Leveraging Machine Learning to Investigate the Impact of NSERC Funding Programs on Research Outcomes

Hamid Vosoughi

A thesis

in

The Department

of

Concordia Institute for Information Systems Engineering

Presented in Partial Fulfillment of the Requirements

For the Degree of Master of Applied Science (Quality Systems Engineering)

Concordia University

Montréal, Québec, Canada

December 2023

CONCORDIA UNIVERSITY
School of Graduate Studies

This is to certify that the thesis prepared

By: Hamid Vosoughi

Entitled: Leveraging Machine Learning to Investigate the Impact of NSERC Funding Programs on Research Outcomes

and submitted in partial fulfillment of the requirements for the degree of

Master of Applied Science (Quality Systems Engineering)

complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the final examining committee:

_____ Chair
Dr. Y. Zeng

_____ Examiner
Dr. Y. Zeng

_____ Examiner
Dr. C. Wang

_____ Supervisor
Dr. A. Schiffauerova

_____ Co-supervisor
Dr. A. Ebadi

Approved by _____
Dr. C. Wang, Director
Concordia Institute for Information Systems Engineering

_____2024/01/09 _____

Dr. M. Debbabi, Dean
Gina Cody School of Engineering and Computer Science

Abstract

Leveraging Machine Learning to Investigate the Impact of NSERC Funding Programs on Research Outcomes

Hamid Vosoughi

This research examines the impact of various funding programs by NSERC on research outcomes. We utilize statistical models and machine learning algorithms trained on the integrated database of researchers' publications and funding to determine the efficacy of NSERC funding programs. We aim to evaluate the effectiveness of different strategies defined by NSERC through funding programs and analyze the impact of various factors. We seek to enhance our understanding, with the aspiration that it will inform the design of more effective programs in the future.

We compare the results of linear regression, random forest, and neural networks. Then, we perform SHAP analysis to identify the most important features within funding programs. We aim to gain insights into the impact of receiving funding through different programs on research outcomes.

We observed that random forest model outperformed the other models for all dependent variables, i.e., future productivity, quality of the publication, and future co-authorships. Subsequently, we examined the significance of independent variables in predicting dependent variables across the funding programs.

For Canada Research Chairs recipients, the impact of their prior work holds greater importance in shaping research outcomes, underscoring a distinctive emphasis on research excellence within this program. In contrast, the impact of career age is lower compared to other programs. Interestingly, within the Discovery Grants program, career age becomes notably influential in predicting future productivity in favor of young researchers. Furthermore, we found an intriguing exception for researchers with a history of large group collaborations within Discovery Grants, where some experience a negative impact on future collaborations. The award amount plays a more important role in shaping the research outcomes of recipients engaged in strategic projects.

Our findings emphasize the importance of allocating funding programs to researchers whose qualifications are aligned with the programs' objectives.

Acknowledgement

I want to convey my profound gratitude to my supervisors, Dr. Schiffauerova and Dr. Ebadi, whose unwavering support and constructive guidance have been indispensable in bringing this research to fruition. Studying under their supervision and contributing to their team has provided me with invaluable insights and growth opportunities.

My deepest appreciation also goes to my family, including my brother Saeid and sister Bahareh, as well as my parents. Their steadfast support has been the bedrock of my journey, and I am truly grateful for their unwavering encouragement and love.

Table of Contents

List of Figures

List of Tables

1. Introduction

Research funding is defined by Huang and Huang (2018) as the financial support granted by relevant funding agencies to researchers following the submission and approval of research proposals. It plays a crucial role in shaping the scientific output of researchers by providing them with the necessary financial resources to pursue their research. The impact of funding on scientific output has been widely investigated in the literature, focusing on its effect on the quantity and quality of publications, as well as its influence on collaboration among researchers and its potential to enhance scientific output.

Based on the literature, most of previous works concluded that funding plays a significant role in enhancing the productivity of researchers in terms of the number of publications (Payne & Siow, 2003; Godin, 2003; Ebadi & Schiffauerova, 2016), and the quality of these publications while considering either their citation counts (Gök, et al., 2016; Ebadi & Schiffauerova, 2016; Alvarez-Bornstein & Bordons, 2021) or the journal's impact factors (Lewison & Dawson, 1998; Wang & Shapira, 2015; Alvarez-Bornstein & Bordons, 2021), and also in enhancing the collaboration among researchers mainly measured through number of co-authors in the publications (Alvarez-Bornstein & Bordons, 2021; Zhao, 2010; Ebadi & Schiffauerova, 2015). However, most previous studies focused on the impact of funding and its level without considering the type of funding and specific objectives of each funding program. This is a research gap which we aim to address in this research. The main objective is to provide a more comprehensive understanding of the impact of different funding programs on scientific output.

To the best of our knowledge, only one study has explored the budget allocation among various funding programs and fields within NSERC (Veletanlic & Sa, 2020). This study aimed to comprehend the objectives of the NSERC organization and its strategy to influence scientific behavior by channeling more funds toward certain fields and programs. However, this research did not examine the effect of different funding programs on scientific output.

Furthermore, most of the studies that have explored the relationship between funding and scientific output have focused on a small period, a specific field, or used a small dataset. For instance, some studies have examined the effect of funding on scientific output in a specific field, while others

have looked at the impact of funding in a particular year. This approach has limited the findings' applicability and made it difficult to draw robust conclusions about the relationship between funding and scientific output. To overcome these limitations, our research uses a large dataset that covers all the fields supported by NSERC in natural science and engineering for the period of 1992 to 2018. By using this comprehensive dataset, we aim to provide a more accurate, representative and comprehensive picture of the relationship between funding and scientific output.

Finally, previous studies have relied mainly on simple regression, bibliometric, and statistical analysis to analyze the data. In contrast, our research proposes employing more sophisticated analysis techniques by comparing different machine learning models with conventional linear regressions to understand which is better suited for our dataset. This approach will allow us to examine the relationship between funding and scientific output in greater detail that may not be detectable through simple statistical analysis.

We will focus on the main funding programs offered by NSERC. Our main objective is to conduct a thorough examination of the effectiveness of the different strategies outlined by NSERC within its funding programs. This involves carefully studying how various factors affect the overall effectiveness of NSERC programs. By gaining deeper insights through this research, we aspire not only to contribute to our current understanding but also to provide valuable insights that can guide the design and implementation of more effective NSERC programs in the future.

Our study covers all the fields supported by NSERC in natural science and engineering for the period of 1992 to 2018. For each of the selected funding programs, we aim to evaluate the importance of input features on 3 different dependent variables, i.e., future productivity, quality of the publication by considering journal impact factor, and collaboration of the funded researchers. To do so, we compare the results of 5-fold cross-validated multiple linear regression with random forest and multilayer perceptron (MLP) neural network to choose our final model. This allows us to understand how these two machine learning models perform compared to the conventional regression model and choose the one that is better suited for our dataset. After deciding about the model, we analyze the importance of input features for predicting each dependent variable using SHAP (SHapley Additive exPlanations) analysis to compare their impact within different funding programs.

The remainder of this thesis is organized as follows: In the next chapter, we will cover the literature about the impact of funding on the productivity, impact, and collaboration of researchers. Then, we will discuss the data and methodology. Next, we will discuss the main findings of our study. Lastly, we will have the conclusion and present our limitations and future work.

2. Literature Review

Funding is considered one of the most important factors affecting research productivity. Indeed, more funding could increase the number of publications and enhance the quality of the published papers (Alvarez-Bornstein & Bordons, 2021; Ebadi & Schiffauerova, 2016). Moreover, funding facilitates collaboration among researchers thereby enabling them to produce more and higher quality papers (Ebadi & Schiffauerova, 2015). As a result, evaluating the effects of funding on scientific output and finding better ways to the allocation of funding is crucial. Hence, in this section, we will discuss the findings of previous studies that have examined the impact of funding. First, we will present different types for funding categorization. This would help us to better understand how funding is distributed among researchers and why different types of funding exist. Subsequently, we will present the studies that investigated the impact of funding on the productivity of recipients of funding, quality of their publications, and collaboration among researchers. These works would highlight the importance of research funding in influencing research output and subsequently enable us to assess the efficiency of resource allocation.

2.1 Categories of funding

In the literature, there are various ways for categorizing different types of funding. Studying funding categories can help us to better realize the need for providing different types of funding for researchers. Guena and Martin (2003) categorized funding based on the mechanism that has been used for the allocation of funds. They discussed that different countries allocate funding in 3 major ways - i.e., performance-based approach, educational size-based approach, or a combination of these two methods. Performance-based funding is allocating funding based on the performance of the institutions, while the educational size approach allocates funds considering the size of institutions. They claimed that a performance-based approach may lead to more responsibility for the universities, and it provides a strategy to logically shift funding from ineffective areas to those where they may be used more effectively. However, it is expensive to collect accurate and comparable information. Moreover, since the performance-based approach promotes competition,

it may discourage the employment of innovative methods regardless of their positive effects on society.

In addition to the method for the allocation of funding, another important element for funding classification is the source of financial resources allocated to researchers. Muscio et al. (2013) distinguished between private and public fundings based on the need that the research would satisfy. They considered the funding as public if the research work is responding to the public interest. On the other hand, funding that would lead to research that can be sold on the market and is a response to the need of a specific organization is defined as private funding. They concluded that there is a positive relationship between public and private funding indicating that they should not be seen as substitutes but rather as complementing sources of support. Therefore, reducing public funding to universities would weaken the university-industry collaboration and limit their potential to raise funding from outside sources.

In a more recent study, Veletanlic and Sa (2020) considered the purpose of providing funding for its classification. However, they took a different approach by focusing on the intended research objectives that the researcher is supposed to pursue rather than just labeling funding as public or private. Hence, they classified funding leveraging on the delegation mode that has been defined. There are 4 main delegation modes. First, blind delegation mode gives scientists control over the financial resources to pursue academic-oriented projects. Second, the network delegation mode aims to create a network of scientists, firms, and end-users that can follow their own objectives and projects. The third one is the incentive delegation mode that is generally allocating funding to pursue political objectives. Finally, the steady-state delegation mode encourages scientists to focus on innovative research areas that are the priorities of the state.

Particularly, Veletanlic and Sa (2020) investigated how the Natural Sciences and Engineering Research Council (NSERC), which is one of the largest Canadian funding organizations, changed its resource allocation between 1991 to 2016. They have found that the annual R&D budget of this agency has increased during the research period from 700 million CAD in 1991 to more than 1 billion CAD in 2016. Moreover, they indicated that from 2006–07 to 2013–14, there is a shift in allocating funding from blind delegation mode to targeted programs for university-industry partnerships and innovation which could be an effect of the federal government’s innovation agenda.

Recognizing the different types of funding is a crucial starting point, but it is also important to understand how funding could affect the research outcomes. Analyzing the impacts of funding could provide valuable insights for policy makers to evaluate the effectiveness of funding and better allocate the resources. Hence, in the following sections, we will summarize the previous studies about the impact of funding on the research outcomes in terms of the productivity of researchers, the quality of their publications, and collaboration among themselves.

2.2 Effects of funding on quality and quantity of publication

Many studies have explored the relationship between funding and scientific output. Lewison and Dawson (1998) argue that receiving funding can positively impact research outcomes because it not only provides financial support for researchers to pursue their work but also means that the research has undergone one or more screening processes. The effect of funding on the productivity of researchers is one of the aspects that has attracted the attention of researchers. It is common in the literature to compare the number of publications to understand how funding affects research output in terms of their productivity (Godin, 2003; Ebadi & Schiffauerova, 2013; Boyack & Börner, 2003; Campbell, et al., 2010).

Most of the researchers found a positive impact of the research funding on the publication quantity (Payne & Siow, 2003; Godin, 2003; Ebadi & Schiffauerova, 2016; Tahmooresnejad, et al., 2015). For example, Payne & Siow (2003) investigated the impact of federal research funding on the research outcomes of 68 universities. They found that an increase of $1 million (1996$) in funding resulted in 10 more articles and 0.2 more patents. In another study, Godin (2003) focused on NSERC as a major federal funding agency in Canada to examine the impact of funding on the research output. They concluded that the positive impact of funding is more pronounced when the amount is above the median. In a similar study on the NSERC funding agency, Ebadi & Schiffauerova (2016) confirmed that funding and productivity are positively correlated; however, their results suggest that the impact of funding is not the same for all the funding programs. Indeed, targeted programs would have a higher impact on the number of publications.

Using a different approach, Ebadi & Schiffauerova (2015) and Tahmooresnejad, et al. (2015) studied the relation between funding and productivity from another perspective. They studied how the past productivity of researchers could affect the amount of funding granted to researchers in the future. Ebadi & Schiffauerova (2015) found a significant positive impact of researchers' past

productivity to secure more funding in the future. This finding is only partially confirmed by Thamooresnejad, et al. (2015) who compared Canadian and American researchers in the field of nanotechnology. They found only a positive effect of past productivity on future funding for American researchers. In contrast, their results suggest that the past productivity of Canadian researchers does not affect their future grants significantly. They claim that this result could be the consequence of focusing on nanotechnology, which was a relatively young research area in Canada at the time. In fact, by considering other research fields, they might have obtained different results.

In addition to the quantity, the impact of funding on the quality of publications has also been investigated. Wang & Shapira (2015) claim that it is important to consider where the funding is coming from since they found that publications that are funded by the EU, the US, and Germany are more likely to be of higher quality. Moreover, Gök et al. (2016) found that the impact of private funding on the quality is often stronger than public funding. In another study, Alvarez-Bornstein & Bordons (2021) claim that even the field of study is important since the impact of funding on the quality for different fields of study is not the same.

