,550.790)	(10,146.240)	(12,780.910)	(12,158.780)	(16,603.410)
Similarity 6 years ago	9,618.029	-3,026.981	1,793.130	1,491.221	-15,861.530
	(8,606.090)	(11,206.330)	(10,371.450)	(11,072.360)	(17,381.340)
Similarity 7 years ago	5,685.765	10,022.330	-362.728	539.506	4,871.550
	(9,324.466)	(8,094.700)	(9,747.519)	(11,066.690)	(11,025.240)
Similarity 8 years ago	-649.783	-5,062.033	-746.849	-3,601.159	-1,817.503
	(9,119.475)	(9,760.721)	(7,844.844)	(8,320.751)	(8,919.667)
Similarity 9 years ago	-375.466	-9,582.774	-11,439.710	-12,745.680	-26,218.280*
	(12,427.090)	(10,620.860)	(11,916.410)	(11,696.270)	(15,722.660)
Similarity 10 years ago		12,631.340	6,938.303	12,684.640*	14,061.430**
		(10,655.460)	(6,521.658)	(6,687.875)	(6,699.283)
Similarity 11 years ago			5,308.380	-512.960	-8,231.404
			(12,080.620)	(8,447.548)	(9,800.837)
Similarity 12 years ago				3,273.254	-8,533.768
				(13,326.100)	(10,084.580)
Similarity 13 years ago					16,016.630
					(12,614.580)
Observations	300	280	260	240	220
R^2	0.176	0.214	0.227	0.235	0.282
Adjusted R^2	0.042	0.075	0.077	0.071	0.112
Note:				*p<0.1; **p<0.05; ***p<0.01	

Figure 36: Results for regression with fixed effects estimation

While the five models in the previous figure revealed the impact of similarities across patent thematics on patenting in subsequent years, adopting a single year perspective, another model was built regrouping similarities across years. This way a *near past* variable and a *distant past* variable was created, with the average similarities across the respective years. The near past independent variable thereby averaged the similarities from 1 year ago to 5 years ago, while the distant past independent variable averaged the similarities from 6 years to 10 years ago. Since the development time of pharmaceutical drug innovation from idea to patenting takes around 10 to 13 years, and we learned from Figure 36 that the lagged similarity of 10 years ago was significant, a 5 year threshold seemed adequate for the distinction between near past and distant past (Ding, Eliashberg, and Stremersch 2014). Essentially these two variables represent periods of maturity, which is why we called the model: model of periods of maturity. This model allows to check the robustness of the results attained from the previous models.

The newly constructed model discloses that the near past, represented by the lagged average similarity of 1 to 5 years ago, has a positive impact on the patenting in subsequent years (Figure 37). The variable is significant at the 5% significance level. While the near past is positively associated with an increase of patents at a given time, our model shows no indication that the distant past has an impact. The variable of a lagged average similarity from 6 to 10 years ago is not significant.



Regression results for model of periods of maturity

	Dependent variable:
	Sum of patent count
Similarity 1-5 years ago	45,002.63**
	(21,802.09)
Similarity 6-10 years ago	-24,000.06
	(25,804.58)
R^2	0.076
Adjusted R^2	-0.052
Note:	*p<0.1; **p<0.05; ***p<0.01

Figure 37: Regression results for model with moving average (periods of maturity)

In the following, we will focus on model (5), and the model of periods of maturity, when comparing the findings. Model (5) was chosen due to the highest R squared from all five models with lagged similarities per individual year.

Similar to model (5), the model looking at periods of maturity uncovers a significant impact of similarities across patent thematics from the near past on the patenting at a given time. Both the significant variable of a lagged similarity from 1 year ago from model (5) and the variable of the near past from the model of periods of maturity, seem to correspond in their finding that the near past has a positive impact on the patenting at a given time. While model (5), and indeed all five models revealed that the lagged similarity of 3 years ago is associated with less patenting, the averaging of similarities across five years, indicates unequivocally that there is a positive impact of synergies across patent thematics of recent years on subsequent patenting, as does the lagged similarity of 1 year ago of model (5). While model (5) initially yielded the similarity of 10 years ago as being associated with an increase in patenting and the similarity across patent classes of 9 years ago as being associated with a decrease in patenting, the new model reveals that synergies across patent classes in the distant past do not seem to have an impact on the patenting behavior of assignees.

First of all, from both models, we can confirm hypothesis (H3) that similarities across classes in a given year have a positive impact on the patent count at one point. From model (5) we assume, that the similarity between patent classes has, on the one hand, a short-term effect and on the other hand a long-term effect. While a higher similarity between classes is associated with a higher patent count in the subsequent year, indicating a short-term effect of the exploitation of synergies between patent classes, there is also a positive long-term effect, showing up after ten years. Thus, the synergies created between classes have something inherent to them that is exploitable for the subsequent year, it could be hypothesized that the resulting patents in the subsequent year are close to the patents of the patent classes in the previous year and hence constitute small innovations, improving the previous ones. This is in line with findings from Figure 32 that indicated a decrease of similarity with time. On the other hand, the long-term effect after ten years might indicate that the synergies between different patent classes in a given year might launch the development of original innovations. The time lag of ten years is in concordance with the time of development for new drugs that Ding, Eliashberg, and Stremersch (2014) indicated. These findings equally correspond to Achilladelis and Antonakis (2001)'s postulation that companies are interested in introducing incremental innovations to their radical innovations since this enables them to delay the competitor's emergence on the market. The time lag of 9 to 13 years, or 8 to 10 as Achilladelis and Antonakis (2001) state, that it takes competitor's to develop an own innovation that is able to compete with the newly commercialized radical innovation of a firm can be continuously extended when patenting incremental innovations. Hence, the increasing patent count after a time lag of a year after the introduction of patents close in patent content could represent incremental innovations.

This interpretation, however only holds up, when looking at model (5). When looking at an impact of time spans and not at similarities of individual years, the interpretation has to be amended. While the model, we called model of periods of maturity, confirmed a positive impact of patent synergies in the near past on patenting behavior, the impact of the distant past does not hold up. Thus, the interpretation that incremental innovations might be responsible for the increased patenting after 1 year might apply for the model of periods of maturity as well. Synergies across patent thematics in the near past might initiate incremental innovations on existing innovations and hence increases the patenting in the short-term. According to the model of periods of maturity, there does not seem to be a long-term effect of patent thematic synergies. Thus according to this model, synegies across patent thematics would not initiate original innovations on the long-term, but incremental innovations on the short-term.

It follows, that with regard to a long-term effect, the results are inconclusive and require further research. With regard to a short-term effect, however, we found that synergies across patent thematics have a positive impact on the patenting in close years. This might indicate that firms, that tend to adopt more and more integrated research approaches, count on incremental innovations arising from the cooperation across research departments. This might be a strategic decision enabling pharmaceutical firms to stay competitive in an environment where the pace of innovation is increasing and where the protection of a firm's innovations is crucial for its success.

7 Discussion

7.1 Main findings

This thesis has described innovation dynamics in the Chinese pharmaceutical industry, leveraging Data Science techniques. The development of a protocol allowed to capture the innovation dynamics to an extent that would not have been possible without computational methods. Algorithms have been developed that enabled us to analyze 228,108 patents, whereof 69,923 Chinese patents. Text mining allowed to extract the relevant information for this thesis from text files. To recapitulate, this thesis' motivation is to show how innovation dynamics in the pharmaceutical industry in China can be described, leveraging Data Science techniques.

First of all, our analysis has shown that indeed China is worth the analysis of its innovative capabilities. While slowly emerging as a topic in research, scholars have not accorded China enough attention. Only a few researchers pointed out China's emergence in global innovation networks and analyzed the national innovation system where its R&D efforts are embedded in (Lundvall 2009; Sesay, Yulin, and Wang 2018; Zhang and Zhou 2015). Our descriptive statistics have highlighted the emergence of China in terms of innovation when comparing patent applications of 39 different patent offices. Thus, the focus of this thesis is highly relevant in the business field. Not only is it more important to learn about China's innovation capabilities for research professionals, but also for managers that have to ensure their firm's survival in highly competitive environments.

As a global industry, the pharmaceutical industry is an interesting unit of analysis when measuring innovation dynamics through a patent analysis. While in a first step, descriptive statistics could confirm that it is of interest to look at China, the motivation focused on the description of innovation dynamics in China. To this end, three distinctive questions were set out.

(1) Where lies the focus of innovation in the Chinese pharmaceutical industry? For the whole time span of 1990 to 2017? For 1990 to 2000? For 2001 to 2017?

An unsupervised (semi-supervised) Latent Dirichlet Allocation (LDA) analysis allowed to model the topics of pharmaceutical innovations in China for the three time spans. Overall, for 1990 to 2017, the topics were on cancer treatments, aminos, and their functioning, diseases (i.e., diabetes), antibodies/immunology, and finally polynucleotides, the specific form of a homo- or heteropolymer. The time span 1990 to 2000 focused instead on cancer, diabetes, antibodies/immunology, hormone treatments. The last topic could not be summarized under a specific direction; the terms included were formula, alkoxy, treatment, salts, and phenyl. Finally, the time span 2001 to 2017 centered around cancer, diabetes, general diseases, aminos, and polynucleotides (specific homo- or heteropolymers). The topics on cancer and diabetes figured in both the earlier time frame and the more recent one, as well as in the whole sample dataset of 24,087 Chinese patents. Apart from cancer and diabetes, in earlier times, Chinese pharmaceutical innovations focused on antibodies/immunology related innovations, as well as hormone treatments. These topics do not appear in the 2001 to 2017 dataset. Instead, innovations touch upon aminos and their functioning, general diseases and polynucleotides (specific homo- or heteropolymers). Therefore, pharmaceutical innovation has changed its focus over time. The hypothesis (H1.2) *China's innovation in earlier times (1990-2000) differs from the* *thematics patents cover after 2000.* has been confirmed since only two topics, cancer, and diabetes, figure in both time spans. The hypothesis (H1.1) *Overall, innovation in the Chinese pharmaceutical industry focuses on topics such cancer, biotech, diabetes, diseases such as liver and cardiovascular diseases; this applies in particular to innovation after 2001.* is generally confirmed. Indeed, cancer, biotech (represented by the topic antibodies/immunology (Chervenak 2005)), diabetes and cardiovascular diseases figure in the 1990 to 2017 dataset and the 2001 to 2017 dataset. Nonetheless, liver diseases and cardiovascular diseases do not figure as separate topics. In conclusion, not only could the expert's opinion on China's recent research focus be confirmed by a data analysis (e.g., Ni et al. 2017; Zhang and Zhou 2017; Shen 2010), but the findings from the LDA analysis were also reasonably well in concordance with the patent thematics identified from the descriptive statistics on the Derwent classes and the manual codes. Finally, the topic modeling has shown that innovations of different therapeutic groups rely to different degrees on the same research. This could be shown by matching the patent abstracts of a therapeutic group against the five topics modeled. Thus, there is collaboration across research fields.

While these results have provided confirmation for the expert's opinions on innovation in the Chinese pharmaceutical industry, it has more so shown that China has been focusing on relevant topics, such as cancer from the beginning. While the literature on R&D internationalization indicated (e.g., Grimes and Miozzo 2015), (1) that MNEs were initially hesitant to patent in China, and that (2) China's innovative capabilities were basic in the past, and (3) MNEs research activities in China are still basic nowadays, our analysis has revealed that China has focused on a key driver of growth in the pharmaceutical industry, the oncology area (Albrecht et al. 2016). Hence, we conclude that innovation in China is not basic. Patent application numbers and patent thematics indicate that China is an important player on the global market and has focused its innovation efforts on critical key areas, such as cancer, since 1990.

The second question of this thesis was answered by the Jaccard similarity measurement (Jaccard 1902) across patent manual code sections/classes. Measuring the similarity of patents across classes allows to capture an aspect of innovation dynamics that has been neglected by researchers so far. This aspect is however essential when describing the dynamics of innovation since it allows to capture cooperation and in a way, knowledge flows across different therapeutic areas and research fields. Thus, the second question was:

(2) What kind of dynamics do we find between patent thematics/classes? More precisely: What is the relationship of patent classes across time?