Although measuring the quality is not as straightforward as measuring the quantity, the most common way is to consider the number of citations received by the funded papers (Payne & Siow, 2003; Zhao, 2010; Ebadi & Schiffauerova, 2015; Wang & Shapira, 2015; Tahmooresnejad, et al., 2015; Gök, et al., 2016; Yan, et al., 2018; Veletanlic & Sa, 2020). Most studies found a positive impact of funding on the citation counts (Zhao, 2010; Gök, et al., 2016; Alvarez-Bornstein & Bordons, 2021; Ebadi & Schiffauerova, 2016). However, some research only partially confirms this impact. For example, Wang & Shaipra (2015) found a positive impact of funding on citation counts, but they suggest that the number of sources in the acknowledgment of publications has a concave impact on the number of citations, growing up to the ideal number of sources, and after that point, it declines. In another study, Yan, et al. (2018) investigated the impact of funding in seven disciplines such as science, engineering, and medicine, and found that the number of citations for funded articles is higher, except in the field of nanotechnology. This is partially aligned with the research of Tahmooresnejad, et al. (2015) who compared the impact of funding on the outcomes of Canadian and American researchers. They concluded that an increase in the amount of grants has a significantly positive effect only on the research quality of American researchers, while it does not affect the research quality of Canadian researchers. This indicates

that it is also important to consider other factors other than the field of study. Ebadi & Schiffauerova (2016) also showed in their study that the impact of funding on the quality is positive but factors like career age or collaboration of researchers can also influence the quality of their work.

In contrast to these studies, Payne & Siow (2003), who compared the research outcomes of 68 universities by considering the amount of federal funding that each university received, found that an increase in federal funding would lead to more articles but not necessarily of higher quality. The different results they obtained may be attributed to their distinct methodology. They evaluated the collective performance of all researchers within a university, in contrast to an individual researcher's performance.

While many studies considered the citation counts as measure of quality, there are some issues related to this method. First, a considerable percentage of articles will not receive a single citation in a five-year window after their publication which makes it difficult to measure their quality. Moreover, using this measure, we are not able to distinguish between negative and positive citations. The other problem is related to self-citations when researchers refer to their previous works. Last but not least, there is a bias toward English-language publications. They are more likely to receive citations compared to non-English articles (Okubo, 1997). Due to the limitations of using citation counts as the measure of quality, some researchers have measured the effect of funding on the quality by comparing the prestige of the journals in which the research has been published. The prestige of the journals has been commonly represented by considering the impact factor of these journals (Ebadi & Schiffauerova, 2013).

Most researchers who used the journal impact factor as the quality measure found that funding can help researchers publish their publications in a higher-quality journal (Lewison & Dawson, 1998; Wang & Shapira, 2015; Alvarez-Bornstein & Bordons, 2021). For example, Lewison & Dawson (1998) in their study on the effect of funding on the journal impact factor of the publications found that the quality of the research output is significantly influenced by the number of funding sources. The positive impact of the number of funding bodies increases even for six or more sources but it is more significant when it increases from zero to one, highlighting the significant impact of receiving funding on research outcomes. However, Alvarez-Bornstein & Bordons (2021) concluded that the positive impact of funding on the journal impact factor may vary for each

discipline. In contrast, Godin (2003) who focused on the NSERC-funded researchers found no impact of funding on the research quality of Canadian researchers. Focusing on one agency (NSERC) for their analysis might explain why their results differ from others.

In a distinct attempt, Ebadi & Schiffauerova (2015) investigated the relationship between funding and quality from a different perspective. By considering NSERC-funded researchers, they studied how quality of previous works can affect the potential of a researcher to receive funding in future. For measuring the quality, they considered both citation counts and journal impact factor. Their results suggest that researchers whose previous studies were of higher quality in terms of citations and journal impact factor have a higher probability of securing more funding in the future.

The summary of the papers that studied the impact of funding on the quantity is shown in Table 1 while Table 2 summarizes the ones investigating the quality of publications.

Data	Findings
Articles from 68 US universities for the period of 1981 to 1998	**Positive** impact of funding on the quantity
NSERC funded researchers from 1990 to 1999	**Positive** impact of funding on the quantity
NSERC funded researchers within the period of 1996 to 2010 and their articles	**Positive** impact of funding on the quantity **Positive** impact of collaboration on the quantity
Publication of 3,684 Canadian scientists & 33,655 US scientists: 1996-2005 from Scopus	**Positive** impact of funding on the quantity in US & Canada **Positive** impact of collaboration on the quantity
NSERC funded researchers within the period of 1996 to 2010 and their articles	**Positive** impact of funding on the quantity **Positive** impact of past productivity of a funded researcher on the quantity **Positive** impact of collaboration on the quantity

Table 1 Summary of the research on quantity of research output

Authors	Year	Findings
Payne & Siow	2003	**Positive** impact of funding on the quality
Godin	2003	**No** impact of funding on the quality
Ebadi & Schiffauerova	2015	**No** impact of funding on the quality **Positive** impact of funding on the quality **Positive** impact of funding on the quality **Positive** impact of number of authors on the quality **Positive** influence on the journal impact factor
Tahmooresnejad, et al.	2015	**Positive** impact of funding on the quality in US **No** impact of funding on the quality in Canada **Positive** impact of collaboration on the quality
Ebadi & Schiffauerova	2016	**Positive** impact of past productivity of a funded researcher on the quality **Negative** impact of career age on the quality **Positive** impact of collaboration on the quality **Positive** impact of funding on the quality **Positive** impact of funding on the quality **Positive** impact of collaboration on the quality **Positive** impact of funding on the quality

Authors	Year	Data
Lewison & Dawson	1998	12,925 papers published from 1988 to 1994 in biomedical field
Godin	2003	NSERC funded researchers from 1990 to 1999
Payne & Siow	2003	Articles from 68 US universities for the period of 1981 to 1998
Zhao	2010	72 funded papers and 194 normal papers in 1998 in the field of Library and Information Science (LIS)
Ebadi & Schiffauerova	2015	NSERC funded researchers within the period 1996 to 2010 and their articles
Wang & Shapira	2015	89,605 nanotechnology publications from August 2008 – July 2009 from WoS
Tahmooresnejad, et al.	2015	Publication and authorship data of 3,684 Canadian scientists & 33,655 US scientists: 1996-2005 from Scopus
Ebadi & Schiffauerova	2016	NSERC funded researchers within the period 1996 to 2010 and their articles
Gok, et al.	2016	Articles published from January 2009 to December 2011 from six different countries
Yan, et al.	2018	Articles and review articles of five journals for 7 disciplines: 2010-2016
Álvarez-Bornstein & Bordons	2021	Spanish researchers publications during 2010-2014 from WoS

Table 2 summary of the research on the quality of research output

2.3 The relationship between funding and collaboration

According to Katz & Martin (1997), scientific collaboration could be defined as the act of cooperation among researchers with a similar objective to develop new scientific findings. The accepted metric in the literature for assessing collaboration is co-authorship because it is perceived as a reliable indicator of cooperative scientific research (De Solla Price, 1963; Ubfal & Maffioli, 2011; Ebadi & Schiffauerova, 2013). However, co-authorship is not the perfect indicator for measuring collaboration since collaborating with another researcher does not always lead to a joint paper. A scenario where two researchers collaborate on a study but ultimately decide to publish their findings independently may serve as an example (Tijssen, 2004; Katz & Martin, 1997; Ebadi

& Schiffauerova, 2013). Moreover, the co-authorship does not always mean that the researchers collaborated as the co-author could be added for various reasons. You can add a co-author for various reason Some researchers have also considered the position of a researcher in a collaboration network as an important factor. Hence, they considered network variables to indicate the collaboration of researchers (Tahmooresnejad, et al., 2015; Ebadi & Schiffauerova, 2015).

As a result of modern science that has a more complicated and interdisciplinary spirit, researchers may be inclined to work together more (Ebadi & Schiffauerova, 2013). Also, Tahmooresnejad, et al. (2015) found a growing trend among researchers to form teams that have members from a variety of fields. In fact, scientific collaboration could lead to a more efficient use of resources because of economies of scale since it would combine various ideas and talents and develop existing skills (Ubfal & Maffioli, 2011; Ebadi & Schiffauerova, 2013). Therefore, it is important to study how the collaboration of researchers could affect their research outcomes.

The researchers mostly found that co-authorship can positively influence the quantity and quality of researchers' work (Zhao, 2010; Wang & Shapira, 2015; Ebadi & Schiffauerova, 2016; Tahmooresnejad, et al., 2015). Ebadi & Schiffauerova (2016) suggest as researchers are involved in larger teams, they become able to allocate the workload among members more effectively and have access to required resources like funding and manpower. Hence, they would be able to work on different projects. Yan, et al. (2018) who compared the funded and non-funded research in seven fields concluded that co-authorship has a significant positive impact on the quality of papers in terms of citation counts except in the field of mathematics where co-authorship leads to more citations only for funded researchers. This highlights the importance of considering other factors like funding and field of study to better capture the impact of collaboration. Furthermore, Tahmooresnejad, et al. (2015) showed that the role of a researcher in a collaboration network is also an important factor influencing the quality of their work. They claimed that being in a better position within co-publication networks has a significant positive impact on the number of their papers and their quality.

Beyond the collaboration itself, some researchers have considered the type of collaboration as another potential factor impacting the quality of research outcomes. Wang & Shapira (2015) explored the impact of collaboration on the research quality, considering the number of affiliations and author countries as well. Their findings suggest that co-authorship positively impacts the

quality. However, the positive impact of author countries was observed only for two countries; beyond that, quality tended to decline. On the other hand, the impact of affiliations followed a U-shaped curve, initially decreasing and then increasing after a certain point. Yan, et al. (2018) took a different approach, examining the impact of institutions as a binary variable, distinguishing between single-institutional and multi-institutional co-authorship. They found that multi-institutional collaboration positively influenced research quality. However, the study did not consider the specific number of institutions involved in the collaboration.

Some researchers also investigated the impact of funding on shaping collaborations among researchers and found a positive impact of funding on co-authorships of publication (Zhao, 2010; Ebadi & Schiffauerova, 2015; Alvarez-Bornstein & Bordons, 2021). In fact, funding agencies may encourage researchers to collaborate more, aiming to enhance their outcomes (Alvarez-Bornstein & Bordons, 2021). For example, Zhao (2010) compared funded and non-funded research in the field of library and information science. They found that funded research, on average, has a higher-level collaboration since the average number of authors per article for funded research is 29% higher compared to non-funded papers. Ebadi & Schiffauerova (2015) discovered that while funding positively influences research collaboration, other factors such as past productivity and quality also play a role in shaping co-authorship. They found that researchers who produced more high-quality papers tend to form larger teams in the future. However, in a distinct study, Ebadi & Shiffauerova (2015) showed that the relationship between funding and collaboration is not unidirectional. They found that collaborating with other researchers could increase the likelihood of securing more funding in the future for researchers.

The summary of the respective papers is shown in Table 3.

Table 3 Summary of research on the collaboration of researchers

Authors	Year	Data	Findings
Zhao	2010	72 funded papers and 194 normal papers in 1998 in the field of Library and Information Science (LIS)	**Positive** impact of funding on the collaboration
Ebadi & Schiffauerova	2015	NSERC funded researchers	**Positive** impact of funding on the collaboration **Negative** impact of career age on the collaboration
Ebadi & Schiffauerova	2015	NSERC funded researchers within the period of 1996 to 2010 and their articles	**Positive** impact of funding on the collaboration **Positive** impact of collaboration on the quality & quantity
Wang & Shapira	2015	89,605 nanotechnology publications during August 2008 – July 2009 from WoS	**Positive** impact of collaboration on the quality
Tahmooresnejad, et al.	2015	Publication and authorship data of 3,684 Canadian scientists & 33,655 US scientists: 1996-2005 from Scopus	**Positive** impact of collaboration on the quality & quantity
Ebadi & Schiffauerova	2016	NSERC funded researchers within the period of 1996 to 2010 and their articles	**Positive** impact of collaboration on the quality & quantity
Yan, et al.	2018	Articles and review articles of five journals for 7 disciplines: 2010-2016	**Positive** impact of collaboration on the quality
Álvarez-Bornstein & Bordons	2021	Spanish researchers' publications during 2010-2014 form WoS	**Positive** impact of funding on the collaboration

As discussed in this section, previous studies explored the impact of funding on scientific output, focusing on its influence on quality, productivity, and collaboration of researchers. Despite valuable insights gained from these studies, a notable gap exists in understanding how different funding programs affect research outcomes. Hence, in this study, we aim to contribute to the existing literature by offering a more comprehensive understanding of the relationship between various funding programs and scientific output. To do so, we decided to use a dataset covering all NSERC-supported fields in natural science and engineering from 1992 to 2018, aiming for a more accurate analysis of the relationship between funding and scientific output. Additionally, our research proposes advanced analysis techniques, including comparing machine learning models with linear regression, to provide a more detailed examination of the relationship between funding and scientific output. In the next section, we will discuss the dataset used for our analysis.

3. Data Collection

As we aim to discover the impact of different funding programs on the scientific outcomes, we decided to focus on NSERC-funded researchers. NSERC is the primary federal funding agency in Canada and supports researchers by providing different funding programs with different objectives. We had several steps for the data collection and preprocessing phase. Three different data sources have been used. First, we have collected the funded researchers in natural science and engineering who were supported by NSERC within the period from 1992 to 2018. The funding dataset includes (but is not limited to) metadata such as the name of the researcher who received funding, funding program, year of the financial support and duration, researcher's affiliation, and amount of funding.

The NSERC dataset used in this research consists of 556,427 records representing the various funding programs awarded by NSERC from 1992 to 2018. It is important to note that individual researchers could have received funding under different programs and in multiple installments. For example, a researcher receiving a Discovery Grant of $100,000 in total over 5 years will have in our database 5 entries of $20,000 each year. The NSERC funding database includes both funding for researchers (grants) and funding for students (scholarships). For the purpose of this research, data related to scholarships were excluded, resulting in a dataset size of 337,329 data points. After removing the scholarships from the original dataset, the number of researchers who received funding from NSERC decreased to 29,394 from 147,327. This reduction is due to the fact that

scholarships comprise a large portion of the funding programs offered by NSERC, and their removal has allowed us to focus on our objective – the impact of research funding. It is also worth noting that the dataset includes only the funding programs offered by NSERC and does not encompass all the funding received by researchers during the study period.

In the next step of the data collection for investigating the impact of NSERC funding on research outcomes, we also collected data on the publications produced by researchers who received NSERC funding during the aforementioned period. To do this, we used Elsevier's Scopus database, which we selected due to its reputation for completeness and accuracy in comparison to other citation databases like Google Scholar and Web of Science (Tahmooresnejad, et al., 2015). Since we had limited access to the database, we had to collect the publications of each researcher separately, which was a time-consuming process. Overall, we collected 2,434,442 publications from researchers who had received funding from NSERC and had publications in Scopus. The publications database contains a range of metadata, including but not limited to the researcher's name, the title of the publication, publication type, publication date, journal name, number of citations, author affiliations, and funding information.

To make our analysis more comparable, we decided to focus specifically on journal articles published by these researchers and removed other publication types such as conference papers or book chapters. This resulted in a final dataset of 1,775,103 articles. It is worth noting that this dataset only includes publications in Scopus and may not include all publications produced by NSERC-funded researchers during the study period.

Third, in addition to collecting data on funding programs and publications, we also obtained information on the journal impact factor from SCImago. We considered two key indicators of journal impact: the Scimago Journal Rank (SJR) indicator and the h-index indicator. The SJR indicator evaluates the prestige of a journal, in a given year, by measuring the average number of weighted citations received by the documents published over the past three years in that journal. The h-index, on the other hand, provides a measure of both the productivity and impact of a journal by considering the number of articles published by the journal and the number of citations received by those articles (SCImago, 2023). We collected information on the name of the journal, its SJR indicator, and its h-index from the SCImago database, and used this data to gain insights into the quality and impact of the journals in which NSERC-funded researchers published their articles.

The process of merging the three datasets into a unified dataset involved combining the NSERC dataset and Scopus dataset based on the names of the researchers. The resulting dataset was linked to the SCImago dataset by matching the name of the journal.

However, when we examined the NSERC dataset, we found that the researchers' names were not always consistent across different instances of funding. For instance, one researcher named Andrew Michael Jones might be listed as Andrew Jones in one instance, AndrewM Jones in another, and AndrewMichael Jones in a third instance. Furthermore, in some cases, a researcher's name might include accents in one instance but not in another. To address these issues, we decided to remove all accents from the researchers' names. This ensured consistency across the dataset and eliminated discrepancies that might arise from different spellings of the same name. Additionally, we adopted the format used in the Scopus dataset, which lists authors' names with their full last names and the first character of their first name. To ensure consistency across the dataset, we adjusted the names of NSERC-funded researchers to match this format. However, to reduce the potential errors, for matching the names of the researchers, we have also ensured that the affiliation of the researcher is the same for the matched names.

After performing the pre-processing steps on the NSERC dataset, we used fuzzy string matching to connect the names of researchers in the NSERC and Scopus datasets and to unify the data from these sources. Fuzzy string matching is a technique that allows for approximate string matching by measuring the similarity between two strings based on their characters, sequences, and lengths. This technique was particularly useful in cases where the spelling of a researcher's name varied slightly across different datasets, as it allowed us to identify and match researchers with similar but not identical names.

In addition to fuzzy string matching, we also used the unique Author ID in the Scopus dataset. The author ID is assigned to each author by Scopus, and it allows for the tracking of a researcher's publications over time, even if they change their name or affiliation. By using this unique identifier, we were able to ensure that the publications recorded for a funded researcher belonged to them only and not to another researcher with a similar name. This step was necessary to avoid any potential errors or confounding factors in our analysis, and to ensure that our results were accurate and reliable.

In the final step of merging the datasets, we needed to connect the SCImago dataset with the NSERC and Scopus datasets. We achieved this by using a string-matching approach to match the journals in the Scopus dataset with those in the SCImago dataset. Fortunately, we found that the names of the journals in these two datasets were very similar and did not require any modification to match them accurately. By linking the datasets in this way, we were able to combine the data from different sources and create a more comprehensive and complete dataset for our analysis.

As the next step, we will present NSERC funding programs and then provide an initial descriptive analysis of our dataset.

4. NSERC Funding Programs

The Natural Sciences and Engineering Research Council of Canada (NSERC) is a prominent funding agency in Canada that was established in 1978 (NSERC, 2023). As Canada's primary federal funding agency in natural science and engineering, NSERC plays an important role in supporting and promoting research excellence in related fields. By providing financial support, resources, and opportunities to researchers at different stages of their careers, NSERC aims to push the boundaries of knowledge and contribute to the growth of research in Canada. Focusing on the improvement of collaboration, fostering the partnership between academia and industry, interdisciplinary research, and knowledge exchange, NSERC seeks impactful discoveries, breakthrough innovations, and helping skilled researchers (NSERC, 2023).

By introducing different types of grants, scholarships, and partnership programs, NSERC supports a diverse range of scientific disciplines and engineering fields. NSERC's funding programs span a wide spectrum of research areas, including physics, chemistry, biology, computer science, engineering, etc. By following multidisciplinary approaches and promoting collaboration, NSERC encourages researchers to address different problems and generate innovative solutions that have positive impacts on society. NSERC not only supports individual researchers and scholars but also actively promotes partnerships and collaborations among academia, industry, and government. By fostering knowledge exchange, technology transfer, and industry-academic collaborations, NSERC facilitates the translation of research outcomes into practical applications that are useful for Canada. Consequently, we have decided to focus on NSERC in this research and compare the effectiveness of its funding programs (NSERC, 2023).

In this study, the NSERC dataset includes funded researchers for the period of 1992 to 2018. By narrowing our scope to funding and grants, we identified a total of 189 distinct programs within the dataset through which researchers received financial support from NSERC during the specified period.

Given the vast number of programs introduced by NSERC, it was not feasible to consider all of them within the scope of our research. Thus, we made a deliberate decision to focus on a subset of programs that include different objectives of NSERC funds and are considered to be more important. By 'importance,' we refer to programs that have consistently provided funding to researchers over time and have supported a substantial number of researchers since some programs have allocated funds for only a brief period or supported a limited number of researchers.

To select the programs for our research, we defined specific criteria to ensure comparability among the programs. First, we required that a program should have allocated funds in at least 15 different years. This criterion was essential to select programs with a sustained presence and impact over time. Additionally, we considered the number of records in our dataset. Thus, we set a threshold of at least 5,000 records associated with a program to ensure comparability and an adequate sample size for our analysis.

Following these considerations, we identified five programs that met these criteria. However, among these programs, one stood out as being distinct in its focus and potential impact. The Research Tools and Instrument funding program specifically provides funds for acquiring tools and instruments necessary for research purposes. Given the program's specific focus on instruments and the varying requirements across different research areas, we found it less comparable to the other programs. Therefore, we excluded this program from our analysis. As a result, the selected programs that we chose for this study are as follows:

- **Discovery Grants Program – Individual:** One of the most recognized funding programs introduced by NSERC is the Discovery Grants Program – Individual. It is specifically developed to support individual researchers in the natural sciences and engineering disciplines. This program focuses on fostering creativity, innovation, and scientific excellence by providing long-term funding stability to researchers. Through this program, NSERC aims to support fundamental research that advances knowledge, addresses

research gaps, and contributes to the overall advancement of science and engineering in Canada (NSERC, 2023).

The Discovery Grants Program - Individual operates on a competitive peer review basis, where researchers submit proposals mentioning their research objectives, methodology, and expected outcomes. By considering different factors like the novelty of the research, the committees assess the scientific merit, potential impact, and feasibility of the proposed research projects. One important mission of the Discovery Grants Program – Individual is its concentration on supporting early-career researchers and enabling them to establish independent research. This program provides opportunities for emerging researchers to build their careers, develop their scientific expertise, and attract additional research funding in the future (NSERC, 2023).

By supporting individual researchers through the Discovery Grants program, NSERC plays a vital role in nurturing a dynamic and innovative research environment in Canada. The program contributes to the advancement of scientific knowledge, the training of highly skilled researchers, and the development of a strong research community that drives innovation and makes contributions to society.

- **Canada Research Chairs:** The Canada Research Chairs (CRC) Program is a prestigious funding opportunity that aims to attract and support outstanding researchers in Canada. The program focuses on enhancing Canada's research excellence by providing funds to researchers with exceptional potential. The program provides long-term funding for chairholders, allowing them to establish and lead research teams, collaborate with partners, and make contributions to their disciplines.

  This program has two types of funding opportunities. Tier 1 chairs are reserved for outstanding, internationally recognized researchers; while Tier 2 Chairs are for exceptional emerging researchers who have the potential to become leaders in their fields.

  Canada Research Chairs program has a competitive selection process. Institutions nominate researchers for positions, and the nominees undergo a comprehensive evaluation by an expert peer review committee. The evaluation assesses the candidate's research track record, the potential for leadership, and the quality and impact of their proposed research program (Chairs, 2023).

- **Collaborative Research and Development Grants:** This program is designed to promote collaboration between academic researchers and industry partners in Canada. It aims to facilitate knowledge transfer, foster innovation, and address research challenges that have practical applications and potential economic benefits. The primary objective of the Collaborative Research and Development Grants program is to support collaborative research projects that involve both academic researchers and industry partners. By bringing together the expertise and resources from academia and industry, the program encourages the development of innovative solutions to real-world problems and aims to translate scientific knowledge into practical applications. Therefore, it provides financial support to research projects that involve a strong collaboration between academic researchers and industry partners.

  The key feature of this program is its focus on industry relevance and commercialization potential. The research projects funded through this program are expected to have a clear path to the application, demonstrating potential economic and positive impacts on society. To apply for this program, academic researchers and their industry partners jointly develop and submit research proposals outlining the project objectives, methodology, expected outcomes, and the roles and contributions of each partner. These proposals undergo evaluation by expert reviewers and committees who assess the scientific merit, technical feasibility, and potential socio-economic impacts of the proposed research projects. (NSERC, 2022)

- **Strategic Projects – Group:** The Strategic Project program in NSERC aims to support research that combines fundamental and applied aspects, with a particular interest in achieving practical applications beyond the university setting. The program's goal is to enhance Canada's economy, society, and/or environment within the next 10 years by increasing research and training in targeted areas.

  The success of a Strategic Project is based on the appropriate skill sets and expertise among the researchers involved. Therefore, it is crucial to clearly describe the roles and time commitments of each research co-applicant and collaborator in the proposal. Additionally,

well-defined collaboration and communication plans should be established to ensure effective coordination and cooperation among the team members.

Candidates should select one target area that aligns with the program's objectives. Only applications whose objectives are aligned with the target areas of interest would be considered. In exceptional cases where a proposal fits the context of the target area but falls outside its specific research topics, researchers can provide a convincing case for consideration as an "Exceptional Opportunity" (NSERC, 2022).

5. Initial descriptive analysis

We did a descriptive analysis to obtain a comprehensive understanding of the selected funding programs. As shown in Table 4, the Discovery Grants program supported a higher number of researchers compared to the other programs. However, the average amount of funding per record is the lowest for this program. This means that the Discovery Grants program is designed to support a large number of researchers rather than providing a substantial funding amount to each individual. Despite the lower average funding per record, the program's annual budget is the highest due to the large number of supported researchers. This highlights the program's importance to NSERC in supporting many researchers and fostering research excellence across various disciplines.

On the other hand, the Canada Research Chairs program and Strategic Projects program recorded the highest average amount of funding per record. These programs focus on exceptional researchers and aim to provide substantial funding to motivate and support their research interests. The higher average funding per record in these programs reflects the emphasis on recognizing and supporting outstanding researchers who can make significant contributions to their respective fields. It is worth noting that the Strategic Projects program has the least number of records in our dataset. This can be attributed to the program's specific focus on targeted areas of interest, limiting the pool of eligible researchers.

Lastly, the Collaborative Research and Development Grants program has supported the second-highest number of researchers and provided more than twice the funding amount per record compared to the Discovery Grants Program. This indicates the significance of the collaboration of industry with academia for NSERC.

Program	Total number of records	Number of distinct funded researchers	Average amount of funding per record	Average annual budget
Discovery Grants Program - Individual	231,550	66,190	30,324 $	260,062,176 $
Collaborative Research and Development	15,295	6,142	72,113 $	40,851,143 $
Canada Research Chairs	13,222	2,540	128,484 $	94,379,132 $
Strategic Projects – Group	8,740	3,092	122,792 $	39,748,540 $

Table 4 Information on selected NSERC funding programs

To better understand the trends existing in the funding allocation within the programs, we have plotted both the average amount of funding per researcher and the annual budget of each program over time. By combining these two perspectives, we can gain a comprehensive understanding of how funding has been allocated within the programs and how it has changed over the years. The funding information is available for the period of 1992 to 2018 for all programs except the Canada Research Chairs program, for which data is available from 2001 to 2018, as it was established in 2000.

As depicted in Figure 1, from 1992 to 1998, the annual budget remained relatively stable, but from 1998 onwards, there was an increasing trend, resulting in almost a doubling of the budget by 2018. A similar trend can be observed for the average amount of funding per researcher. From 1992 to 1998, the average funding per researcher followed a comparable pattern to the annual budget. Subsequently, there was an increase in the average funding amount until 2000, after which it remained relatively stable until 2009. From 2009 onwards, there was again an increasing trend in the average funding allocated to the researchers. Overall, the average funding granted to each researcher in each year increased by approximately 30 percent from 1992 to 2018.