Our measurement has shown that patent classes are similar to other classes with average similarities ranging between 0% and 25%. Thereby, patent classes have increased their similarity with other classes over the years. In particular, from 1990 to 2000 similarities increased steadily. From 2000 on, the average Jaccard similarities per class ranged between 17% and 25%. A Jaccard similarity of 25% can be considered high since it means that 25% of the text is present in another text. Since the range across classes is close, we can assume that all classes collaborate to a similar extent with other classes. How can we explain the growing measurement of the similarities? One could assume that firms in China became more conscious about collaborating across departmental borders, which is in line with the emergence of the integrated and System Integration and Networking models (SIN) that proclaimed the integration of departments for the benefit of innovation (Rothwell 1993). In the same vein, the literature on organizational economics and R&D organization promoted an integrated network model where different R&D departments dispersed inside and across firm-boundaries collaborate (Gassmann and von Zedtwitz 1999). In case of knowledge flows between subsidiaries, Mudambi (2002) called these collaborative knowledge flows. Since the data measured the similarity across time, it indicated not only an increasing collaboration across classes in a given year but also an increasing reliance on recent innovations. An increasingly competitive environment leading to an increasing pace of innovation and shortening of the life cycle of pharmaceutical innovation might account for the growing reliance on former innovations and collaboration across research fields (Albrecht et al. 2016).

The Jaccard similarity measurement allowed us to identify the role of the time component in the relationship across patent classes. It has been shown that with an increasing time lag, the average Jaccard similarities of classes decreased. This decrease indicates that innovations rely less and less on previous innovations the more time passes. Researchers seem to exploit existing innovations and their knowledge from other domains the closer the other knowledge is in time. After a time lag of 23 years, the similarities do not seem to decrease anymore with an average Jaccard similarity of approximately 3%. Thus, we assume that after 23 years, the patents do not rely on the previous innovations anymore. Finally, the patent count was introduced into the analysis, since we wanted to assess whether the fact of cooperating and leveraging knowledge across patent thematics has an impact on future patenting behavior. The third question was dedicated to this analysis, but more precisely looked at similarities across a given year.

As mentioned, this thesis constitutes a proof of concept, revealing the explanatory power of the similarity between patent thematics. Thus, the relationship between the similarity of patent classes and the patent count when taking into account a time component has been the topic of a separate analysis answering the third question:

(3) What is the relationship between the similarity of patent classes and the patent count when taking into account the time lag?

The econometric analysis of a panel linear model with twoway fixed effect estimation highlighted the relationship between lagged similarities and the patent count. The results indicated four statistically significant variables: the similarity of 1 year ago (statistically significant at the 1% level), of 3 years ago (at the 5% level), of 9 years ago (at the 10% level), and of 10 years ago (at the 5% level). While the relationship between the lagged Jaccard similarity and the patent count is positive for the variables *similarity of 1 year ago* and *similarity of 10 years ago*, it is negative for the other two. Thus, after one and ten years, an increasing Jaccard similarity between two manual code sections is associated with an increase in patent count. In consequence, the hypothesis (H3) *We propose that similarities across classes in a given year have a positive impact on the patent count at one point* has been confirmed. Nonetheless, there have been two statistically significant variables that exhibited a negative relationship with the patent count. With regards to the two positive variables, we assume that the lagged Jaccard similarity has a short-term and long-term effect on the patent count. We assume that similarities between patent classes characterize knowledge flows between patent research fields and hence stand for the creation of synergies across classes. With respect to the short-term effect, these synergies might have something inherent to them that is exploitable in a subsequent year, where more patents are filed. The increased patenting might result from incremental innovations and thus improving, and relying on the innovations of a year ago. Resulting in the increased pace of innovating, the competitive environment of the pharmaceutical industry might have fostered this development of relying on knowledge flows across patent thematics for incremental innovations. Indeed, Achilladelis and Antonakis

(2001) explained that firms are interested in introducing incremental innovations as a form of delaying a competitor's emergence on the market. This idea of a short-term effect will have to be cause for future research. The long-term effect, on the other hand, might indicate that the synergies between different patent classes in a given year might launch the development of original innovations. The increasing competition in the industry might be the reason for this effect as well. We assume that the time lag of 10 years creates original innovations since it is in concordance with Ding, Eliashberg, and Stremersch (2014)'s indication of the development times for new drugs. Finally, an econometric model was built regrouping the similarities in a variable of the near past and a distant past. This allowed in a way to check the robustness of the previous short-term and long-term effect against periods of maturity. Indeed, the results indicated that the synergies across patent thematics of the near past (1 to 5 years ago) were associated with an increase in patent count. The distant past, however was not significant. Hence this model confirmed a positive short-term effect, that indicates towards incremental innovations originating from patent thematic synergies in the near past. The long-term effect is however ambiguous, since the new model, that functioned as a kind of robustness check did not yield the distant past as significant. Further research is needed on these hypothesis the econometric models raised. Although it has some limitations, this econometric methodology seems to be an interesting research protocol and would require further investigation. Hence, it is a proof of concept of using similarity scores in an econometric model. Overall, it raises a new light on the innovation process, by adopting a more modular approach than existing innovation models, such as the integrated model (Rothwell 1994). Concerning the specific case of the pharmaceutical industry in China and its innovation process, it moreover introduces a time component.

In conclusion, the application of Data Science techniques has enabled us to describe innovation dynamics from a multitude of angles. Thereby, attention has been devoted to the Chinese pharmaceutical industry's research focus and synergies across patent classes, in particular. Our findings have in part confirmed the literature's indication but have even more so raised new avenues to describe innovation dynamics that are beneficial for scholars and business professionals equally.

7.2 Theoretical and practical contributions

Most of the publications on innovation in developing countries focus on the macro environment, explaining the reasons for and strategies of developing countries to foster innovation or explaining firm strategies to innovate. China, while gaining attention from business professionals has thereby only gained limited attention from scholars, and mostly from a qualitative perspective (Ovide 2017). Thus, this thesis pointed out the lack of research on innovation dynamics in China, in particular in the pharmaceutical industry. The lack of literature on this topic might thereby be due to the availability of data. How to describe innovation dynamics? Data Science techniques enabled us to get access to a variety of data that was not accessible to researchers before. Using a patent proxy, computational methods made the analysis of 228,108 pharmaceutical patents possible. The code developed to extract patent information from text files is part of a protocol that has been developed to analyze innovation dynamics. This protocol is the main contribution of this thesis since researchers and business professionals equally can use it to analyze the industry of their interest. In fact, the repetition of the analysis for another industry and/or country would allow generalizing the results beyond the pharmaceutical industry in China. Moreover, the code and analysis could be used to look at a firm's innovation capabilities.

On a theoretical level, both the introduction of the similarity variable in the innovation process as well as the role of the time component in the relationship between similarities across patent classes and the patent count are of importance. Knowledge flows between research fields that indicate some form of cooperation increase the patenting of firms after a year or increase it in the near futur. We call this short-term effect. This finding could be base for avenues for future research. Overall, we have proven the usefulness of a similarity measurement in an econometric model and provided a detailed description of patent thematics in the Chinese pharmaceutical industry and its dynamics.

In conclusion, a part of the importance of this thesis emanates from the Data Science techniques that were employed and from the relevance of this neglected topic. Another part emanates from its contribution to the academic literature on economics of innovation, innovation management and knowledge pipelines as well as managers' decision-making process with regards to innovation management.

7.3 Limitations of the study

The results are prone to certain limitations that will be explained hereafter. First of all, the literature review has shown that there is a lack of prior research, which made the development of hypotheses difficult and thus is responsible for the – in part – exploratory character of the thesis. Moreover, despite scholars having shown the significance of patents as a proxy for innovation (Hagedoorn and Cloodt 2003; Arundel and Kabla 1998; Frost and Zhou 2000; Hadengue, de Marcellis-Warin, and Warin 2015), the approach is not able to cover all innovations. Firms might not patent some innovations. In addition, some might argue that patenting is losing its importance in times of open innovation models; many scholars, however, have argued otherwise (Chesbrough 2003; Chesbrough 2006; Wang, Vanhaverbeke, and Roijakkers 2012; West 2007). Finally, while it generally holds true that not all innovations are being patented, the pharmaceutical industry is in particular propensive to patenting their innovations (Arundel and Kabla 1998). Therefore, we assume that the results of this thesis are fairly accurate when analyzing innovation in the pharmaceutical industry.

Despite enabling us to picture innovation dynamics, the patent approach has the disadvantage of not enabling us to differentiate between the types of innovations. The patent approach might have the advantage of including not only new drug innovations but also innovations with regard to components and processes, it entails, however, that we cannot distinguish between the different types of innovations and hence limits the explanatory power of our results.

We chose to look at all pharmaceutical patents filed in China. While this allows to capture the dynamics in the Chinese pharmaceutical industry, it does not allow to identify the source of these innovations. Patents can be filed at several patent offices. Thus, our results describe the Chinese pharmaceutical industry but do not allow to estimate whether the innovation is originating from China. Given global innovation networks, open innovation models and frequent exchange of researchers this is in our opinion, neglectable. Nonetheless, it would be an interesting avenue for future research.

As has been mentioned in the results of the LDA analyses, the final comparison of the topics modeled with the framed therapeutic groups is merely a first indication of across which therapeutic groups the topics fit. An improvement of the algorithm would allow to compare the topics to all patents, even if they are in more than one therapeutic group. It was refrained from modeling the topics from the dataset that contained duplicate patents since a patent being present more than once would have influenced the classification process and thus the topics modeled.

With regards to the analysis of similarities across patent classes, results are limited, since, due to feasibility, patent classes were compared to each other, not single patents. In future research, similarities among single patents could be compared, which allows to take into account more specific attributes of patents and would avoid looking at average values.

Finally, we were not able to control for whether a patent in a patent class was present in several patent classes. Therefore the measurement of similarity might be biased due to patents being in more than one group. This potential bias is however not an issue in the econometric part since we were interested in looking at whether synergies across patent classes were reflected in the patenting amount in subsequent years. Whether these synergies originate due to being the same patent or due to similar research areas touched upon in the patent, it indicates that two patent classes experience links with each other.

While these limitations exist, future research could tackle most of them. And although they might limit the results of this thesis, our research exhibits a large explanatory power nonetheless. The results of the computational analysis of 69,923 patents in the Chinese pharmaceutical industry might be limited in a few regards, but the Data Science approach allows for the first time to analyze this amount of textual data, that could not be explored to the same extent with traditional methodological approaches. Finally, this thesis contained several proofs of concept that therefore lay the ground for future research.

7.4 Avenues for future research

The results of this thesis raise fruitful areas for future research on innovation dynamics in an industry. For instance, it would be interesting to check the present results across borders. How would the results look like if analyzing the pharmaceutical industry in the US for instance? The US, lying at the center of the network of the global pharmaceutical industry might be an interesting unit of research (Hu et al. 2015). A comparison between the results of the US and the Chinese pharmaceutical industry would enable us to identify whether China has indeed been focusing on specific topics, that it is especially apt to treat, or whether the global pharmaceutical industry is of such a global character that borders and country specificities do barely matter. The literature on R&D internationalization and knowledge pipelines indicated that global innovation networks had been created that result in knowledge flowing across borders easily (e.g., Cantwell and Mudambi 2011; Gassmann and von Zedtwitz 1999; Mudambi and Swift 2011; Turkina and van Assche 2018). Nonetheless, researchers have shown that geographical proximity plays a role (e.g., Gugler, Keller, and Tinguely 2015; Hadengue, de Marcellis-Warin, and Warin 2015; Jaffe, Trajtenberg, and Henderson 1993; Williamson and Masten 1999). Moreover, according to Porter (1990) countries dispose of certain comparative advantages that influence firm behavior. The Chinese government actively fostering innovation in China is a prominent example. In conclusion, it is of importance to repeat the present analysis for other countries and to compare all results with each other. This repetition would be beneficial for the generalizability of the results of the Jaccard and econometric analysis. It could confirm the importance of the similarity in the innovation process.