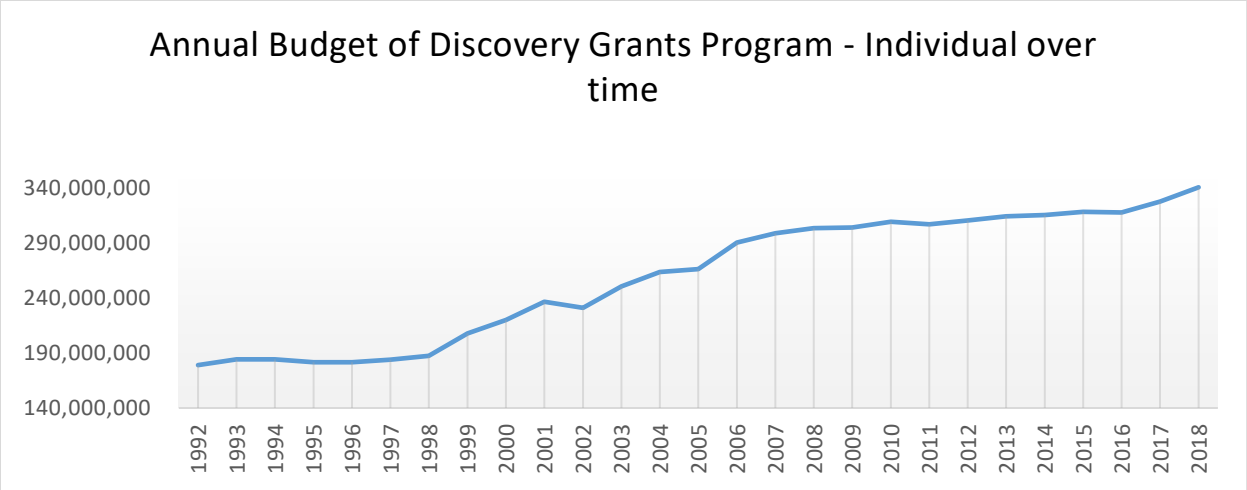


Figure 1 Annual budget of Discovery Grants Program – Individual over time

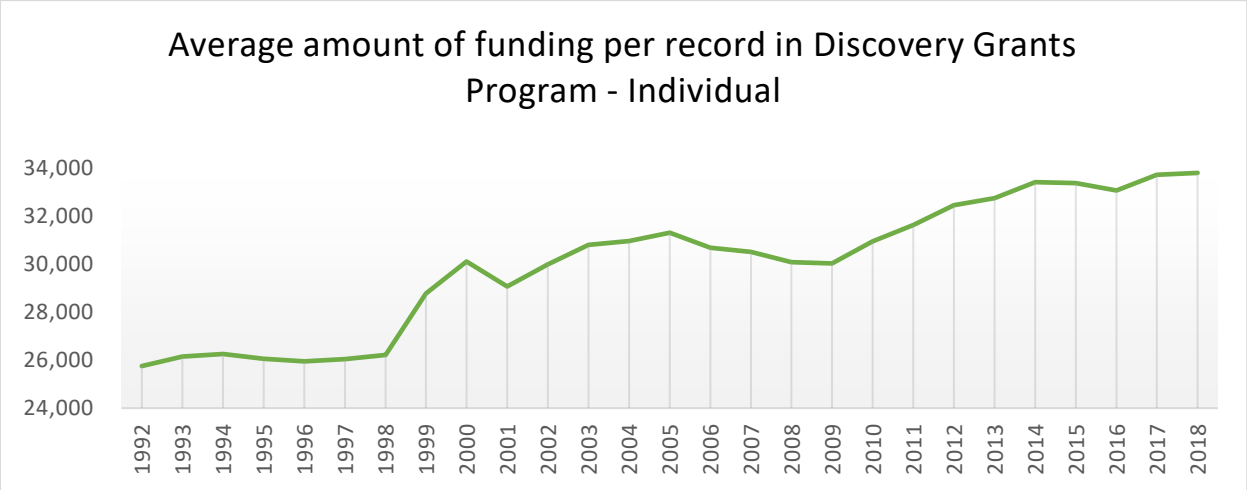


Figure 2 Average amount of funding per record in Discovery Grants Program – Individual

These two graphs suggest that the increase in the budget for the Discovery Grants program was not only intended to provide higher funding amounts to researchers but also to support a greater number of researchers. In fact, the relatively smaller increase in the average funding per researcher compared to the budget increase supports this notion. It could highlight the program's objective of supporting a broad number of researchers and promoting a diverse range of research projects within the natural sciences and engineering fields. The increase in the budget over the years can also reflect the importance placed on fostering research excellence and contributing to scientific advancements in NSERC.

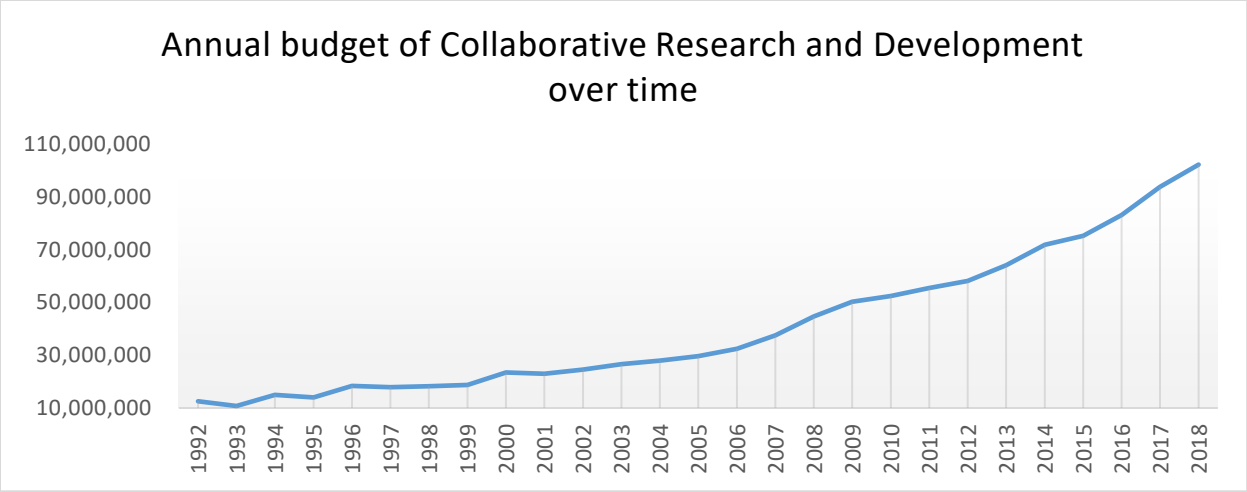


Figure 3 Annual budget of Collaborative Research and Development over time

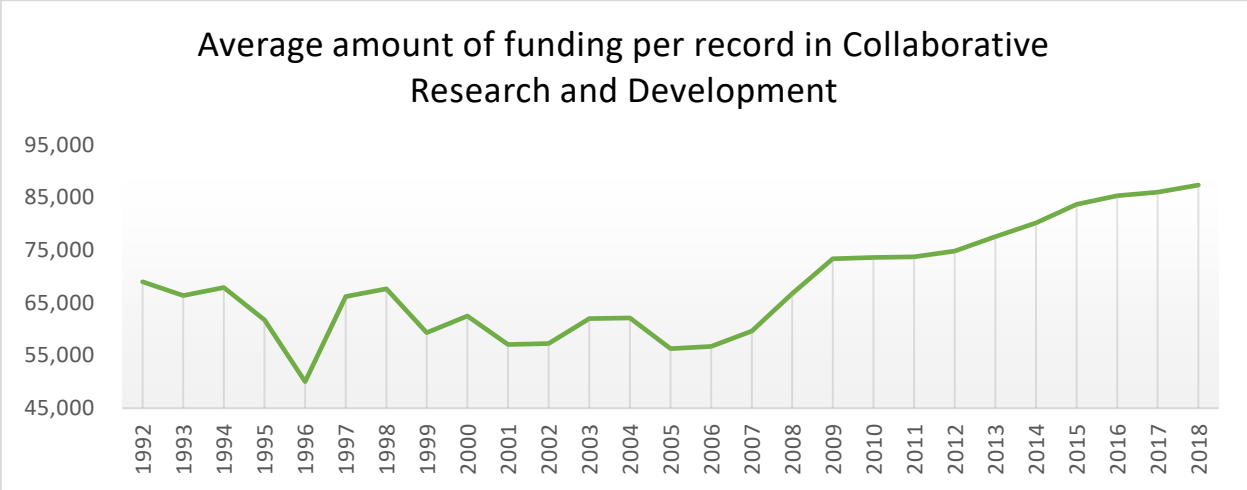


Figure 4 Average amount of funding per record in Collaborative Research and Development

Figure 3 and Figure 4 depict the annual budget and average funding per researcher for the Collaborative Research and Development (CRD) program from 1992 to 2018. During this period, the annual budget for the CRD program displayed a consistently increasing trend, coming from 12 million CAD to 102 million CAD. This substantial increase in the program's budget indicates the significant investment made by NSERC to support collaborative research initiatives between academia and industry in this period. Examining the average funding per researcher under the CRD program, we observe fluctuations from 1992 to 2006, after which a consistent upward trend emerges. The average funding per researcher gradually increased each year and reached an average of 87,000 CAD by the end of the period.

Since even the increase in the average funding per researcher after 2006 is not proportionate to the overall budget increase during the same period, we can conclude that the increase in the annual budget for the CRD program was primarily directed towards supporting a greater number of researchers rather than solely providing a higher amount of funds to researchers. Overall, the increasing trend in the annual budget and the moderate growth in the average funding per researcher for the CRD program demonstrate NSERC's commitment to promoting collaborative research endeavors and facilitating knowledge exchange between academia and industry.

The Canada Research Chairs (CRC) program displayed a distinct trend in terms of budget allocation and funding distribution compared to the other programs. Upon its establishment, the annual budget for the CRC program experienced a steady increase until 2009, after which it remained relatively stable until 2018. Examining the average allocation of funds to researchers under the CRC program, we observe an increasing trend from the program's establishment to 2005. During this period, the average amount of funding granted to researchers gradually increased, reflecting the program's focus on supporting exceptional researchers and motivating their research endeavors. However, starting from 2005, there was a slight decrease in the average funding allocation until 2009, after which it remained relatively steady.

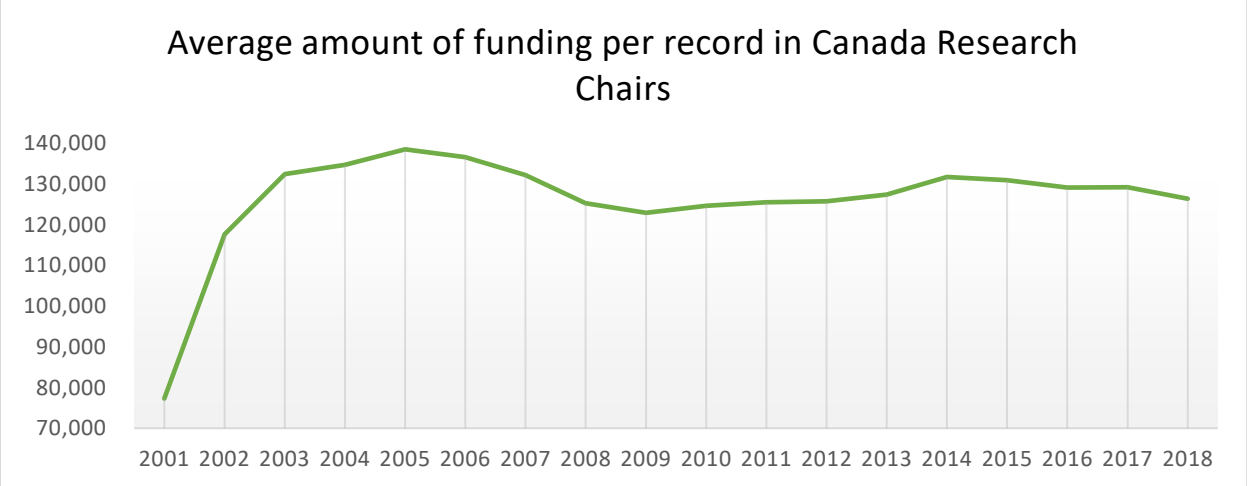


Figure 5 Average amount of funding per record in Canada Research Chairs

This trend of increasing funding followed by a period of stability in both the total budget and average allocation to researchers could be attributed to the program's establishment phase, during which there was a need to attract and support outstanding researchers. Once the program matured and achieved its intended goals, the budget and funding allocation became more consistent,

potentially indicating a balance between sustaining ongoing research projects and attracting new exceptional researchers. It is worth noting that the stability in the total budget and average funding per researcher suggests that the CRC program aimed to provide a consistent level of support to the selected researchers, ensuring their research activities can be conducted effectively without significant fluctuations in financial resources.

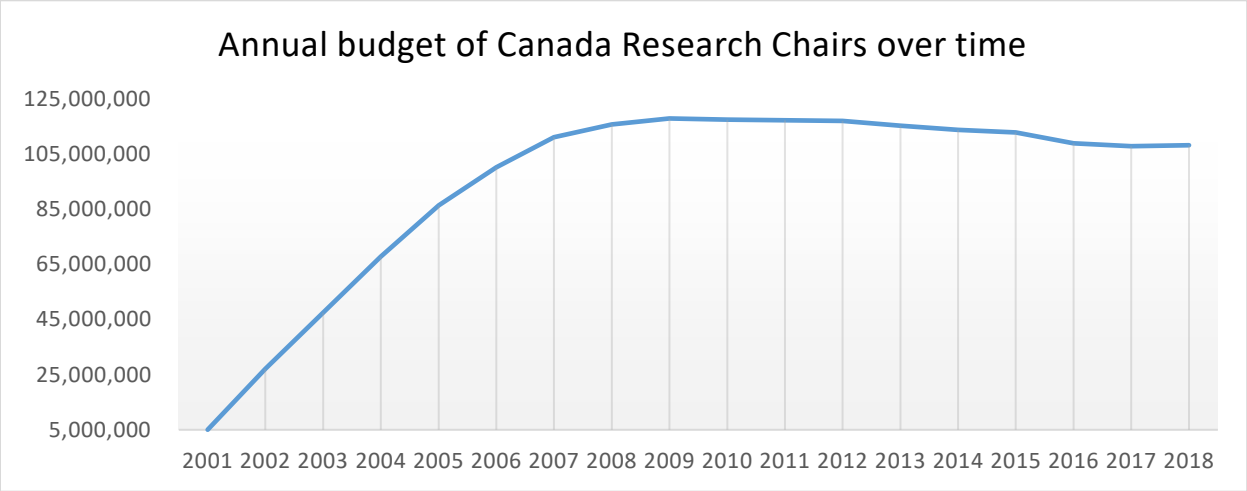


Figure 6 Annual budget of Canada Research Chairs over time

As depicted in Figure 7, the annual budget for the Strategic Projects - Group program remained relatively stable from 1992 to 1998. However, a notable increase in the budget can be observed from 1998 onwards, reaching a peak of 60 million CAD in 2010. Subsequently, there was a significant decline in the budget, dropping to 34 million CAD in 2014, followed by a subsequent increase to 45 million CAD by 2018. While the program's annual budget exhibited fluctuations, it is important to note that the average amount granted to successful applications showed a consistently increasing trend throughout the entire period, increasing from 84,000 CAD in 1992 to 177,000 CAD in 2018.