The comparison of Jaccard similarities across time would equally allow researchers to compare the US and China and assess whether China's patents have become more dissimilar to US patents in time. R&D internationalization scholars would hypothesize that China's innovation output in terms of patents might have been initially very close to the US output, but has become more dissimilar since the country gained innovation capabilities on its own and might leverage its national innovation system that is growing in strength.

As has been evoked in the limitations section, it would be interesting to repeat the analysis on the subset of Chinese pharmaceutical patents filed first at the Chinese patent office, the SIPO. The analysis on the subset of Chinese pharmaceutical patents would allow to compare its results with the results of the whole dataset. Moreover, the results of the topic modeling through the LDA would be in particular interesting, since they allow to assess on what topics patents originating in China focus on.

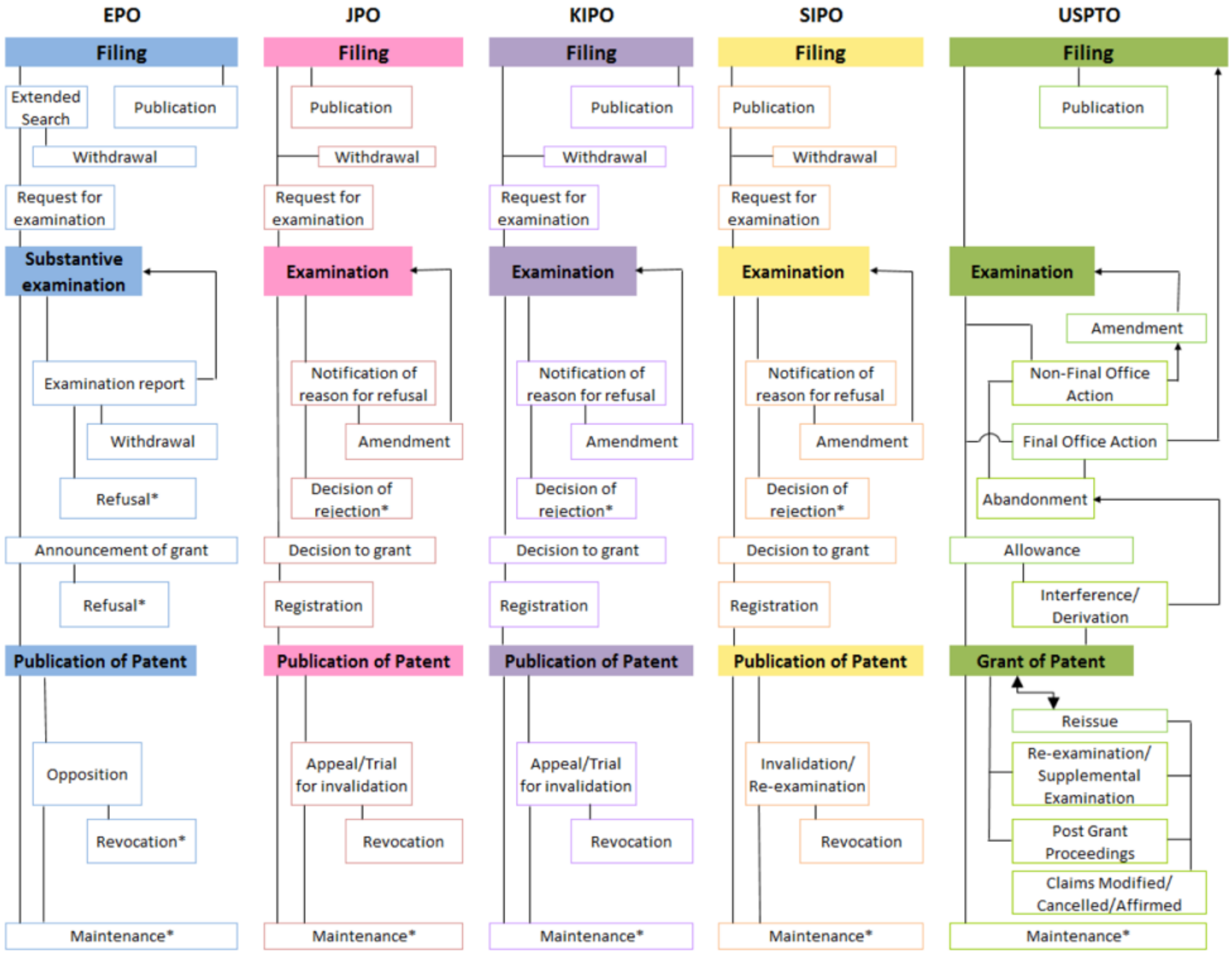
8 Conclusion

This research describes innovation dynamics in the Chinese pharmaceutical industry, leveraging Data Science techniques. We developed a protocol based on these techniques, which allows the researcher to reproduce our research in different settings quickly. With regard to our results, first, text mining of 228,108 pharmaceutical patent has foremost revealed the emergence of China as an innovator on a global scale. The Derwent World Patent Index provided access to a unique, extensive and up-to-date patent database of patents in .html and .txt format, which were converted in a dataset through text mining. Further analysis of the patent data, revealed China's research focus on cancer treatments, aminos and their functioning, diseases (i.e., diabetes), antibodies/immunology, and finally polynucleotides, the specific form of a homo- or heteropolymer. These were the research fields that emerged when computationally classifying 69,923 Chinese pharmaceutical patents from 1990 to 2017. When distinguishing between time frames, antibodies/immunology is a topic appearing in earlier patents from 1990 to 2000, while the topics aminos and their functioning, general diseases and polynucleotides (specific homo- or heteropolymers) figure among more recent innovations, from 2001 to 2017. In a next step we looked at the similarity between patent thematics that represented knowledge flows between research fields, a modular perspective that has not been adopted neither in the literature on economics of innovation nor on knowledge pipelines. Collaborative knwoledge flows only gained attention across subsidiaries (Gassmann and von Zedtwitz 1999; Mudambi 2002; Rothwell 1994). More specifically, we assumed that knowledge flowing between research fields, created certain synergies between these fields that can be measured by the similarity of patent thematics. The analysis of similarities across patent classifications, measured by the Jaccard similarity index, revealed that innovations rely less and less on previous innovations the more time passes. Finally, an econometric analysis has shown that the similarity across classes has – what we call – a short-term effect on the patent count. A higher patent count is associated with the similarity of one year ago, and the similarities in the near past. Thus, we hypothesize that the short-term effect might be due to the patenting of incremental innovations with the smallest delay. This would slow down competitors in an increasingly competitive environment. The results are inconclusive with regards to a long-term effect or the effect of similarities in the distant past. While the similarity of 10 years ago indicates a long-term effect on the patent count, which might indicate the patenting of original innovations that draw on innovations from ten years ago, the analysis from the perspective of periods of maturity did not confirm this finding. Foremost, the results have shown that synergies across patent thematics can have a positive impact on patenting in subsequent years. Thus, we have proven the usefulness of a variable "similarity" for the innovation process, which had not been basis for research yet.

This research extends the literature on economics of innovation, knowledge pipelines and finally innovation in developing countries, a field that has not received enough attention to date. It increases our understanding of developing countries moving towards innovative capabilities. While former studies concentrated on describing the national innovation system and foreign MNE's R&D internationalization to China, we are, to our knowledge, the first leveraging a unique, vast dataset of patent data for the analysis of specificities of innovation in China. Researchers in economics and international business have to realize the immense possibilities Data Science techniques can offer them. This thesis would not have been feasible without the said techniques. Indeed, research questions that could previously not be covered, such as the questions of this thesis, can nowadays be tackled by Data Science techniques. Apart from its theoretical contribution, this thesis supports managers' decision-making process with regards to innovation management.

Further studies are needed to extend the initial understanding of innovation dynamics in the pharmaceutical industry in China and to allow the generalizability of our results. In particular, the comparison with other countries and industries would help with theory development.

9 Appendix



* Decision may be appealed

Figure 38: Patent procedures at the world's five largest IP offices (the IP5)

Source: EPO, JPO, KIPO, SIPO, USPTO, 2016, p. 75

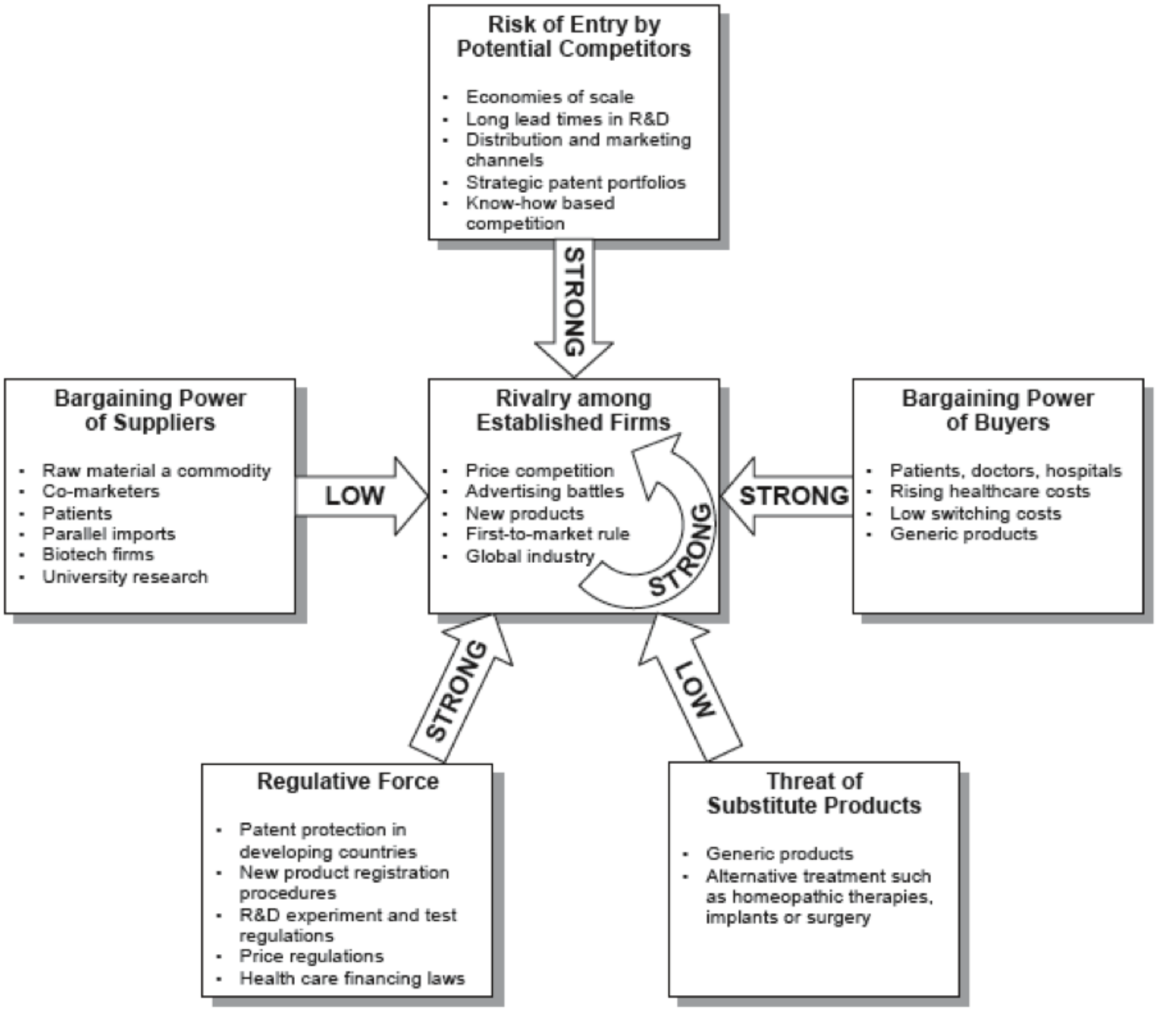

Risk of Entry by Potential Competitors
Economies of scale
Long lead times in R&D
Distribution and marketing channels
Strategic patent portfolios
Know-how based competition
STRONG
Bargaining Power of Suppliers
Raw material a commodity
Co-marketers
Patients
Parallel imports
Biotech firms
University research
LOW
Rivalry among Established Firms
Price competition
Advertising battles
New products
First-to-market rule
Global industry
STRONG
STRONG
Bargaining Power of Buyers
Patients, doctors, hospitals
Rising healthcare costs
Low switching costs
Generic products
STRONG
LOW
Regulative Force
Patent protection in developing countries
New product registration procedures
R&D experiment and test regulations
Price regulations
Health care financing laws
Threat of Substitute Products
Generic products
Alternative treatment such as homeopathic therapies, implants or surgery

Figure 39: Analysis of Porter's Five-Forces for the pharmaceutical industry

Source: Gassmann (2008): p. 23; after Porter (1985)

Table 5: Choice of keywords for the extraction of therapeutic groups

Therapeutic group	**Keywords**
Antiparasitic	*parasit*
Anti Cancer	[Cc]ancer*; [Oo]ncolog*; *cancer*; [Mm]alignant tumor*
Biotechnology	[Gg]enetic* [Ee]ngineer*; [Bb]iotech*; [Bb]io [Tt]ech*; [Bb]iopharmaceut*; [Bb]ioengineer*; [Bb]io* [Ee]ngineer*; [Bb]io-tech*; [Gg]ene editing*; DNA* [Cc]omput*
Hormonal	[Hh]ormon*
Genitourinary	[Gg]enitourinary*; [Gg]enital; [Uu]rinary*; [Uu]rogenital*; [Uu]rolog*
Blood and Clotting	[Bb]lood*; [Cc]oagulation; [Cc]lot*
Sensory	[Ss]ensory; SPD
Respiratory	[Rr]espirat*
Dermatological	[Dd]ermatolog*
Immunological	[Ii]mmunological*; *immun*; [Ii]mmun
Cardiovascular	[Cc]ardiovascular*; CVD*; [Cc]irculatory*
Musculoskeletal	[Mm]usculoskelet*; MSD*
Alimentary Metabolic	[Aa]liment*; [Mm]etaboli*; [Dd]iabet*
Anti Infective	[Aa]nti-Infective*; [Ii]nfecti*
Neurological	[Nn]eurolog*; [Nn]ervous [Ss]ystem*

Note: The letters in brackets, signify that both upper- and lowercase letters have to be taken into account. The star * means that all beginnings or endings of the word are considered by the code. [Dd]iabet*, for instance takes into account both diabetes and diabetic.

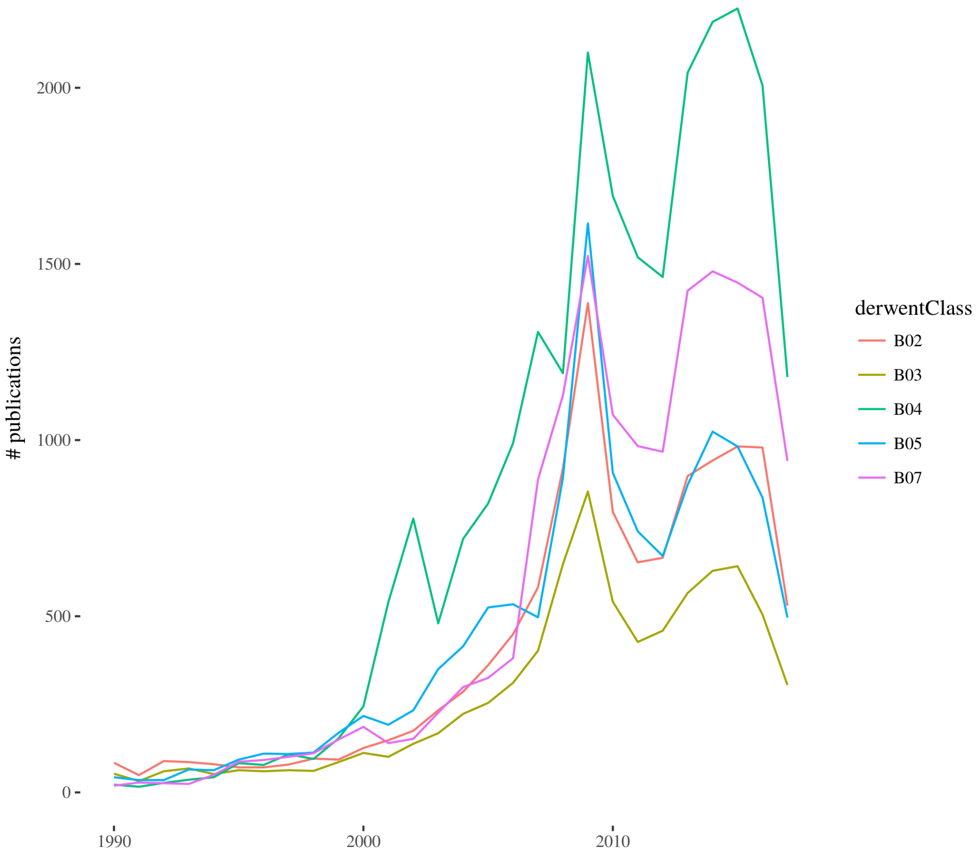


Figure 40: Development of pharmaceutical Derwent classes (1990-2017)

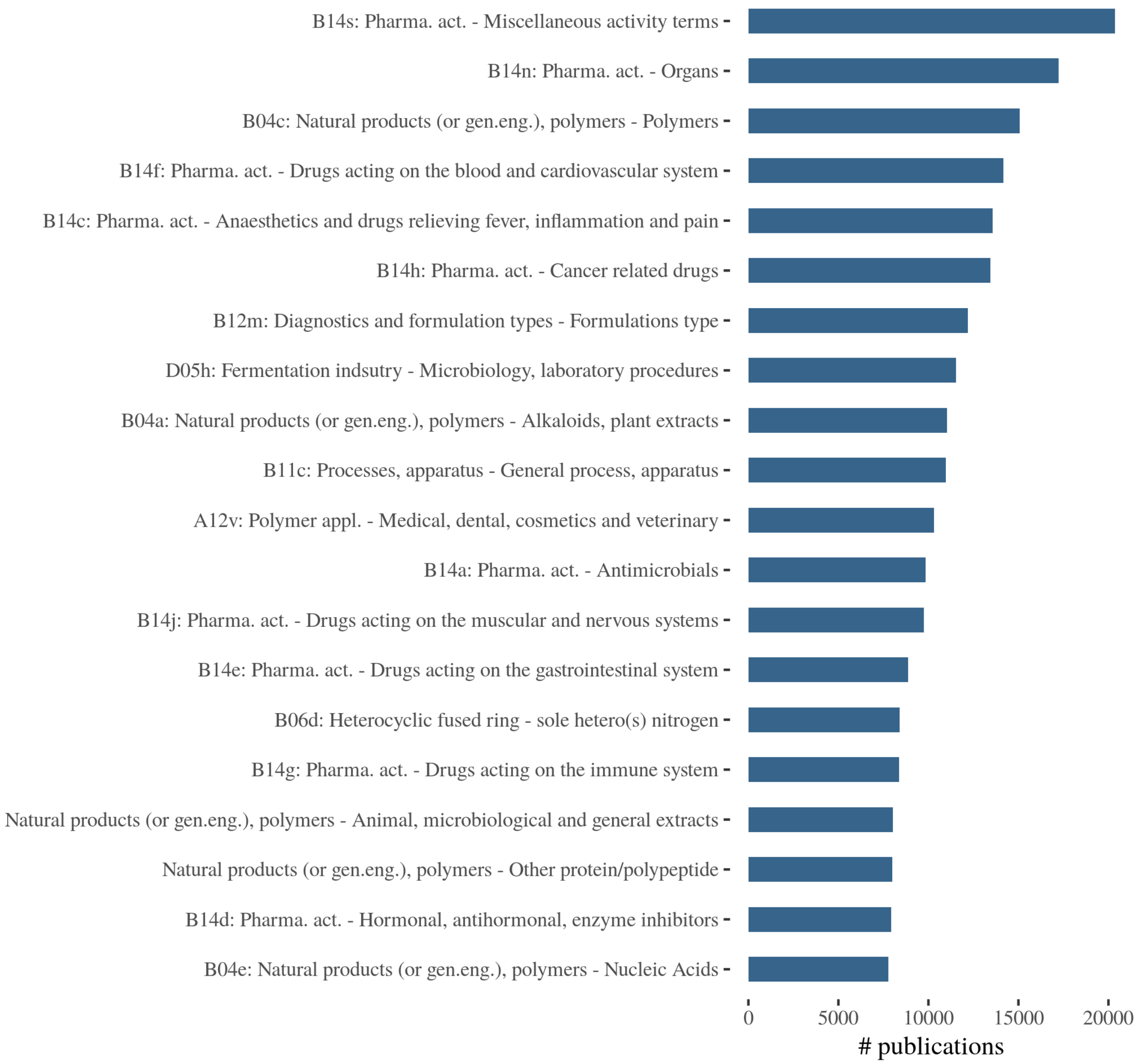


Figure 41: Top 20 manual code sections in terms of numbers of patents (1990-2017)

Table 6: Patent-issuing authorities included in the dataset

Patent-issuing authorities	
AU	Australia
AT	Austria
BE	Belgium
BR	Brazil
CA	Canada
CN	China
CS	Czechoslovakia
CZ	Czech Republic
EP	European Patents - European Patent Office
FI	Finland
FR	France
DE	Germany
DD	Germany - East
HK	Hong Kong
HU	Hungary
IN	India
IL	Israel
JP	Japan
KR	Korea
MX	Mexico
NL	Netherlands
NO	Norway
WO	Patent Cooperation Treaty - World Intellectual Property Organisation
PH	Philippines
PL	Poland
PT	Portugal
RO	Romania
RU	Russian Federation
SG	Singapore
SK	Slovakia
ZA	South Africa
ES	Spain
SE	Sweden
CH	Switzerland
TW	Taiwan
TH	Thailand
GB	United Kingdom
US	United States of America
VN	Vietnam

Table 7: Signification of top 20 manual code sections in China

Manual code sections	Signification	
B14-N	Pharmaceutical activities	Organs
D05-H	Fermentation industry	Microbiology, laboratory procedures
B04-A	Natural products (or genetically engineered), polymers	Alkaloids, plant extracts
B14-S	Pharmaceutical activities	Miscellaneous activity terms
B14-F	Pharmaceutical activities	Drugs acting on the blood and cardiovascular system
B04-C	Natural products (or genetically engineered), polymers	Polymers
B14-C	Pharmaceutical activities	Anaesthetics and drugs relieving fever, inflammation and pain
B04-E	Natural products (or genetically engineered), polymers	Nucleic acids
B12-M	Diagnostics and formulation types	Formulations type
B14-J	Pharmaceutical activities	Drugs acting on the muscular and nervous systems
B14-H	Pharmaceutical activities	Cancer related drugs
B11-C	Process, apparatus	General process, apparatus
B14-A	Pharmaceutical activities	Antimicobials
B14-E	Pharmaceutical activities	Drugs acting on the gastrointestinal system
B10-A	Aromatics and cycloaliphatics (mono and bicyclic only), aliphatics	Rarer chemical groups general
B14-G	Pharmaceutical activities	Drugs acting on the immune system
B14-D	Pharmaceutical activities	Hormonal, antihormonal, enzyme inhibitors
A12-V	Polymer applications	Medical, dental, cosmetics and veterinary
B07-D	Heterocyclics, mononuclear	Sole hetero(s) nitrogen
B06-D	Heterocyclic fused ring	Sole hetero(s) nitrogen

Source: Derwent World Patents Index (2018)