The trends indicate that NSERC has made a deliberate effort to provide increasing support to researchers who are successful in securing funding through this program. The fluctuations in the program's budget could be attributed to various factors, such as the evolving needs and priorities of NSERC. For instance, during periods of budgetary increase, it is likely that there was a greater demand for research in specific areas or a larger pool of researchers eligible to participate in the program. Conversely, budgetary decreases may reflect adjustments based on changing research priorities or resource availability.

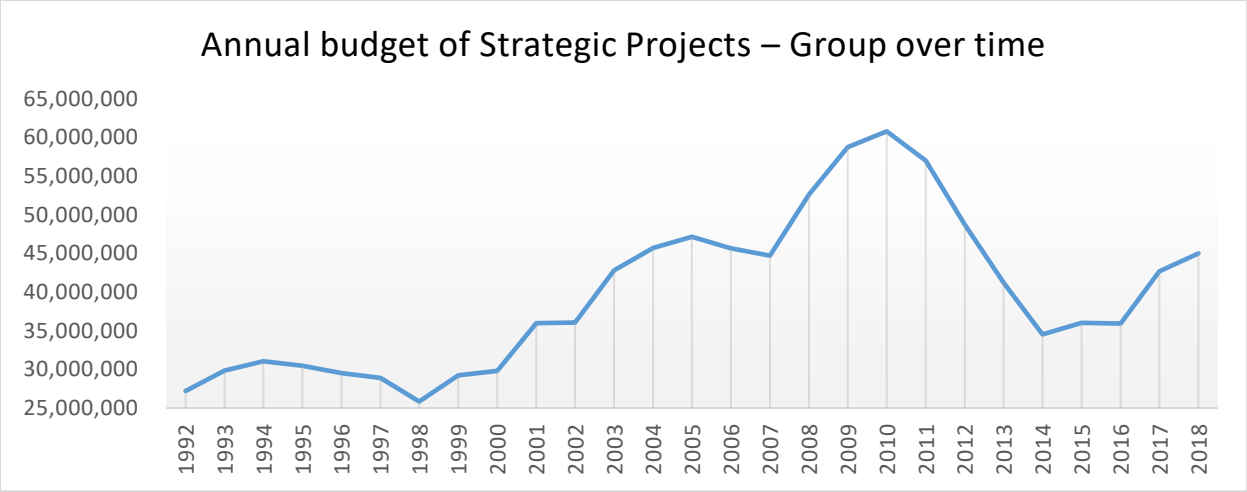


Figure 7 Annual budget of Strategic Projects – Group over time

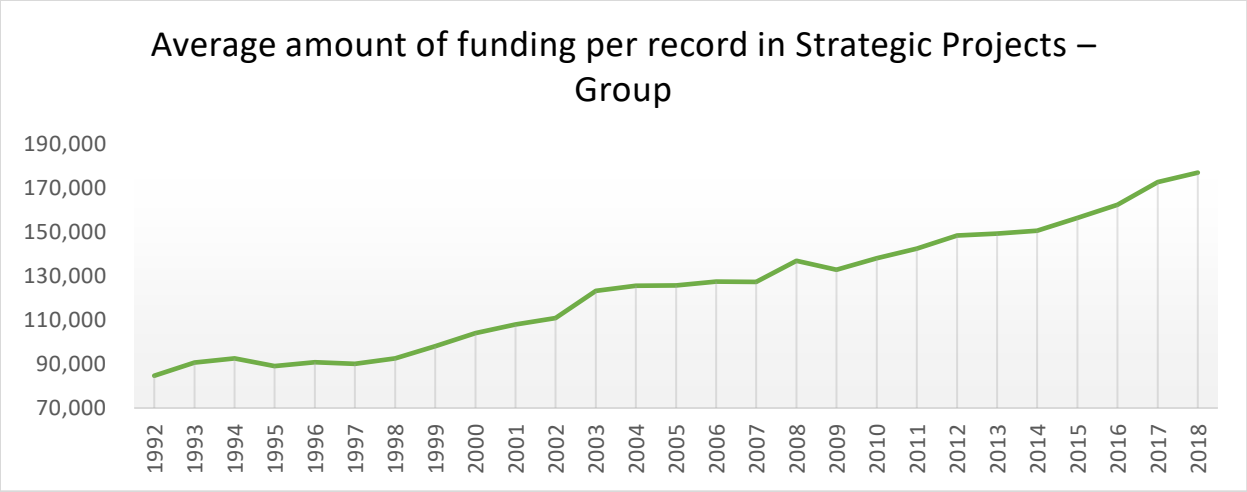


Figure 8 Average amount of funding per record in Strategic Projects – Group

The next step in our descriptive analysis is to compare the provinces that received funding under each program. By doing this, we can have a better understanding of the distribution of NSERC research funding across different regions in Canada. All four programs covered 10 provinces in Canada indicating a nationwide approach to research funding. Figure 9 provides an overview of the percentage of total spending for selected NSERC programs in Canada's provinces. While the allocation of funding seems to be related to the population of provinces, there are differences in the allocation of funds within each program.

Prince Edward Island consistently receives the lowest share of funding across all programs ranging from 0.07% to 0.35%. This could be attributed to several factors, including a smaller population, fewer research institutions, and potentially fewer research projects that align with the focus areas

of NSERC programs. Similarly, Newfoundland and Labrador, New Brunswick, Manitoba, Nova Scotia, and Saskatchewan also receive relatively lower amounts of funding compared to provinces with larger populations and more established research infrastructure.

Ontario received the largest share of funding in all four programs, ranging from 34.41% in Collaborative Research and Development to 40.5% in the Discovery Grants program. Ontario's large share of funding across all four programs, including the highest share in the Discovery Grants program, could indicate the province's robust research system. The Discovery Grants program is designed to provide support to a wide range of researchers and research projects across various disciplines. Ontario's large share in this program suggests that it has a significant number of researchers who successfully secure funding through this program.

Quebec, as the second-largest recipient of funding, received substantial shares in all programs, including 22.76% in Discovery Grants, 25.09% in Canada Research Chairs, 31.05% in Strategic Projects, and 33.18% in Collaborative Research and Development. Quebec's high share of funding in the Collaborative Research and Development program which is almost equal to the share of Ontario suggests a strong collaboration between industry and academia in this province. Also, Quebec's substantial share of funding in the Strategic Projects program indicates that researchers in the province are actively engaged in NSERC areas of interest. This could suggest that Quebec researchers are aligned with NSERC's priorities and are making significant contributions in addressing key challenges.

British Columbia's share of funding ranged from 10% in Collaborative Research and Development to 14.68% in Strategic Projects. British Columbia's higher share of funding (14.68%) in Strategic Projects signifies the province's strength in areas of interest identified by NSERC. Finally, Alberta’s share varies from 8.69% in Strategic Projects to 13.54% in Collaborative Research and Development. The relatively high share of funding in Collaborative Research and Development program could imply the significant involvement of Alberta-based researchers in collaborative efforts between academia and industry.

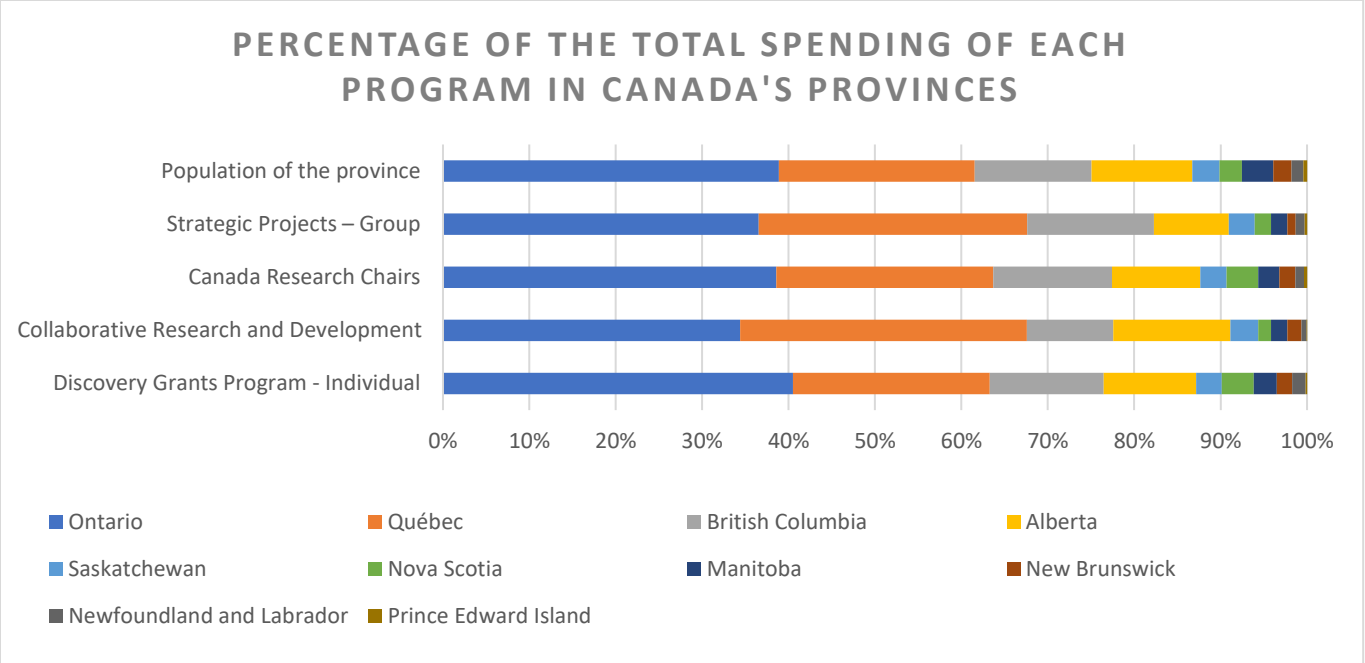


Figure 9 Percentage of the total spending of each program in Canada's provinces

In this section, we conducted a descriptive analysis to investigate the changes in budget allocation and the number of researchers supported by NSERC for the selected programs from 1992 to 2018. Our exploration is constrained by the absence of data on the number of applications, preventing a determination of the trend in success rates during this period. Despite this limitation, our analysis provides an understanding about the broader patterns and shifts within the specified timeframe. These descriptive findings serve as an initial step for our analysis. In the following section, we will discuss the methodology employed for our study, aiming to identify and assess the various factors influencing researchers' outcomes.

6. Methodology

As discussed before, this research aims to investigate the impact of funding programs on research impact, researchers' productivity, and collaboration. To achieve this goal, we compared three different machine learning models on the whole dataset as well as on four subsets of the data, each consisting of one of the funding programs selected for the research.

We chose multiple linear regression as the statistical baseline to understand the relationships between variables. This allowed us to better understand the impact of the independent variables

on the dependent variable. We then compared the results for multiple linear regression, random forest, and MLP neural network. This comparison was crucial in determining the best machine-learning model for our regression problem. To test our models and ensure their reliability, we divided the dataset into a train set (90% of the data) and a test set (10% of the data). After splitting the data, we used 5-fold cross-validation on the train set to evaluate the performance of the model. This approach allowed us to test the performance of our models on different subsets of the data and determine which of the models is better suited for our regression problem.

Since the study aimed to understand the impact of funding programs on three different aspects of scientific activities (impact, productivity, and collaboration), we performed separate analyses for each of these sections. For each analysis, we considered one of the aspects as the dependent variable and built our models separately. This allowed us to better understand the impact of each funding program on each of the three dependent variables.

In summary, our methodology involved selecting multiple linear regression as the statistical baseline, comparing the results of 5-fold cross-validation on multiple linear regression, random forest, and MLP neural network, performing separate analyses for impact, productivity, and collaboration, and validating our models using 5-fold cross-validation on the training set.

6.1 Variables

6.1.1 Dependent Variables

The variables used in our models are key to understanding the impact of funding on collaboration, productivity, and impact. For each of these three terms, we selected a dependent variable. For all the dependent variables, we considered the publications for a 3-year and 5-year window after they received an installment of funding from NSERC. After testing both time windows in our models, we observed more robust results when employing the 3-year time window.

1) 3-year future co-authorship:

We used co-authorship as a proxy to measure the collaboration of funded researchers. Using co-authorship as a proxy for collaboration, we can estimate the average number of authors working with a funded researcher over time. To do so, we calculated the average number of co-authors for the publications of researchers for a 3-year period after they received an installment of funding from NSERC. In other words, we considered the articles that a certain

funded researcher published in the next 3 years after receiving the funding and calculated the average number of authors for these articles.

2) 3-year future productivity:

The other objective of our research was to analyze the productivity of researchers who received funding. To measure this, we quantified the number of publications that the researchers produced after receiving funding from NSERC. We calculated the number of publications for a 3-year window after receiving the funding. This variable will allow us to understand how funding impacts the productivity of researchers.

3) 3-year future SJR:

Another aspect of our research is the examination of the impact of work produced by the funded researchers. To accomplish this, we used the average SJR of the journals in which researchers published their publications within a 3-year after receiving funding, as our dependent variable to measure the impact of their work.

6.1.2 Independent Variables

As previously mentioned, we will apply our models to the entire dataset, as well as to subsets of the dataset that contain information on each funding program individually. The independent variables will remain the same across all models, with the exception that we will also include a one-hot encoded funding program categorical variable when conducting the models on the entire dataset.

1) Funding program:

The funding program is the only categorical variable used in our models when performed on the whole dataset. To analyze the impact of each funding program on the dependent variable when we include all rows of the dataset in the analysis, we used a binary transformation of the funding program using one-hot encoded transformation. This transformation allowed us to include the funding program variable as an independent variable in our models and helped us understand the impact of each funding program on the dependent variable.

2) 3-year past productivity:

Another independent variable that we used in our models is the past productivity of funded researchers. We calculated the number of articles that each researcher had published in the three years prior to receiving funding, and we chose to use this as our independent variable. By analyzing the impact of past productivity on collaboration, productivity, and quality, we aimed to better understand how the previous productivity of researchers could affect their future outcomes.