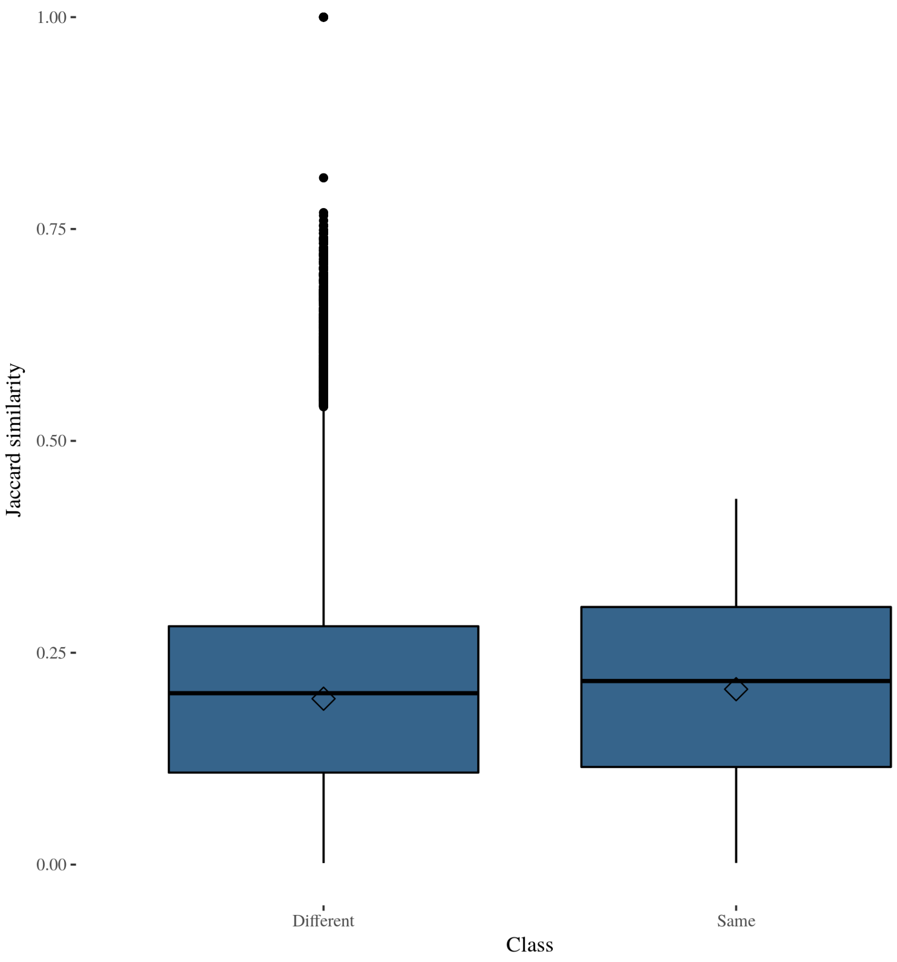


Figure 42: Distribution of Jaccard similarity for different/same classes

Table 8: Manual code classes with the Jaccard similarity 1

Class 1	Class 2	Jaccard similarity
b14d1991	b14f1991	1
b14j1990	b14n1990	1
b14h1993	b14n1993	1
b14d1991	b14s1991	1
b14f1991	b14s1991	1

In 1991, the manual code classes B14-D, B14-S and B14-F contained the same patents. These sections, all on pharmaceutical activities, represent B14-S, *Miscellaneous activity terms*, B14-D, *Hormonal, antihormonal, enzyme inhibitors* and B14-F, *Drugs acting on the blood and cardiovascular system*. Thus, in 1991 the therapeutic groups cardiovascular and hormonal were close in research. In 1990, B14-J and B14-N were congruent. Thus, since B14-J, *Drugs acting on the muscular and nervous system* and B14-N, *Organs*, contained the same patents in 1990, these therapeutic groups were closely linked. In 1993, B14-N, *Organs*, and B14-H, *Cancer related drugs* were congruent. These findings illustrate linkages between therapeutic groups and allow to describe dynamics between sections.

Table 9: Manual code classes with high Jaccard similarity (condition: time lag, different classes)

Class 1	Class 2	Jaccard similarity
a12v2010	b12m2011	0.3923
b14c2008	b14e2009	0.3921
b14n2014	b14s2013	0.3915
b14f2008	b14j2009	0.3912
b14e2009	b14f2008	0.3908
b14c2008	b14j2009	0.3901
a12v2009	b12m2008	0.3897
b14j2009	b14n2008	0.3894
a12v2010	b12m2008	0.3892
a12v2010	b12m2009	0.3886

When taking a sample of the classes compared to each other across time and classes, the finding from Figure 31 is reflected. The highest Jaccard similarities across classes and time register a time lag of one year (Appendix: Table 9). It is noticeable that the highest Jaccard similarity is around 0.3923. Thus, while similar to some extent, there is a large part that is rather dissimilar.

10 Bibliography

"A Closer Look at the 13th Five-Year Plan." 2016. The Economist Intelligence Unit, June.

Abbas, Assad, Limin Zhang, and Samee U. Khan. 2014. "A Literature Review on the State-of-the-Art in Patent Analysis." World Patent Information 37 (Journal Article): 3–13. doi:10.1016/j.wpi.2013.12.006.

Achilladelis, Basil, and Nicholas Antonakis. 2001. "The Dynamics of Technological Innovation: The Case of the Pharmaceutical Industry." Research Policy 30 (4): 535–88. doi:10.1016/S0048-7333(00)00093-7.

Aghion, Philippe, and Rachel Griffith. 2005. Competition and Growth: Reconciling Theory and Evidence. Book, Whole. Cambridge, Mass: MIT Press.

Aghion, Philippe, Nick Bloom, Richard Blundell, Rachel Griffith, and Peter Howitt. 2005. "Competition and Innovation: An Inverted-U Relationship." The Quarterly Journal of Economics 120 (2): 701–28. doi:10.1093/qje/120.2.701.

Aghion, Philippe, Peter Howitt, Maxine Brant-Collett, and Cecilia García-Peñalosa. 1998. Endogenous Growth Theory. Book, Whole. Cambridge, Mass: MIT Press.

Albrecht, Björn, Philippe Menu, Jeff Tsao, and Kevin Webster. 2016. "The Next Wave of Innovation in Oncology."

Alschner, Wolfgang, and Dmitriy Skougarevskiy. 2016. "Mapping the Universe of International Investment Agreements." Journal of International Economic Law 19 (3): 561–88. doi:10.1093/jiel/jgw056.

Alschner, Wolfgang, Julia Seiermann, and Dmitriy Skougarevskiy. 2017. "Text-as-Data Analysis of Preferential Trade Agreements: Mapping the PTA Landscape." SSRN Scholarly Paper ID 2999800. Rochester, NY: Social Science Research Network.

Altenburg, Tilman, Hubert Schmitz, and Andreas Stamm. 2008. "Breakthrough? China's and India's Transition from Production to Innovation." World Development 36 (2): 325–44. doi:10.1016/j.worlddev.2007.06.011.

Altmann, P., and R. Engberg. 2016. "Frugal Innovation and Knowledge Transferability." RESEARCH-TECHNOLOGY MANAGEMENT 59 (1): 48–55. doi:10.1080/08956308.2016.1117323.

Ames, Edward. 1961. "Research, Invention, Development and Innovation." The American Economic Review 51 (3): 370–81.

Arasaratnam, Ajanthy, and Gary Humphreys. 2013. "Emerging Economies Drive Frugal Innovation." BULLETIN OF THE WORLD HEALTH ORGANIZATION 91 (1): 6–7. doi:10.2471/BLT.13.020113.

Argyres, Nicholas S., and Brian S. Silverman. 2004. "R&D, Organization Structure, and the Development of Corporate Technological Knowledge." Strategic Management Journal 25 (8/9): 929–58. doi:10.1002/smj.387.

Arundel, Anthony, and Isabelle Kabla. 1998. "What Percentage of Innovations Are Patented? Empirical Estimates for European Firms." Research Policy 27 (2): 127–41. doi:10.1016/S0048-7333(98)00033-X.

Awate, S., MM Larsen, and R. Mudambi. 2015. "Accessing Vs Sourcing Knowledge: A Comparative Study of R&D Internationalization Between Emerging and Advanced Economy Firms." JOURNAL OF INTERNATIONAL BUSINESS STUDIES 46 (1): 63–86. doi:10.1057/jibs.2014.46.

Baltagi, Badi H., ed. 2001. A Companion to Theoretical Econometrics. Blackwell Companions to Contemporary Economics. Malden, Mass: Blackwell.

Barnard, Helena. 2010. "Overcoming the Liability of Foreignness Without Strong Firm Capabilities the Value of Market-Based Resources." Journal of International Management 16 (2): 165–76. doi:10.1016/j.intman.2010.03.007.

Bathelt, Harald, Anders Malmberg, and Peter Maskell. 2004. "Clusters and Knowledge: Local Buzz, Global Pipelines and the Process of Knowledge Creation." Progress in Human Geography 28 (1): 31–56. doi:10.1191/0309132504ph469oa.

Blei, David M, Andrew Y. Ng, and Michael I. Jordan. 2003. "Latent Dirichlet Allocation." Journal of Machine Learning Research 3: 993–1022.

Boutellier, Roman, Oliver Gassmann, and Maximilian von Zedtwitz. 2008. Managing Global Innovation: Uncovering the Secrets of Future Competitiveness. 3rd ed. Book, Whole. New York;Berlin; Springer.

Bruche, Gert. 2009a. "A New Geography of InnovationChina and India Rising." Transnational Corporations Review 1 (4): 24–27.

———. 2009b. "The Emergence of China and India as New Competitors in MNCs' Innovation Networks." Competition and Change 13 (3): 267–88. doi:10.1179/102452909X451378.

Cammack, Richard. 2006a. Oxford Dictionary of Biochemistry and Molecular Biology. 2nd ed. Oxford: Oxford Univ. Press.

———. 2006b. "Polynucleotide." Oxford Dictionary of Biochemistry and Molecular Biology. Oxford: Oxford Univ. Press.

———. 2006c. "Polypeptide." Oxford Dictionary of Biochemistry and Molecular Biology. Oxford: Oxford Univ. Press.

Cantwell, JA, and R. Mudambi. 2011. "PHYSICAL ATTRACTION AND THE GEOGRAPHY OF KNOWLEDGE SOURCING IN MULTINATIONAL ENTERPRISES." GLOBAL STRATEGY JOURNAL 1 (3-4): 206–32. doi:10.1111/j.2042-5805.2011.00024.x.

Carter, Charles Frederick, and B. R. Williams. 1957. Industry and Technical Progress: Factors Governing the Speed of Application of Science. Book, Whole. New York;London; Oxford University Press.

"CFDA." 2017. http://eng.sfda.gov.cn/WS03/CL0755/.

Chan, Jeremy. 2015. "China's Innovation Paradox." Perspectives: Policy and Practice in Higher Education 19 (1): 23–27. doi:10.1080/13603108.2014.992999.

Chervenak, Matthew. 2005. "China: Moving Towards Innovation in Pharma." Drug Discovery Today 10 (17): 1127–30. doi:10.1016/S1359-6446(05)03579-8.

Chesbrough, Henry. 2017. Open Innovation: The New Imperative for Creating and Profiting from Technology. Book, Whole. Place of publication not identified: Skillsoft.

Chesbrough, Henry William. 2006. Open Business Models: How to Thrive in the New Innovation Landscape. Book, Whole. Boston, Mass: Harvard Business School Press.

Chesbrough, HW. 2003. "The Era of Open Innovation." MIT SLOAN MANAGEMENT REVIEW 44 (3): 35–41.

Chiaroni, Davide, Vittorio Chiesa, and Federico Frattini. 2008. "Patterns of Collaboration Along the Bio-Pharmaceutical Innovation Process." Journal of Business Chemistry 5 (1): 7–22.

"China Strengthens Judicial Protection of Intellectual Property Through Specialized IPR Courts." 2017. Intellectual Property Protection in China, May.

Cockburn, Iain M., and Rebecca M. Henderson. 1998. "Absorptive Capacity, Coauthoring Behavior, and the Organization of Research in Drug Discovery." The Journal of Industrial Economics 46 (2): 157–82. doi:10.1111/1467-6451.00067.

Cohen, Wesley M., and Daniel A. Levinthal. 1990. "Absorptive Capacity: A New Perspective on Learning and Innovation." Administrative Science Quarterly 35 (1): 128–52.

Cooke, Philip N., Hans-Joachim Braczyk, and Martin Heidenreich. 2004. Regional Innovation Systems: The Role of Governance in a Globalized World. Routledge.

Cornell University, INSEAD, and WIPO. 2017. "Global Innovation Index 2017: Innovation Feeding the World." Ithaca, Fontainebleau,; Geneva.

"Corruption Perceptions Index 2016 - Transparency International." 2016. https://www.transparency.org/news/feature/corruption_perceptions_index_2016#table.

Crescenzi, R., A. Rodriguez-Pose, and M. Storper. 2012. "The Territorial Dynamics of Innovation in China and India." Journal of Economic Geography 12 (5): 1055–85. doi:10.1093/jeg/lbs020.

Criscuolo, Paola, Rajneesh Narula, and Bart Verspagen. 2005. "Role of Home and Host Country Innovation Systems in R&d Internationalisation: A Patent Citation Analysis." Economics of Innovation and New Technology 14 (5): 417–33. doi:10.1080/1043859042000315285.

Croissant, Yves. 2017. "Package 'Plm'."

Croissant, Yves, and Giovanni Millo. 2007. "Introduction to Plm."

———. 2008. "Panel Data Econometrics in R: The Plm Package." Journal of Statistical Software 27 (2). doi:10.18637/jss.v027.i02.

Daemmrich, AA, and ME Bowden. 2005. "A Rising Drug Industry." CHEMICAL & ENGINEERING NEWS 83 (25): 28–28.

Dang, JW, and K. Motohashi. 2015. "Patent Statistics: A Good Indicator for Innovation in China? Patent Subsidy Program Impacts on Patent Quality." CHINA ECONOMIC REVIEW 35 (Journal Article): 137–55. doi:10.1016/j.chieco.2015.03.012.

Dasgupta, Partha, and Joseph Stiglitz. 1980. “Industrial Structure and the Nature of Innovative Activity.” The Economic Journal 90 (358): 266–93. doi:10.2307/2231788.

De Queiroz, Gabriela, Oliver Keyes, David Robinson, and Julia Silge. 2018. “Package ‘Tidytext’.”

“Derwent World Patents Index.” n.d. Clarivate. https://clarivate.com/products/derwent-world-patents-index/.

Ding, Min, Jehoshua Eliashberg, and Stefan Stremersch. 2014. Innovation and Marketing in the Pharmaceutical Industry: Emerging Practices, Research, and Policies. International Series in Quantitative Marketing. Springer.

Dranove, David. 2003. “Introduction to Fixed Effects Models,” 10.

Dunning, John H. 1998. “Location and the Multinational Enterprise: A Neglected Factor?” Journal of International Business Studies 29 (1): 45–66. doi:10.1057/palgrave.jibs.8490024.

———. 2009. “Location and the Multinational Enterprise: A Neglected Factor?” Journal of International Business Studies 40 (1): 5–19. doi:10.1057/jibs.2008.74.

Edquist, Charles, Leif Hommen, and Maureen McKelvey. 2001. Innovation and Employment: Process Versus Product Innovation. Cheltenham: Edward Elgar Publishing.

Fabrizio, Kira R. 2006. “The Use of University Research in Firm Innovation.” In Open Innovation: Researching a New Paradigm. Oxford University Press.

Fagerberg, Jan. 1994. “Technology and International Differences in Growth Rates.” Journal of Economic Literature 32 (3): 1147–75.

———. 1996. “Technology and Competitiveness.” Oxford Review of Economic Policy 12 (3,): 39–51.

Fagerberg, Jan, and Martin Srholec. 2008. “National Innovation Systems, Capabilities and Economic Development.” Research Policy 37 (9): 1417–35. doi:10.1016/j.respol.2008.06.003.

Fagerberg, Jan, and Bart Verspagen. 2002. “Technology-Gaps, Innovation-Diffusion and Transformation: An Evolutionary Interpretation.” Research Policy 31 (8-9): 1291–1304. doi:10.1016/S0048-7333(02)00064-1.

Fan, Guanghua, Fan Peilei, and Ming Lu. 2012. “China’s Regional Inequality in Innovation Capability, 1995-2006.” China & World Economy 20 (3): 16–36.

Fan, Peilei. 2014. “Innovation in China.” Journal of Economic Surveys 28 (4): 725–45. doi:10.1111/joes.12083.

Feinerer, I., K. Hornik, and D. Meyer. 2008. “Text Mining Infrastructure in R.” JOURNAL OF STATISTICAL SOFTWARE 25 (5): 1–54.

Freeman, Christopher. 1982. The Economics of Industrial Innovation. 2nd – ed. Book, Whole. Cambridge, Mass: MIT Press.

———. 1987. Technology, Policy, and Economic Performance: Lessons from Japan. Book, Whole. London: Pinter.

“From Vision to Decision - Pharma 2020.” 2012. PWC.

Frost, Tony, and Changhui Zhou. 2000. "The Geography of Foreign R&D Within a Host Country: An Evolutionary Perspective on Location-Technology Selection by Multinationals." International Studies of Management & Organization 30 (2): 10–43. doi:10.1080/00208825.2000.11656786.

Fu, Xiaolan, and Yundan Gong. 2011. "Indigenous and Foreign Innovation Efforts and Drivers of Technological Upgrading: Evidence from China." World Development 39 (7): 1213–25. doi:10.1016/j.worlddev.2010.05.010.

Furman, Jeffrey L., Margaret K. Kyle, Iain Cockburn, and Rebecca M. Henderson. 2005. "Public & Private Spillovers, Location and the Productivity of Pharmaceutical Research." Annales d'Économie et de Statistique, no. 79/80: 165–88.

Gassmann, Oliver. 2008. Leading Pharmaceutical Innovation [Ressource Électronique]: Trends and Drivers for Growth in the Pharmaceutical Industry. 2nd ed.

Gassmann, Oliver, and Maximilian von Zedtwitz. 1999. "New Concepts and Trends in International R&D Organization." Research Policy 28 (2): 231–50. doi:10.1016/S0048-7333(98)00114-0.

Gassmann, Oliver, Angela Beckenbauer, and Sascha Friesike. 2012. Profiting from Innovation in China. Book, Whole. Berlin, Heidelberg: Springer.

Ge. 2017. "Alibaba, Tencent Included in Fortune Global 500 List for the First Time." South China Morning Post, July.

"Generic Market Share: Change in Unit Volume Worldwide by Region 2006-2016." 2018. Statista. https://proxy2.hec.ca:2554/statistics/864200/unit-volume-change-in-generics-market-share-worldwide-by-region/.

"Generic Pharma Industry Top Acquisitions U.S. 2015." 2018. Statista. https://proxy2.hec.ca:2554/statistics/596736/top-us-generic-pharmaceutical-industry-acquisitions/.

"Global Sector Report - Pharmaceuticals." 2017. Euler Hermes Economic Research.

Godin, Benoit, and Joseph P. Lane. 2013. "Pushes and Pulls: Hi(S)Tory of the Demand Pull Model of Innovation." Science, Technology, & Human Values 38 (5): 621–54. doi:10.1177/0162243912473163.

Godin, Benoît. 2006. "The Linear Model of Innovation: The Historical Construction of an Analytical Framework." Science, Technology, & Human Values 31 (6): 639–67. doi:10.1177/0162243906291865.

———. 2009. "National Innovation System: The System Approach in Historical Perspective." Science, Technology, & Human Values 34 (4): 476–501. doi:10.1177/0162243908329187.

Govindarajan, Vijay, and Chris Trimble. 2012. Reverse Innovation: Create Far from Home, Win Everywhere. Book, Whole. Boston: Harvard Business Press.

Grant, Robert M. 1996. "Toward a Knowledge-Based Theory of the Firm." Strategic Management Journal 17 (S2): 109–22. doi:10.1002/smj.4250171110.

Griliches, Zvi. 1990. "Patent Statistics as Economic Indicators: A Survey." Journal of Economic Literature 28 (4): 1661–1707.

Grimes, Seamus, and Debin Du. 2013. “Foreign and Indigenous Innovation in China: Some Evidence from Shanghai.” European Planning Studies 21 (9): 1357–73. doi:10.1080/09654313.2012.755829.

Grimes, Seamus, and Marcela Miozzo. 2015. “Big Pharma’s Internationalization of R&D to China.” European Planning Studies 23 (9): 1873–94. doi:10.1080/09654313.2015.1029442.

Grün, Bettina, and Kurt Hornik. 2011. “Topicmodels: An R Package for Fitting Topic Models.” Journal of Statistical Software 40 (13): 30.

Guan, Jiancheng, and Ying He. 2007. “Patent-Bibliometric Analysis on the Chinese Science Technology Linkages.” Scientometrics 72 (3): 403–25. doi:10.1007/s11192-007-1741-1.

Gugler, Philippe, Michael Keller, and Xavier Tinguely. 2015. “The Role of Clusters in the Global Innovation Strategy of MNEs Theoretical Foundations and Evidence from the Basel Pharmaceutical Cluster.” Competitiveness Review 25 (3): 324–+. doi:10.1108/CR-09-2014-0033.

Gupeng, Zhang, and Chen Xiangdong. 2012. “The Value of Invention Patents in China: Country Origin and Technology Field Differences.” China Economic Review 23 (2): 357–70. doi:10.1016/j.chieco.2012.02.002.

Hadengue, Marine, Nathalie de Marcellis-Warin, and Thierry Warin. 2015. “Reverse Innovation and Reverse Technology Transfer: From Made in China to Discovered in China in the Pharmaceutical.” Management International 19 (4): 49.

———. 2017. “Reverse Innovation: A Systematic Literature Review.” International Journal of Emerging Markets 12 (2): 142–82. doi:10.1108/IJoEM-12-2015-0272.

Hagedoorn, John, and Myriam Cloodt. 2003. “Measuring Innovative Performance: Is There an Advantage in Using Multiple Indicators?” Research Policy 32 (8): 1365–79. doi:10.1016/S0048-7333(02)00137-3.

Hagedoorn, John, and Ann-Kristin Zobel. 2015. “The Role of Contracts and Intellectual Property Rights in Open Innovation.” Technology Analysis & Strategic Management 27 (9): 1050–67. doi:10.1080/09537325.2015.1056134.

Haour, Georges, and Dominique Jolly. 2014. “China: The Next Innovation Hot Spot for the World.” Journal of Business Strategy 35 (1): 2–8. doi:10.1108/JBS-05-2013-0037.

“Health at a Glance 2017 - OECD Indicators.” 2017. Text. Health at a Glance. OECD.

Hu, Albert G.Z., Peng Zhang, and Lijing Zhao. 2017. “China as Number One? Evidence from China’s Most Recent Patenting Surge.” Journal of Development Economics 124 (January): 107–19. doi:10.1016/j.jdeveco.2016.09.004.

Hu, Albert Guangzhou, and Gary H. Jefferson. 2009. “A Great Wall of Patents: What Is Behind China’s Recent Patent Explosion?” Journal of Development Economics 90 (1): 57–68. doi:10.1016/j.jdeveco.2008.11.004.

Hu, Mei-Chih, and John A. Mathews. 2008. “China’s National Innovative Capacity.” Research Policy 37 (9): 1465–79. doi:10.1016/j.respol.2008.07.003.

Hu, Yuanjia, Thomas Scherngell, Lan Qiu, and Yitao Wang. 2015. “R&D Internationalisation Patterns in the Global Pharmaceutical Industry: Evidence from a Network Analytic Perspective.” Technology Analysis & Strategic Management 27 (5): 532–49. doi:10.1080/09537325.2015.1012058.

Huang, Kenneth G. 2010. “China’s Innovation Landscape.” Science, New Series 329 (5992): 632–33.

Huang, Shufang. 2012. “How Can Innovation Create the Future in a Catching-up Economy?: Focusing on China’s Pharmaceutical Industry.” Journal of Knowledge-Based Innovation in China 4 (2): 118–31. doi:10.1108/17561411211235721.

“ISIC 2423 of Pharmaceuticals, Medicinal Chemicals and Botanical Products.” 2017. Mondo International Data.

Jaccard, Paul. 1902. “Lois de distribution florale dans la zone alpine.” Bulletin de La Société Vaudoise Des Sciences Naturelles 38 (144): –. doi:http://dx.doi.org/10.5169/seals-266762.