3) 3-year past citation counts:

Similarly, we incorporated the number of citations received by the articles written by researchers three years before receiving funding as an independent variable in our models. By analyzing the impact of past citation counts on collaboration, productivity, and quality, we aimed to gain insights into how the citation impact of researchers' prior work contributes to their future outcomes. This variable allowed us to examine whether a higher number of citations received by their previous articles correlates with increased collaboration, productivity, and higher-quality research after receiving funding.

4) 3-year past co-authorship:

Another independent variable we considered is the 3-year past co-authorship. This variable represents the level of collaboration among funded researchers by calculating the average number of authors per article in the three years prior to receiving funding. It serves as a proxy to measure the extent of collaboration among researchers during that period. By analyzing the impact of this variable on collaboration, productivity, and quality, we aimed to understand how pre-existing co-authorship patterns affect researchers' future outcomes.

5) 3-year past SJR:

To evaluate the quality of the work produced by funded researchers in terms of journal impact factor, we introduced the average of SJR over the past 3 years as an independent variable. SJR is a metric that measures the quality and impact of scientific journals. By examining the 3-year past SJR, we aimed to understand how the prior journal impact factor of researchers' publications influences their future outcomes. This variable allowed us to investigate the relationship between the quality of the journals in which researchers have previously published and their subsequent collaboration, productivity, and overall research quality. While we also

considered the h-index as a measure of research impact, we found that SJR was better suited for our research problem. The SJR metric provides a comprehensive evaluation of journal quality, taking into account the influence of both the journals in which researchers publish and the citations their articles receive.

6) Career age:

We also considered the career age of the researchers as an additional independent variable. Career age represents the duration of a researcher's career in terms of their time in the field of study. To calculate career age, we subtracted the date of their first publication from the date they received funding. In cases where researchers had no prior publications, the career age was assigned a value of zero. By incorporating career age, we aimed to examine how the length of a researcher's career impacts their collaboration, productivity, and overall research outcomes after receiving funding. This variable allowed us to explore the potential influence of experience and accumulated knowledge on the researchers' subsequent performance and success.

7) Award amount:

The final independent variable we considered is the "award amount" which represents the amount of funding received by a researcher in a specific year under a specific funding program. Since the funding amounts within each program are typically close to each other, this variable holds particular significance when we perform our model on the entire dataset, as it helps us understand the potential impact of the funding amount on the researchers' outcomes. Incorporating the award amount as an independent variable allows us to differentiate and assess the influence of varying funding levels on collaboration, productivity, and research quality. By including this variable in our analysis, we can examine how the financial resources provided through funding programs contribute to the outcomes of researchers in a comprehensive manner.

6.2 Data Analysis Methods

6.2.1 Multiple Linear Regression

To achieve the objectives of our research, we used multiple linear regression as the statistical baseline for our data analysis. By choosing the linear regression model as our baseline, we can

assess whether the advanced machine learning models offer improved predictive accuracy or capture additional complexities that cannot be adequately captured by linear regression alone.

Simple linear regression is the starting point of this framework that assumes an approximately linear relationship between the dependent variable and the predictor. The equation for simple linear regression is (James, et al., 2013; Vercellis, 2009):

$$Y = \beta_0 + \beta_1 X + \varepsilon \qquad (1)$$

In this equation, Y represents the dependent variable, X represents the predictor variable, β_0 is the intercept term that is the value of Y when X = 0, β_1 is the coefficient associated with X which is the slope of the line, and ε represents the error term accounting for all the variability that the linear regression would miss. However, many real-world scenarios require the consideration of multiple predictors. This is where multiple linear regression comes into play as an extension of simple linear regression. In multiple linear regression, the relationship between the dependent variable Y and multiple predictors $X_1, X_2, ..., X_n$ is modeled. Therefore, the equation for multiple linear regression takes the following form (James, et al., 2013; Vercellis, 2009):

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \varepsilon \qquad (2)$$

In this equation, Y represents the dependent variable, while X_1, X_2, ..., X_n represent the independent variables. The coefficients β_0, β_1, ..., β_n denote the weights associated with each predictor, and ε represents the error term accounting for the variability not explained by the model. The extension from simple linear regression to multiple linear regression allows us to capture the effects of multiple predictors simultaneously and consider their combined impact on the dependent variable.

As discussed before, our aim in this study is to investigate the impact of different funding programs on the productivity of researchers, the impact of their work, and collaboration among them. For each of our dependent variables, first, we seek to analyze the influence of independent variables on the dependent variables across all funding programs by applying our model to the whole dataset. By doing so, we study the relative importance of different independent variables in predicting the output. In this model, we include funding programs as the categorical independent variable to capture the potential impact of receiving different types of funding. To do so, we use the "OneHotEncoder" package in python, transforming the funding programs into binary dummy

variables. As in dummy variables, it is common to remove one of the levels to avoid multicollinearity, we excluded the Canada Research Chairs program in transforming the funding programs variable to use it as the baseline for comparison against the other three programs. Consequently, we have now in our model Discovery Grants Program - Individual, Strategic Projects - Group, and Collaborative Research and Development programs as our independent variables. As a result, our regression model is as follows:

$$DependentVar = \beta_0 + \beta_1 * PastProductivity + \beta_2 * Citations + \beta_3 * CoAuthor + \beta_4 * SJR + \beta_5 * CareerAge + \beta_6 * AwardAmount + \beta_7 * DiscoveryGrants + \beta_8 * StrategicProject + \beta_9 * CollaborativeResearch \quad (3)$$

Second, we study the impact of independent variables influence the outcomes within each specific funding program by focusing on subsets of the dataset that pertain to each funding program. Therefore, the three dummy variables in the abovementioned equation will be removed from our model. Hence, the regression model became:

$$DependentVar = \beta_0 + \beta_1 * PastProductivity + \beta_2 * Citations + \beta_3 * CoAuthor + \beta_4 * SJR + \beta_5 * CareerAge + \beta_6 * AwardAmount \quad (4)$$

6.2.2 Random Forest

Random forest is a machine learning algorithm that can be used for both classification and regression tasks. For a regression problem, random forest creates an ensemble of decision trees and combines their predictions to generate a final output (Hastie, et al., 2009). Unlike the traditional linear regression model, random forest is a powerful method to deal with complex, nonlinear relationships between predictors and the target variable.

Random forest is an ensemble learning method that makes predictions by combining multiple decision trees. The number of decision trees in the ensemble is a hyperparameter that is set before the learning process. For building each decision tree, a subset of the dataset would be randomly selected which is called bootstrapping. This process would lead to diversity in the training data resulting in a smaller risk of overfitting. Moreover, for each split in a decision tree, a random subset of the features would be considered. This is another hyperparameter that is called the maximum number of features (Hastie, et al., 2009; James, et al., 2013; Muller & Guido, 2018).

The next step is the recursive partitioning in which each decision tree is constructed. Building decision trees starts with the entire dataset, selects the best feature, and split point that minimizes a pre-defined criterion, such as the mean squared error (MSE). Then, the data is divided into two subsets based on the split point. The process is recursively repeated for each resulting subset until it meets the termination condition. The user defines the stopping condition which could be criteria such as the maximum depth of the tree or the minimum number of samples required for splitting a node. When all the trees are created, each individual tree makes predictions. For a regression problem, by aggregating the prediction of all trees, the final output is obtained. Usually, the output is calculated as the average amount, or the weighted average of the predictions made by individual trees (Hastie, et al., 2009; James, et al., 2013; Muller & Guido, 2018).

To apply random forest model on our dataset, we tuned the hyperparameters to find the optimal combination for predicting each dependent variable. To enhance computational efficiency and reduce the overall execution time, first, we ran random forest model by increasing the number of trees included in the model to find the number for which the performance metrics are not improved anymore. Then, we used grid search for comparing different combinations of hyperparameters to find the best combination. As for each dependent variable, first, we applied our models on the whole dataset and subsequently on the subsets of the data that includes each of the funding programs, we tuned the hyperparameters for each scenario separately. For each scenario, the values of the hyperparameters are summarized in the following tables:

Future Productivity					
	Whole Dataset	Discovery Grants	Strategic Projects	Collaborative Research & Development	Canada Research Chairs
n_estimators	25	25	25	25	25
max_depth	20	10	10	20	20
min_samples_split	5	10	5	5	5
min_samples_leaf	2	2	2	2	2
max_features	0.5	None	0.5	0.5	0.5
Future Impact					
n_estimators	50	50	50	50	50

max_depth	20	10	20	20	10
min_samples_split	2	5	5	5	5
min_samples_leaf	4	2	4	4	1
max_features	0.5	0.5	sqrt	1	log2
Future Collaboration					
n_estimators	50	50	50	50	50
max_depth	10	10	10	10	10
min_samples_split	5	5	2	2	2
min_samples_leaf	4	1	1	4	3
max_features	0.5	1	1	1	log2

Table 5 Hyperparameters of random forest models

6.2.3 Artificial Neural Network

The multilayer perceptron artificial neural network is a machine learning technique that we considered in order to compare the results with the other two methods. Artificial neural network could be seen as an alternative because of its functionalities and advantages. This method effectively deals with non-linear relationships between input variables and the target output. As a result, it could offer better prediction results while enabling the efficient processing of large datasets and potentially handling high-dimensional inputs.

The idea behind the artificial neural network is inspired by the biological neurons in humans' and animals' brains. Figure 10 shows an artificial neuron, where in the input-output relationship we have multiple inputs and a single output (Hagan, et al., 2014).

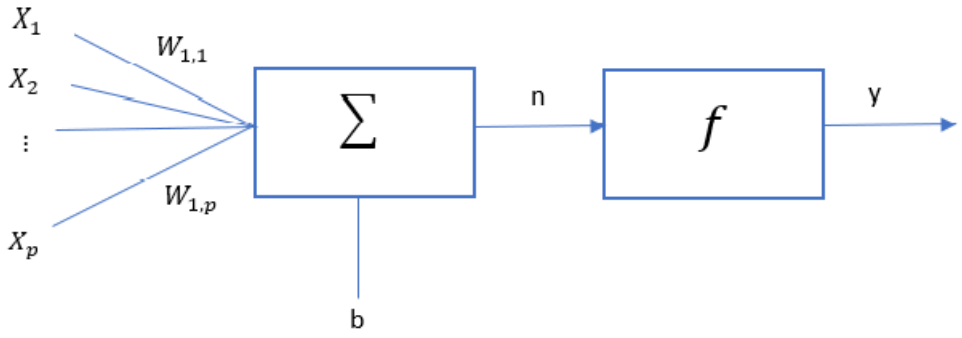


Figure 10 An artificial neuron

The equation for this neuron showing its input-output relationship is:

$$y = f(n) = f\left(\sum_{i=1}^{p} w_{1,i}x_i + b\right) = f(Wx + b) \qquad (5)$$

In this equation y is the output of the neuron, f denotes the activation function, b is the bias parameter, W represents the weights vector, and x indicates the input vector assuming that the neuron has p inputs. As it is shown in Figure 10, the weighted sum of the inputs would be added to the bias. The output y would be the result of an activation function f applied to this summation. W and b are the learnable parameters while the activation function is a specific linear or non-linear function that should be determined in accordance with the problem's specifications. Some of the commonly used activation functions are shown in Figure 11. It is possible to form networks that can create more complex models by grouping the neurons (Kůrková, 1992). The number of layers and the number of nodes are set by the user based on the problem that should be tackled.

In neural networks, the learning process is used to update the learnable parameters (i.e., weights and biases). The backpropagation algorithm is used in a neural network for the training process (Hagan, et al., 2014; Goh, 1995). After defining the appropriate cost function like mean squared errors, cross-entropy, etc. the weights and biases are updated in a way that minimizes the loss associated with the cost function. To do so, an initial value would be assigned to each parameter using different methods like random generation of values. Then, weights are updated iteratively in a way to find the local minimum for the cost function. An iteration is a backpropagation on a single data point. When backpropagation is done for the whole training dataset, it is called an epoch. Backpropagation would update the values of the weights so that after each iteration the value of the cost function is decreased. The weights of the model are updated using the gradient descent method as shown in the following formula:

$$w_{i,j}^{new} = w_{i,j}^{old} + \Delta w_{i,j}^{old} = w_{i,j}^{old} - \alpha \frac{\partial C}{\partial w_{i,j}^{old}} \qquad (6)$$

In this formula, $w_{i,j}$ is the weight of j^{th} node in the i^{th} layer. C is the cost function and α represents the learning rate. The learning rate is another hyperparameter which should be set by the user to define the steps by which the weights should be updated. Determining an appropriate learning rate could be critical since a large learning rate might not lead to the optimal solution because of

overshooting the minimum while small learning rates are time consuming. An alternative solution could be to start with large learning rates and then reduce it when we approach the minimum.