Jaffe, Adam B., Manuel Trajtenberg, and Rebecca Henderson. 1993. “Geographic Localization of Knowledge Spillovers as Evidenced by Patent Citations.” The Quarterly Journal of Economics 108 (3): 577–98. doi:10.2307/2118401.

Jalfin, Shai. 2017. “Made in China: The Past, Present and Future of Chinese IPR.” Intellectual Property Watch, June.

Janne, O, and J Cantwell. 1999. “Technological Globalisation and Innovative Centres: The Role of Corporate Technological Leadership and Locational Hierarchy.” Research Policy 28 (2-3): 119–44. doi:10.1016/S0048-7333(98)00118-8.

Jiang, Marshall S., Oana Branzei, and Jun Xia. 2016. “DIY: How Internationalization Shifts the Locus of Indigenous Innovation for Chinese Firms.” Journal of World Business 51 (5): 662–74. doi:10.1016/j.jwb.2016.07.005.

Jin, Jun, Shanchao Wu, and Jin Chen. 2011. “International University-Industry Collaboration to Bridge R&D Globalization and National Innovation System in China.” Journal of Knowledge-Based Innovation in China 3 (1): 5–14. doi:10.1108/17561411111120837.

Jindra, Björn, Axèle Giroud, and Joanna Scott-Kennel. 2009. “Subsidiary Roles, Vertical Linkages and Economic Development: Lessons from Transition Economies.” Journal of World Business 44 (2): 167–79. doi:10.1016/j.jwb.2008.05.006.

Jou, YP, Gilbert Wu, and Warren Chan. 2015. “China’s Transformation into a Global Patent Powerhouse - Challenges and Trends.” Intellectual Asset Management Magazine, no. 74: 27–31.

Jun, Sunghae, and Sang Sung Park. 2013. “Examining Technological Innovation of Apple Using Patent Analysis.” Industrial Management & Data Systems 113 (6): 890–907. doi:10.1108/IMDS-01-2013-0032.

Kline, Stephen, and Nathan Rosenberg. 1986. “An Overview of Innovation.” In The Positive Sum Strategy: Harnessing Technology for Economic Growth, 275–306. National Academy Press.

Klossek, A., BM Linke, and M. Nippa. 2012;2010; "Chinese Enterprises in Germany: Establishment Modes and Strategies to Mitigate the Liability of Foreignness." JOURNAL OF WORLD BUSINESS 47 (1): 35–44. doi:10.1016/j.jwb.2010.10.018.

Kuemmerle, Walter. 1997. "Building Effective R&D Capabilities Abroad." Harvard Business Review, March.

Lamoreaux, N., and K. Sokoloff. 1999. "Inventive Activity and the Market for Technology in the United States, 1840-1920." Working Paper No. 7107. National Bureau of Economic Research.

Leiponen, Aija, and Constance E. Helfat. 2011. "Location, Decentralization, and Knowledge Sources for Innovation." Organization Science 22 (3): 641–58. doi:10.1287/orsc.1100.0526.

Li, Xibao. 2009. "China's Regional Innovation Capacity in Transition: An Empirical Approach." Research Policy 38 (2): 338–57. doi:10.1016/j.respol.2008.12.002.

Lin, Faqin, Chao Zhang, and Lin Wang. 2013. "Vertical Spillover Effects of Multinationals on Chinese Domestic Firms via Supplier-Customer Relationships." China & World Economy 21 (6): 37–57. doi:10.1111/j.1749-124X.2013.12045.x.

Lloyd, Ian. 2017. "Pharma R&D Annual Review 2017." Pharma intelligence - Informa UK.

Luan, Chunjuan, and Tienan Zhang. 2011. "Innovation in China: A Patentometric Perspective (1985-2009)." Journal of Knowledge-Based Innovation in China 3 (3): 184–97. doi:10.1108/17561411111167854.

Lundvall, Bengt- Ake. 1992. Nationa Systems of Innovation: Towards a Theory of Innovation and Interactive Learning. London: Pinter.

Lundvall, Bengt-Ake. 2009. Handbook of Innovation Systems and Developing Countries: Building Domestic Capabilities in a Global Setting. Book, Whole. Cheltenham, Glos, UK;Northamption, MA, USA; Edward Elgar.

Lundvall, Bengt-Åke. 2007. "National Innovation SystemsAnalytical Concept and Development Tool." Industry & Innovation 14 (1): 95–119. doi:10.1080/13662710601130863.

Maclaurin, W. Rupert. 1953. "The Sequence from Invention to Innovation and Its Relation to Economic Growth." The Quarterly Journal of Economics 67 (1): 97–111. doi:10.2307/1884150.

Madani, Farshad, and Charles Weber. 2016. "The Evolution of Patent Mining: Applying Bibliometrics Analysis and Keyword Network Analysis." World Patent Information 46 (Journal Article): 32–48. doi: 10.1016/j.wpi.2016.05.008.

Mansfield, Edwin. 1968. The Economics of Technological Change. Book, Whole. New York: W.W. Norton.

"Market Share Top Pharma Companies Rx Drugs Sales Globally 2024." 2018. Statista. https://www.statista.com/statistics/309425/prescription-drugs-market-shares-by-top-companies-globally/.

Motohashi, Kazuyuki. 2015. "Measuring Multinationals' R&D Activities in China on the Basis of a Patent Database: Comparing European, Japanese and US Firms." China & World Economy 23 (6): 1–21. doi: 10.1111/cwe.12133.

Mudambi, R., L. Piscitello, and L. Rabbiosi. 2014. "Reverse Knowledge Transfer in MNEs: Subsidiary Innovativeness and Entry Modes." LONG RANGE PLANNING 47 (1-2): 49–63. doi:10.1016/j.lrp.2013.08.013.

Mudambi, Ram. 2002. "Knowledge Management in Multinational Firms." Journal of International Management 8 (1): 1–9. doi:10.1016/S1075-4253(02)00050-9.

———. 2008. "Location, Control and Innovation in Knowledge-Intensive Industries." Journal of Economic Geography 8 (5): 699–725. doi:10.1093/jeg/lbn024.

Mudambi, Ram, and Tim Swift. 2011. "Leveraging Knowledge and Competencies Across Space: The Next Frontier in International Business." Journal of International Management 17 (3): 186–89. doi:10.1016/j.intman.2011.05.001.

Mullin, Theresa. 2017. "China Joins ICH in Pursuit of Global Harmonization of Drug Development Standards." US FDA Voice.

"National Programs for Science and Technology." 2017. China Through A Lens. http://www.china.org.cn/english/features/China2004/107131.htm.

Ni, Jingyun, Junrui Zhao, Carolina Oi Lam Ung, Yuanjia Hu, Hao Hu, and Yitao Wang. 2017. "Obstacles and Opportunities in Chinese Pharmaceutical Innovation." Globalization and Health 13 (Journal Article). doi:10.1186/s12992-017-0244-6.

Nilsson, Andreas. 2013. "Innovation Proxy - A Study of Patent and Economic Growth in China."

OECD. 2008. "The Internationalisation of Business R&D: Evidence, Impacts and Implications." 9264044043;9789264044043; Paris: OECD.

"OECD Statistics - Science, Technology and Patents." 2017. http://stats.oecd.org/.

OECD/EUIPO. 2016. "Trade in Counterfeit and Pirated Goods: Mapping the Economic Impact." Paris: OECD Publishing.

OECD/Eurostat. 2005. Oslo Manual: Guidelines for Collecting and Interpreting Innovation Data. 3rd Edition. Paris: OECD Publishing.

Ovide, Shira. 2017. "Who's the Copycat Now?" Bloomberg.com, June.

Papageorgiadis, Nikolaos, Adam R. Cross, and Constantinos Alexiou. 2014. "International Patent Systems Strength 19982011." Journal of World Business 49 (4): 586–97. doi:10.1016/j.jwb.2013.12.011.

Park, Walter G. 2008. "International Patent Protection: 19602005." Research Policy 37 (4): 761–66. doi:10.1016/j.respol.2008.01.006.

Patel, P., and M. Vega. 1999. "Patterns of Internationalisation of Corporate Technology: Location Vs. Home Country Advantages." Research Policy 28 (2-3): 145–55. doi:10.1016/S0048-7333(98)00117-6.

Penner-Hahn, Joan, and J. Myles Shaver. 2005. "Does International Research and Development Increase Patent Output? An Analysis of Japanese Pharmaceutical Firms." Strategic Management Journal 26 (2): 121–40. doi:10.1002/smj.436.

Perri, Alessandra, Vittoria G. Scalera, and Ram Mudambi. 2017. “What Are the Most Promising Conduits for Foreign Knowledge Inflows? Innovation Networks in the Chinese Pharmaceutical Industry.” Industrial and Corporate Change 26 (2): 333–55. doi:10.1093/icc/dtx004.

Pénin, Julien. 2011. “Open Source Innovation: Towards a Generalization of the Open Source Model Beyond Software.” Revue d’économie Industrielle, no. 136: 65–88. doi:10.4000/rei.5184.

Pietrobelli, Carlo, and Roberta Rabellotti. 2011. “Global Value Chains Meet Innovation Systems: Are There Learning Opportunities for Developing Countries?” World Development 39 (7): 1261–9. doi:10.1016/j.worlddev.2010.05.013.

Pilat, Dirk, Koen de Backer, Ester Basri, Sarah Box, and Mario Cervantes. 2009. “The Development of Global Innovation Networks and the Transfer of Knowledge.” In Innovation and Growth, 85–105. Paris: OECD Publishing. doi:10.1787/9789264073975-7-en.

Plane, Dan, and Scott Livingston. 2017. “Procedures and Strategies for Anti-Counterfeiting: China - World Trademark Review.” World Trademark Review. http://www.worldtrademarkreview.com/Intelligence/Anti-counterfeiting/2017/Country-chapters/China.

“Poisson Regression | R Data Analysis Examples.” 2018. Idre: Institute for Digital Research and Education at UCLA. https://stats.idre.ucla.edu/r/dae/poisson-regression/.

Porter, Michael E. 1990. The Competitive Advantage of Nations. Book, Whole. New York: Free Press.

Posner, M. V. 1961. “International Trade and Technical Change.” Oxford Economic Papers, New Series 13 (3): 323–41.

Prahalad, Coimbatore Krishnarao. 2006. The Fortune at the Bottom of the Pyramid. Book, Whole. Upper Saddle River, N.J: Wharton School Pub.

Radjou, Navi, Jaideep C. Prabhu, Simone Ahuja, and Kevin Roberts. 2012. Jugaad Innovation: Think Frugal, Be Flexible, Generate Breakthrough Growth. Book, Whole. San Francisco, CA: Jossey-Bass.

Ramirez, P. 2006. “The Globalisation of Research in the Pharmaceutical Industry: A Case of Uneven Development.” Technology Analysis & Strategic Management 18 (2): 143–67. doi:10.1080/09537320600624006.

Rana, Preetika, Amy Dockser Marcus, and Wenxin Fan. 2018. “China, Unhampered by Rules, Races Ahead in Gene-Editing Trials.” Wall Street Journal, January.

“Revenues of the Top 10 Pharmaceutical Markets Worldwide in 2016.” 2017. Statista.

Roberts, Edward B. 1988. “WHAT WE’VE LEARNED: MANAGING INVENTION AND INNOVATION.” Research Technology Management 31 (1): 11–29.

Rogers, Everett Mitchell. 2003. Diffusion of Innovations. 5th – ed. Book, Whole. New York: Free Press.

Romer, Paul M. 1990. “Endogenous Technological Change.” Journal of Political Economy 98 (5): S71–S102. doi:10.1086/261725.

Rothwell, R., C. Freeman, A. Horlsey, V. T. P. Jervis, A. B. Robertson, and J. Townsend. 1974. “SAPPHO Updated - Project SAPPHO Phase II.” Research Policy 3 (3): 258–91. doi:10.1016/0048-7333(74)90010-9.

Rothwell, Roy. 1992. “Successful Industrial Innovation: Critical Factors for the 1990s.” R&D Management 22 (3): 221–40. doi:10.1111/j.1467-9310.1992.tb00812.x.

———. 1993. “The Changing Nature of the Innovation Process.” Technovation 13 (1): 1–2. doi:10.1016/0166-4972(93)90009-K.