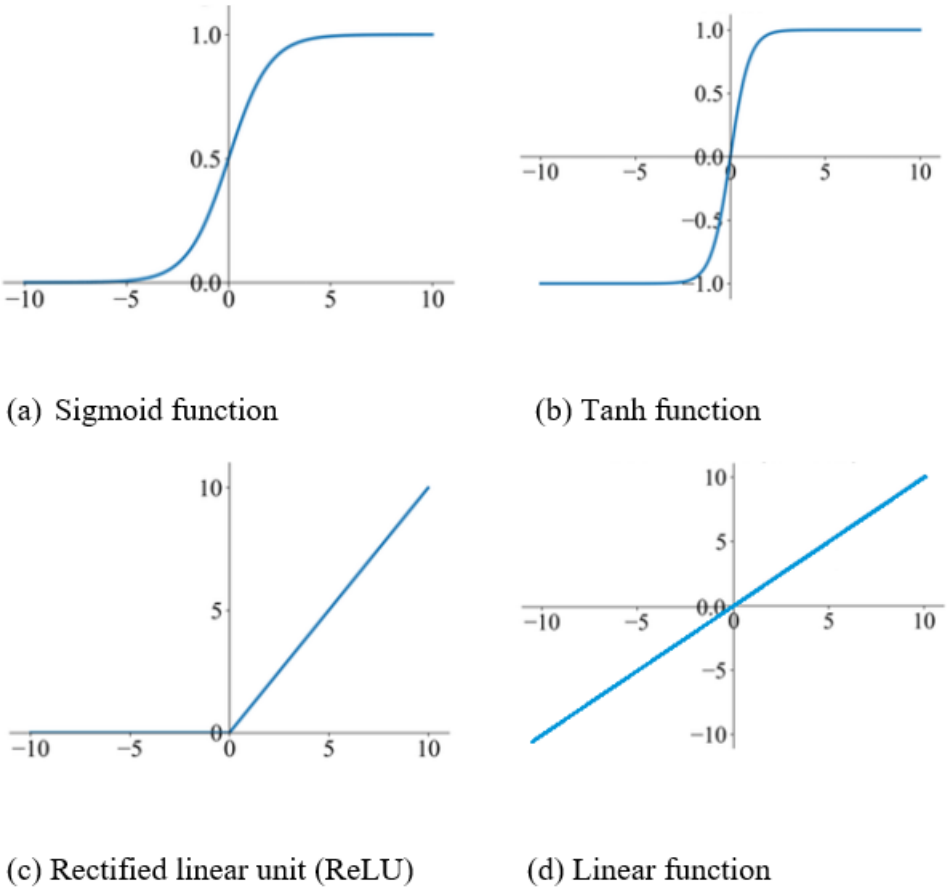


Figure 11 Examples of common activation functions

For applying neural network in predicting each of dependent variables, we tuned the neural network's hyperparameters. We used Keras Tuner with a random search approach to identify the most influential factors. Through an iterative process, we explored different configurations to determine the optimal network architecture. We found that the optimal model for all the three dependent variables within our dataset remained the same. The only variation among the models pertained to the number of epochs, which ranged from 30 to 40. We determined that a network with 2 hidden layers yielded the optimal performance. Layer 1 comprised 32 neurons, layer 2 had 16 neurons, and the output layer featured 1 neuron. The ReLU (Rectified Linear Unit) activation

function delivered the best results for both hidden layers. Other hyperparameters are determined as follows:

- Batch Size: 32
- Learning rate: 0.001
- Optimizer: Adam
- Loss function: Mean Squared Error (MSE)

7. Results and Discussion

In this section, first, we discuss the correlation matrix of the variables included in our models. By doing so, we aim to obtain an understanding of the linear relationships among these variables. The correlation analysis provides valuable insights into the relationships between the dependent variables (i.e., future productivity, future SJR, and future co-authorships) and their corresponding past performance variables.

As shown in Figure 12, the results indicate that a researcher's past performance in each area has a substantial linear relationship with their future outcomes. Future collaboration, which is represented by co-authorship, exhibits a strong 0.7 correlation with past co-authorship. Similarly, future productivity, measured by the number of publications, shows a notable 0.69 correlation with past productivity. Moreover, past SJR of researchers has a strong correlation (0.64) with future SJR. Apart from these three, the value of other correlation coefficients is relatively small that shows the linear correlation of the other variables is weak. Therefore, in general, the variables in our dataset exhibit low correlations with each other.

Given the generally low correlations among the variables, we can understand that using linear models alone might not be sufficient to fully capture the complex and non-linear relationships in the data. Therefore, we have chosen linear regression as a starting point for our analysis, but to better capture the relationships among the variables, we have performed MLP neural network and random forest to compare their results with linear regression and understand which model is better fitted for our dataset. By combining the strengths of linear regression with the non-linear capabilities of MLP neural network and random forest, our research analysis aims to yield more accurate predictions and a deeper understanding of the relationships among the variables.

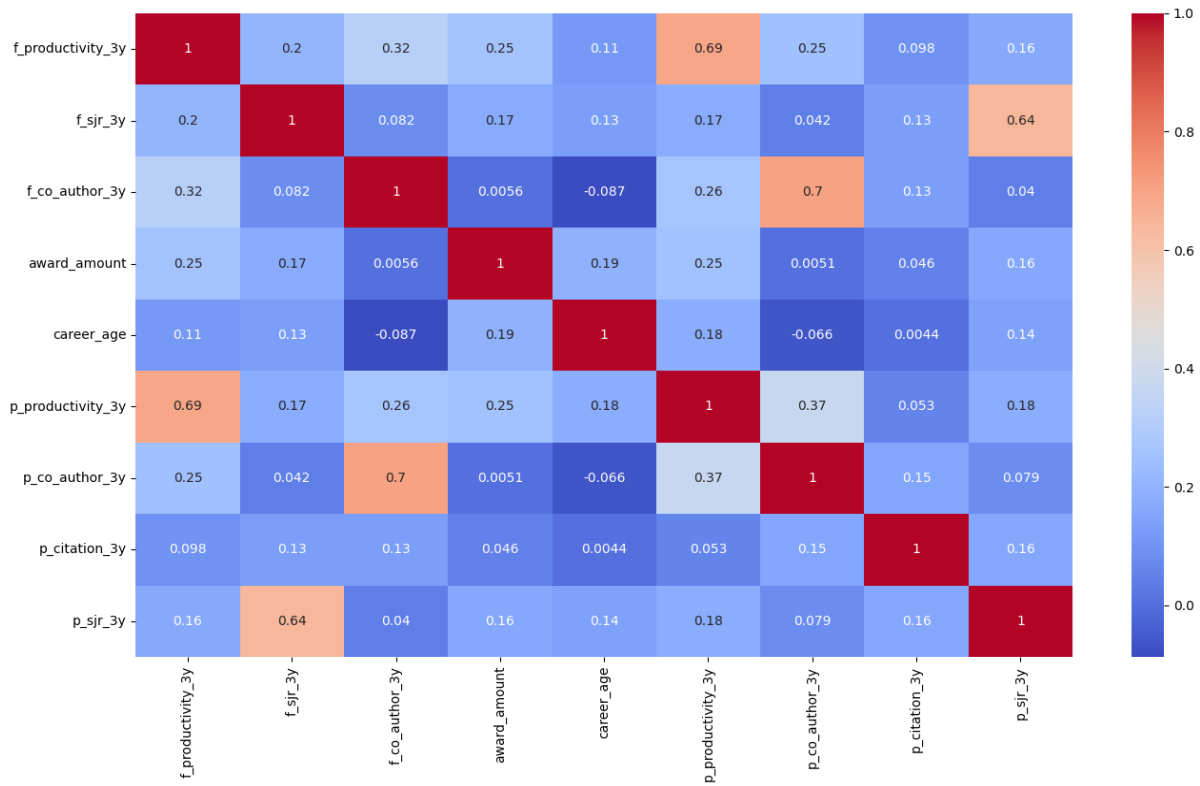


Figure 12 Correlation matrix of the variables

In the following sections, we will discuss the results of applying linear regression, random forest, and MLP neural network to our dataset. Our dataset comprises information from four selected funding programs. First, we seek to understand how each factor is important in predicting the dependent variables across all funding programs by applying the models to the whole dataset. By doing so, we will be able to determine the relative importance of different independent variables in predicting the dependent variables. This approach would allow us to have a comprehensive view of the factors that play a role in determining the outcomes of the funding programs. Second, we aim to explore how these independent variables influence the outcomes within each specific funding program by focusing on subsets of the dataset that pertain to each funding program. By doing so, we can gain deeper insights into how the significance of different factors varies across different programs.

For each of the sections, we first provide the results of linear regression as the baseline and the coefficients of the variables. Then, we compare the performance of 5-fold cross-validation linear regression, random forest, and MLP neural network. For comparing the precisions, we have considered four metrics: Mean Squared Error (MSE), Root Mean Square Error (RMSE), Mean

Absolute Error (MAE), and R-squared. Based on these metrics, we would identify the model that is better suited to analyze the results. After deciding about the model, we would present the significance of the input variables and compare the results across the funding programs.

7.1 Results on Future Productivity

The first dependent variable that we have analyzed is the future productivity of researchers. We started by cleaning and normalizing the dataset to ensure data accuracy and consistency. Once the data was prepared, we performed multiple linear regression analysis on the whole dataset. In this model, we included funding programs as the categorical independent variable to capture the potential impact of receiving different types of funding.

For performing the linear regression as the baseline, we have performed 5-fold cross-validation on the 90% of the data to train the model and test it on the remaining 10% of data. The F-statistic obtained from the regression analysis is 2.642e+04 and the p-value associated with the F-statistic is 0. This result indicates that the model is statistically significant, suggesting that at least one of the independent variables has a significant effect on the future productivity of researchers. The R-squared of the model is 0.496 indicating that around 50% of the variability in the future productivity of researchers is explained by the independent variables included in the model. Therefore, this model is not very good at capturing the variability of the future productivity of researchers. Considering the evaluation of the model on the test set, we have examined different metrics to assess its performance. The results of these metrics on the evaluation set for each part of the 5-fold cross-validation and test set are summarized in the following table:

	Mean Squared Error (MSE)	Root Mean Square Error (RMSE)	Mean Absolute Error (MAE)	R-Squared (R^2)
Fold 1	94.02	9.70	4.68	0.52
Fold 2	104.92	10.24	4.71	0.47
Fold 3	102.31	10.11	4.67	0.48
Fold 4	96.72	9.83	4.64	0.50
Fold 5	97.09	9.85	4.69	0.50
Test set	98.75	9.94	4.68	0.49

Table 6 Performance of linear regression on evaluation set for each part of the 5-fold cross-validation and test set for future productivity

The results of the multiple linear regression are presented in Table 7. Since all the p-values related to the variables are around 0, we can conclude that all the coefficients of the variables included in the model are statistically significant and significantly different from zero. Regarding the funding programs, we considered the Canada Research Chairs program as the baseline for comparison to other funding programs with it. The positive coefficients related to the Collaborative Research and Development grants and Strategic Projects suggest that researchers receiving these two funding programs might publish more articles compared to those granted the Canada Research Chairs. On the other hand, the negative value of the coefficient of Discovery Grants program shows that the researchers receiving this funding have published fewer articles compared to the other programs included in the dataset.

Variable	Coef.	Std. Err.	t	P > |t|	[95% Conf. Interval]
Const.	10.5218	0.111	94.793	0.000	10.304 0.739
Award_amount	0.8617	0.029	29.506	0.000	0.804 0.919
Career_age	-0.3324	0.021	-15.607	0.000	-0.374 -0.291
P_productivity_3y	9.4659	0.023	410.505	0.000	9.421 9.511
P_co_author_3y	-0.2384	0.022	-10.711	0.000	-0.282 -0.195
P_sjr_3y	0.4091	0.021	19.331	0.000	0.368 0.451
P_citation_3y	0.7438	0.021	35.943	0.000	0.703 0.784
prgrm_Collaborative Research and Development Grants	0.8771	0.131	6.685	0.000	0.620 1.134
prgrm_Discovery Grants Program - Individual	-1.0873	0.117	-9.264	0.000	-1.317 -0.857
prgrm_Strategic Projects - Group	0.4594	0.145	3.161	0.000	0.175 0.744

Table 7 Summary of the results of the multiple linear regression for future productivity

Based on the regression results, we can understand that the most significant factor in predicting the future productivity of researchers is their past productivity. The coefficient associated with past

productivity stands out as considerably higher compared to the coefficients of other variables, indicating a strong relationship between these two factors. It was expected since productive researchers are supposed to remain productive in the future as well.

Among the other variables, the award amount has the highest coefficient. This can highlight that receiving more funding can play a crucial role in enabling researchers to sustain or enhance their productivity. Additionally, citation counts and the prestige of the journal (SJR) are the other variables with positive coefficients. The impact of the citation counts on future productivity is almost twice the coefficient of the SJR. This result shows that the quality of previous works also positively impacts the productivity of researchers in the near future. The higher impact of citation counts compared to journal prestige might suggest that receiving attention and recognition from peers is more influential in driving researchers' future productivity than publishing in high-prestige journals. It underscores the value of research impact and how being cited by other researchers can contribute significantly to the growth and productivity of researchers.

Moreover, the results reveal that the career age of the researchers and past co-authorship have negative coefficients implying that the higher their amount the lower would be the future productivity of researchers. In fact, we can conclude that as the career age of researchers increases, their future productivity may experience a decline. Furthermore, the small negative impact of the past co-authorship suggests that extensive involvement in co-authorship, particularly in large groups of researchers, may not necessarily translate into increased productivity in the future.

In the next step, we want to compare the predictive capabilities of the selected models that are 5-fold cross-validation multiple linear regression, random forest, and MLP neural network. To optimize the performance of each model, we conducted hyperparameter tuning for both neural network and random forest, aiming to identify the best combination of factors for each model while optimizing computational efficiency.

To ensure a fair comparison among the models, we selected the best-performing results of each model after hyperparameter tuning. This approach guarantees that each model is operating at its optimal configuration. The performance of the models was evaluated on the test set after completing the training process. By employing a separate test set, we ensured an unbiased assessment of their predictive ability on unseen data. Table 8 presents a summary of the models' performance based on the defined evaluation metrics.

	MSE	RMSE	MAE	R-Squared
Linear regression	98.75	9.94	4.68	0.49
Random forest	54.59	7.39	4.03	0.72
MLP neural network	68.16	8.26	4.20	0.65

Table 8 Summary of the models' performance for future productivity on the test set

By evaluating the performance of the models, it is evident that random forest shows better suitability for our dataset. While a higher R-squared value is preferred to indicate better performance, the opposite holds true for the other metrics, where a lower value means enhanced model performance. As a result, we can conclude that random forest outperforms the other alternative models across all four defined metrics. Consequently, we decided to focus on this model for further analysis.

After deciding about the model, the next step is to consider the importance of the input variables on the model's output. In this pursuit, we leveraged on SHAP values, a tool that provides a quantified assessment of the impact of variables on the predicted outcome. This approach enables us to gauge the relative importance of each independent variable in influencing the model's overall performance. (Lundberg & Lee, 2017)

The beeswarm plot depicted in Figure 13 provides a visual presentation of the SHAP values associated with our final model. Within this plot, the relative influence of the independent variables on the predicted outcome is illustrated, ranging from the most significant (located in the upper left quadrant) to the least influential (situated in the lower left quadrant).