———. 1994. “Towards the Fifth-Generation Innovation Process.” International Marketing Review 11 (1): 7–31. doi:10.1108/02651339410057491.

Rothwell, Roy, and A. B. Robertson. 1973. “The Role of Communications in Technological Innovation.” Research Policy 2: 204–25.

Samara, E., P. Georgiadis, and I. Bakouros. 2012. “The Impact of Innovation Policies on the Performance of National Innovation Systems: A System Dynamics Analysis.” TECHNOVATION 32 (11): 624–38. doi:10.1016/j.technovation.2012.06.002.

Schmid, Jon, and Fei-Ling Wang. 2017. “Beyond National Innovation Systems: Incentives and China’s Innovation Performance.” Journal of Contemporary China 26 (104): 280. doi:10.1080/10670564.2016.1223108.

Schmookler, Jacob. 1966. Invention and Economic Growth. Book, Whole. Cambridge, Mass: Harvard University Press.

Schumpeter, Joseph Alois. 1934. The Theory of Economic Development: An Inquiry into Profits, Capital, Credit, Interest, and the Business Cycle. Vol. XLVI. Book, Whole. Cambridge, Mass: Harvard University Press.

———. 1950. Capitalism, Socialism, and Democracy. 3rd ed. Book, Whole. New York: Harper.

Sesay, Brima, Zhao Yulin, and Fang Wang. 2018. “Does the National Innovation System Spur Economic Growth in Brazil, Russia, India, China and South Africa Economies? Evidence from Panel Data.” Open Access, 12.

Shang, Qingyan, Jessie P.H. Poon, and Qingtang Yue. 2012. “The Role of Regional Knowledge Spillovers on China’s Innovation.” China Economic Review 23 (4): 1164–75. doi:10.1016/j.chieco.2012.08.004.

Shen, Zhu. 2010. “CHINA 2020: Walled in NO More.” Pharmaceutical Executive 30 (12): 82–88.

Shi, Lizheng, Heidi Y. Yang, Gang Cheng, and Qingyue Meng. 2014. “Time Trends and Determinants of Pharmaceutical Expenditure in China (1990-2009).” Pharmacoeconomics 32 (3): 257–64. doi:10.1007/s40273-013-0072-3.

Silge, Julia, and David Robinson. 2017. Text Mining with R. O’Reilly.

Simard, C., and Joel West. 2006. “Knowledge Networks and the Locus of Innovation.” In Open Innovation: Researching a New Paradigm, pp. 220–40. Oxford University Press.

Singh, Jasjit. 2008. “Distributed R&D, Cross-Regional Knowledge Integration and Quality of Innovative Output.” Research Policy 37 (1): 77–96. doi:10.1016/j.respol.2007.09.004.

Someren, Taco C. R. van, and Shuhua Someren-Wang. 2013. Innovative China: Innovation Race Between East and West. Book, Whole. Berlin, Heidelberg: Springer.

Stelmaszak, Marta, and Philipp Hukal. 2017. "When Data Science Meets Social Sciences: The Benefits of the Data Revolution Are Clear but Careful Reflection Is Needed." LSE Impact Blog.

Sun, Yifei. 2000. "Spatial Distribution of Patents in China." Regional Studies 34 (5): 441–54. doi:10.1080/00343400050058693.

Sun, Yutao, and Fengchao Liu. 2010. "A Regional Perspective on the Structural Transformation of China's National Innovation System Since 1999." Technological Forecasting and Social Change 77 (8): 1311–21. doi:10.1016/j.techfore.2010.04.012.

Suroso, Edy, and Yudi Azis. 2015. "Defining Mainstreams of Innovation: A Literature Review." In. doi:10.2991/iceb-15.2015.55.

Taferner, Benjamin. 2017. "A NEXT GENERATION OF INNOVATION MODELS? AN INTEGRATION OF THE INNOVATION PROCESS MODEL BIG PICTURE." Review of Innovation and Competitiveness 3 (3): 14.

Tallman, Stephen, Mark Jenkins, Nick Henry, and Steven Pinch. 2004. "Knowledge, Clusters, and Competitive Advantage." The Academy of Management Review 29 (2): 258–71. doi:10.5465/AMR.2004.12736089.

Tang, Mingfeng, and Caroline Hussler. 2011. "Betting on Indigenous Innovation or Relying on FDI: The Chinese Strategy for Catching-up." Technology in Society 33 (1-2): 23–35. doi:10.1016/j.techsoc.2011.03.001.

"Top 100 Pharmaceutical & Biotech Companies (Global)." 2015. Statista.

"Top Pharma Companies by Rx Sales and R&D Spending 2016." 2017. Statista.

Torres-Reyna, Oscar. 2007. "Panel Data Analysis Fixed and Random Effects Using Stata (V. 4.2)." Princeton University.

"Trade Map - List of Exporters for the Selected Product in 2016 (Pharmaceutical Products)." 2017. Trade Map. http://www.trademap.org/Country_SelProduct.aspx?nvpm=1|||||30|||2|1|1|2|1|1|2|1|1.

"Trade Map - List of Importers for the Selected Product in 2016 (Pharmaceutical Products)." 2017. Trade Map. http://www.trademap.org/Country_SelProduct.aspx?nvpm=1|||||30|||2|1|1|1|1|1|2|1|1.

"Trade Map - List of Supplying Markets for the Product Imported by China in 2016." 2017. Trade Map. http://www.trademap.org/Country_SelProductCountry.aspx?nvpm=1%7C156||||30|||2|1|1|1|1|1|2|1|1.

Tseng, Yuen-Hsien, Yu-I Lin, and Chi-Jen Lin. 2007. "Text Mining Techniques for Patent Analysis." Information Processing and Management 43 (5): 1216–47. doi:10.1016/j.ipm.2006.11.011.

Turk, S. 2016. "Industry Report 32541aCA Brand Name Pharmaceutical Manufacturing in Canada." IBIS World.

Turkina, Ekaterina, and Ari van Assche. 2018. "Global Connectedness and Local Innovation in Industrial Clusters." Working Paper CIRANO, 55.

"U.S. Retail Sales of Homeopathic and Herbal Remedies 2011-2017." 2018. Statista. https://proxy2.hec.ca:2554/statistics/466508/us-retail-sales-of-homeopathic-and-herbal-remedies/.

Un, C. Annique. 2011. “The Advantage of Foreignness in Innovation.” Strategic Management Journal 32 (11): 1232–42. doi:10.1002/smj.927.

“UNESCO Science Report: Towards 2030.” 2016. UNESCO.

University of Arizona. 2000. “Antibody Structure.” The Biology Project - Immunology. http://www.biology.arizona.edu/immunology/tutorials/antibody/structure.html.

Utterback, James M, Thomas J Allen, J Herbert Hollomon, and Marvin A Sirbu. 1976. “The Process of Innovation in Five Industries in Europe and Japan.” IEEE TRANSACTIONS ON ENGINEERING MANAGEMENT, 7.

Uzzi, Brian. 1997. “Social Structure and Competition in Interfirm Networks: The Paradox of Embeddedness.” Administrative Science Quarterly 42 (1): 35–67.

van Eck, Nees, Ludo Waltman, Jan van den Berg, and Uzay Kaymak. 2006. “Visualizing the Computational Intelligence Field [Application Notes].” IEEE Computational Intelligence Magazine 1 (4): 6–10. doi:10.1109/MCI.2006.329702.

Vanhaverbeke, Wim. 2006. “The Interorganizational Context of Open Innovation.” In Open Innovation: Researching a New Paradigm. Oxford University Press.

Vernon, Raymond. 1966. “International Investment and International Trade in the Product Cycle.” The Quarterly Journal of Economics 80 (2): 190. doi:10.2307/1880689.

Veugelers, Reinhilde. 2017. “The Challenge of China’s Rise as a Science and Technology Powerhouse.” Policy Contribution, no. 19 (July).

von Hippel, Eric, and Georg von Krogh. 2006. “Free Revealing and the Private-Collective Model for Innovation Incentives.” R & D Management 36 (3): 295–306. doi:10.1111/j.1467-9310.2006.00435.x.

von Zedtwitz, Maximilian, and Oliver Gassmann. 2002. “Market Versus Technology Drive in R&D Internationalization: Four Different Patterns of Managing Research and Development.” Research Policy 31 (4): 569–88. doi:10.1016/S0048-7333(01)00125-1.

Waltman, Ludo, and Nees Jan van Eck. 2013. “A Smart Local Moving Algorithm for Large-Scale Modularity-Based Community Detection.” The European Physical Journal B 86 (11). doi:10.1140/epjb/e2013-40829-0.

Wang, Yuandi, Wim Vanhaverbeke, and Nadine Roijakkers. 2012. “Exploring the Impact of Open Innovation on National Systems of Innovation Theoretical Analysis.” Technological Forecasting & Social Change 79 (3): 419–28. doi:10.1016/j.techfore.2011.08.009.

Warin, Thierry. 2018. “Mondo International.” Mondo International. http://mondointl.cirano.qc.ca/.

West, Joel. 2007. “The Economic Realities of Open Standards: Black, White and Many Shades of Gray.” In Standards and Public Policy, pp. 87–122. Cambridge: Cambridge University Press.

West, Joel, and Scott Gallagher. 2006. “Patterns of Open Innovation in Open Source Software.” In Open Innovation: Researching a New Paradigm. Oxford University Press.

Wickham, H. 2017. “Package ‘Tidyverse’.”

Wickham, Hadley. 2014. “Tidy Data.” Journal of Statistical Software 59 (1): 1–23. doi:10.18637/jss.v059.i10.

Williamson, Oliver E., and Scott E. Masten. 1999. The Economics of Transaction Costs. Book, Whole. Cheltenham: E. Elgar Pub. Ltd.

Wooldridge, Jeffrey. 2000. Introductory Econometrics - A Modern Approach. South-Western College Publishing.

“World Intellectual Property Indicators 2017.” 2017. Geneva: World Intellectual Property Organization.

Xiwei, Zhong, and Yang Xiangdong. 2007. “Science and Technology Policy Reform and Its Impact on China’s National Innovation System.” Technology in Society 29 (3): 317–25. doi:10.1016/j.techsoc.2007.04.008.

Ying, Ying, Yang Liu, and Cong Cheng. 2016. “R&D Activities Dispersion and Innovation: Implications for Firms in China.” Asian Journal of Technology Innovation 24 (3): 361–77. doi:10.1080/19761597.2016.1265457.

Yueh, Linda. 2009. “Patent Laws and Innovation in China.” International Review of Law and Economics 29 (4): 304–13. doi:10.1016/j.irle.2009.06.001.

Zaheer, Srilata. 1995. “Overcoming the Liability of Foreignness.” The Academy of Management Journal 38 (2): 341–63. doi:10.2307/256683.

Zedtwitz, Max, Simone Corsi, Peder Veng Søberg, and Romeo Frega. 2015. “A Typology of Reverse Innovation.” Journal of Product Innovation Management 32 (1): 12–28. doi:10.1111/jpim.12181.

Zeschky, MB, S. Winterhalter, and O. Gassmann. 2014. “From Cost to Frugal and Reverse Innovation: Mapping the Field and Implications for Global Competitiveness.” RESEARCH-TECHNOLOGY MANAGEMENT 57 (4): 20–27. doi:10.5437/08956308X5704235.

Zhang, Fangning, and Josie Zhou. 2017. “What’s Next for Pharma Innovation in China.” McKinsey.

Zhang, Y. Philip, and Michelle M. Deng. 2008. “Enforcing Pharmaceutical and Biotech Patent Rights in China.” Nature Biotechnology 26 (11): 1235–40. doi:10.1038/nbt1108-1235.

Zhang, Yingying, and Yu Zhou. 2015. The Source of Innovation in China: Highly Innovative Systems. Book, Whole. Basingstoke: Palgrave Macmillan.

Zucker, Lynne G., Michael R. Darby, and Marilynn B. Brewer. 1998. “Intellectual Human Capital and the Birth of U.S. Biotechnology Enterprises.” The American Economic Review 88 (1): 290–306.