Based on the SHAP analysis, we can claim that the most important determinant in forecasting future productivity is the past productivity of researchers. The positive impact of past productivity is line with finding of Ebadi & Schiffauerova (2016) who claimed that in addition to the fact that it is probable for researchers to maintain or increase their productivity in the future, past productivity could help researchers to secure more funding in the future which can increase their chance to publish more articles. The Figure 13 displays a concentrated cluster of instances characterized by low past productivity (shown as blue data points) exhibiting small negative SHAP values. In contrast, instances characterized by high past productivity show notably positive SHAP values. This indicates the substantial impact of high past productivity on the prediction of future

productivity. On the other hand, though lower past productivity leads to a lower predicted number of publications in the future, the magnitude of this adverse impact is smaller.

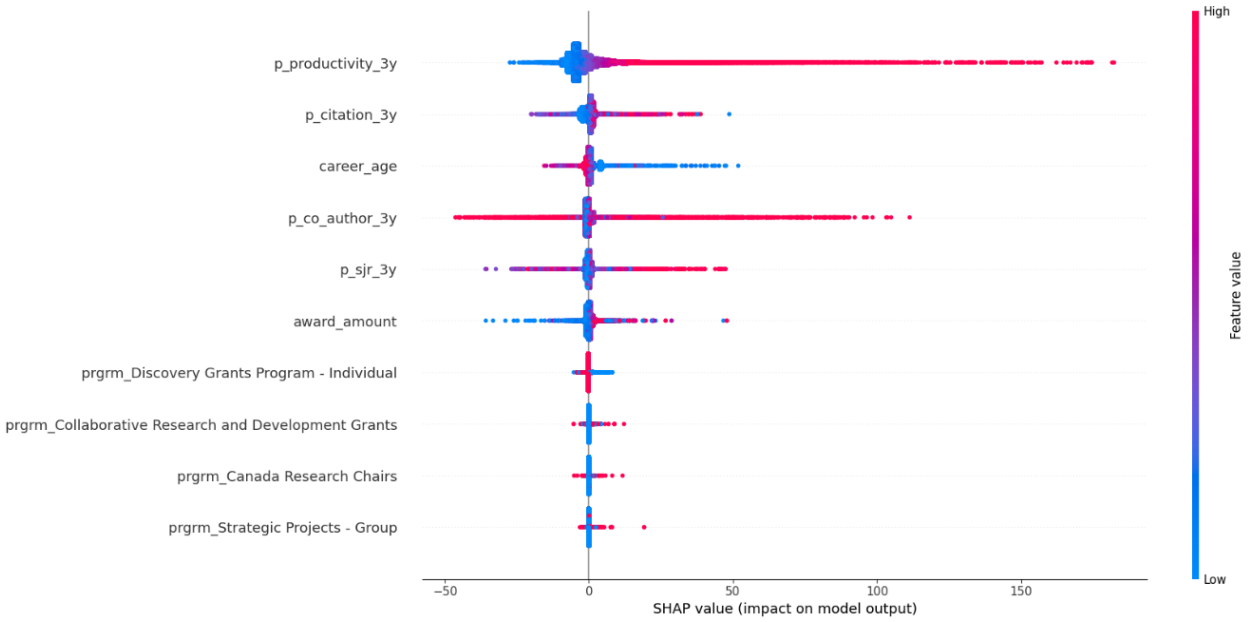


Figure 13 Beeswarm plot for future productivity – Entire dataset

The second important variable is the past citation counts of the researchers. Like the impact of past productivity, a high number of citations would have a positive impact and a low number would negatively affect the future productivity but in a balanced way. This finding is line with Ebadi & Schiffauerova (2016) who also found a positive impact of citations on the future productivity of researchers.

SHAP values of career age show a different pattern where the high values for career age negatively affect the output, but the low values would have a positive effect. As a result, we can claim that young researchers are more likely to be productive. This is in contrast with finding of Ebadi & Schiffauerova (2016) who found an overall positive impact of career age on the productivity. The difference in the results could be because of using different methodologies. After citation counts, the next important variables are co-authorship and SJR. Surprisingly, among numeric independent variables, the award amount is the least important. It is worth noting that the lower impact of award amount could be a result of considering only funded researchers granted by NSERC. By comparing

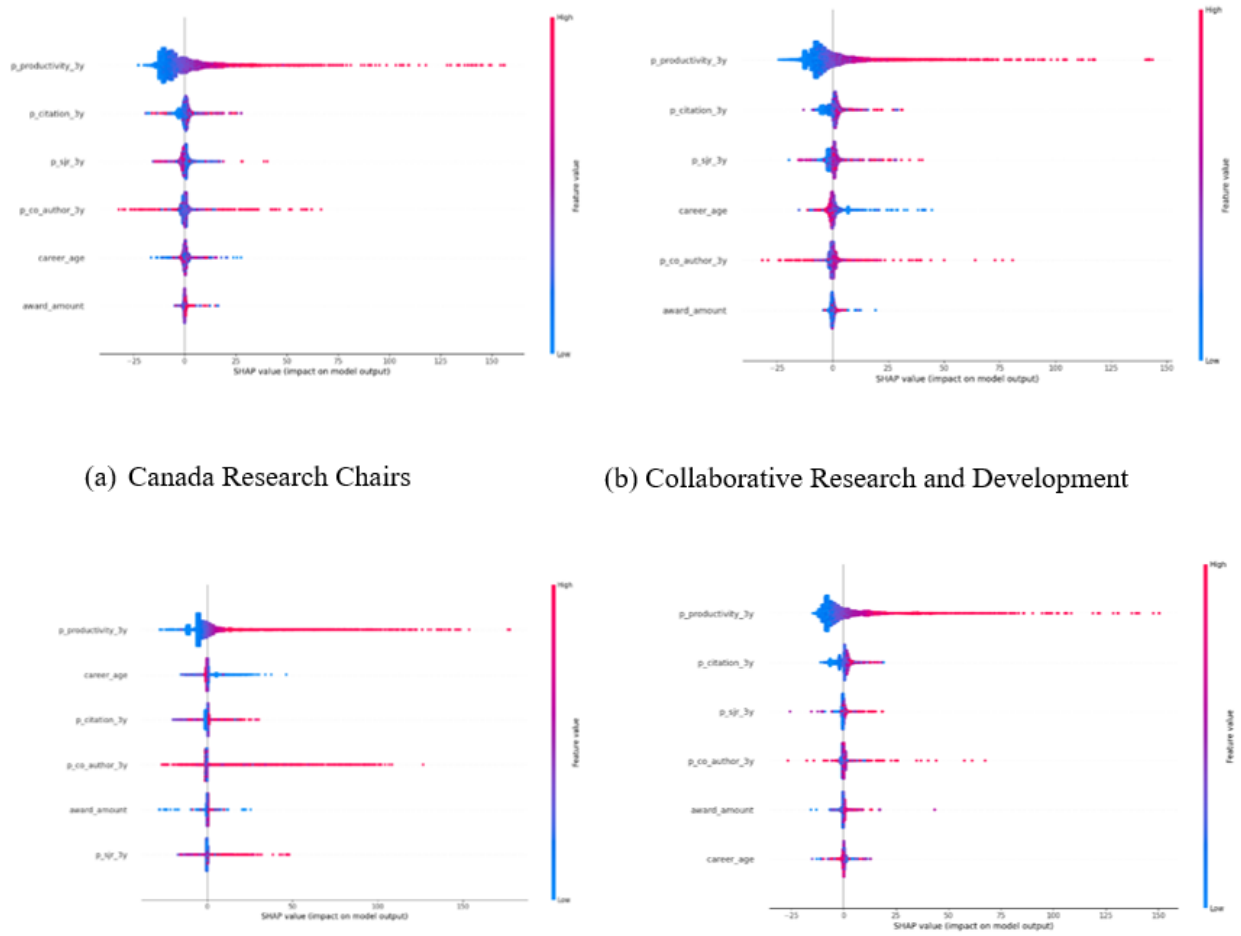
non-funded researchers or considering other funding agencies the impact of award amounts could be much higher.

Considering the type of the program, red points show that the researchers received the specified funding program while blue dots represent that the amount for the program was zero which indicate that the researcher did not receive the funding under this specific program. By analyzing the beeswarm plot, the importance of the Discovery Grant program is higher compared to other funding programs. For this program, most of the red points are located at zero SHAP value or small negative values while blue points are mostly on the positive side of the plot. This could show us that among the selected programs, the recipients of the Discovery Grants program are less productive compared to other programs. In fact, the most important thing about the selected funding program is that whether the researcher received it under Discovery Grants program or not. If they received the funding under other funding programs, it could positively affect their future productivity. As you can see, for the other three programs blue points are centered around zero SHAP value which indicate that not belonging to these program does not affect the outcome but receiving funding under these programs usually has a positive SHAP value.

In the next step of our analysis, we applied the best-performing model to distinct subsets of the dataset, each corresponding to a particular funding program. By doing so, we would like to analyze the significance of variables in forecasting the research productivity of researchers benefiting from the respective funding programs and compare the results for these programs. Across all funding programs, the model that consistently exhibited superior performance was the random forest. Employing this model, we conducted SHAP analysis to find out the importance of variables. The beeswarm plot related to each funding program is shown in the Figure 14.

In all four programs examined within this study, a consistent trend emerges, showing that the most important determinant of the researchers' future productivity is their past productivity as demonstrated by the model's performance across the entire dataset as well. Among these programs, except for the Discovery Grants program, the second most influential factor is the number of citations, followed by the SJR. This indicates the important role played by the quality of previous research in shaping future productivity across these programs. Nonetheless, this paradigm only partially holds true for the Discovery Grants program. For this program, citation counts are positioned as the third most influential factor, while SJR emerges as the least influential. This

distinction shows that in predicting the future productivity of researchers awarded funding through the Discovery Grants program, gaining attention from other researchers could be more crucial than publishing in highly prestigious journals.



(a) Canada Research Chairs (b) Collaborative Research and Development

(c) Discovery Grants Program – Individual (d) Strategic Projects - Group

Figure 14 Beeswarm plots of each program for future productivity

Considering the Discovery Grants program, the second important variable influencing future productivity is the career age of researchers. This finding suggests that young researchers who have been recipients of the Discovery Grants program have a higher likelihood of publishing more papers in the near future. In the context of the Collaborative Research and Development program, after researchers' past productivity and the quality of their work, career age is the subsequent factor in predicting their productivity. This implies that being in the early stages of one's career and securing funding through the CRD program can increase the probability of publishing more

articles. In contrast, for the Strategic Projects and Canada Research Chairs programs, the impact of career age on researcher productivity is less influential (sixth factor in Strategic Projects and fifth factor in Canada Research Chairs). Moreover, in these two programs, being a young researcher does not necessarily increase the likelihood of publishing more articles in the future as young researchers have both positive and negative SHAP values.

In the case of Canada Research Chairs and Collaborative Research and Development programs, it is evident that the variable of award amount has the least significance when predicting researchers' productivity. Similarly, when considering the Discovery Grants and Strategic Projects programs, the award amount assumes a position as the fifth variable out of six in terms of importance. This observation underscores the potential for award amount to have a greater impact on the productivity of researchers within these latter two programs in comparison to the former two. Notably, this effect is particularly pronounced within the Strategic Projects program, where higher award amounts are associated with notably higher SHAP values, indicating a greater influence on research productivity. For all the programs, collaboration represented by co-authorship has a moderate impact on the productivity ranking as the fourth or fifth important variable in predicting the future productivity.

7.2 Results on Future Impact Measured by SJR

The next phase of our research involves assessing the influence of various variables on the impact of researchers' future work. To achieve this objective, we chose to concentrate on the SJR of the journals in which researchers publish their articles three years after receiving funding through the chosen programs.

Similar to the examination of future productivity, our initial step involved conducting a 5-fold cross-validation multiple linear regression analysis on the training dataset, which served as the baseline for our analysis. Subsequently, we evaluated the model's performance on the test dataset. The F-statistic of the regression model yielded a value of 2.014e+04, with a corresponding p-value of 0. The R-squared coefficient, when applied to the training dataset, equals 0.435. The performance of the model for each fold within the cross-validated framework, along with its performance on the test dataset, is presented in Table 9.

	Mean Squared Error (MSE)	Root Mean Square Error (RMSE)	Mean Absolute Error (MAE)	R-Squared (R^2)
Fold 1	0.47	0.69	0.50	0.43
Fold 2	0.47	0.68	0.49	0.43
Fold 3	0.46	0.68	0.50	0.44
Fold 4	0.46	0.68	0.49	0.43
Fold 5	0.46	0.68	0.50	0.44
Test set	0.48	0.69	0.50	0.42

Table 9 Performance of linear regression on evaluation set for each part of the 5-fold cross-validation and test set for future SJR

Table 10 presents a comprehensive summary of the outcomes of our multiple linear regression analysis, with the SJR serving as the dependent variable. It is noteworthy that the p-values associated with all the variables in our model are notably small, indicating all the variables included in the model are statistically significant and different from zero. To maintain consistency throughout our analysis, we have selected the Canada Research Chairs program as the baseline against which we assess the impact of other funding programs.

After excluding Canada Research Chairs, the coefficients related to all the other 3 programs are negative. This implies that researchers who receive funding from these programs are less inclined to publish their articles in higher-quality journals when compared to those who have been granted funding through the Canada Research Chairs program. Notably, the negative magnitude of the coefficients for the Collaborative Research and Development and Strategic Projects programs is relatively higher, suggesting that the adverse impact of these programs on research quality is greater.

Considering the numerical variables, the average SJR of journals related to past articles has the highest coefficient, implying that the foremost factor influencing the publication of papers in reputable journals is the researchers' prior history of publishing in highly prestigious journals. The other variables, while still impactful, show relatively lower coefficients in comparison. Following past SJR, the next most influential variable is past productivity, indicating that researchers with a history of greater productivity are more likely to publish their articles in higher-quality journals. Subsequently, we have award amount, career age, citation counts, and co-authorship as

contributing factors. The coefficient associated with co-authorship is negative, indicating that collaborating with larger research groups might have an adverse effect on SJR and reduce the likelihood of publishing papers in higher-quality journals. In contrast, the coefficient for citation counts is relatively small but positive, suggesting that while citations have a positive impact on SJR, their influence is relatively small.

Variable	Coef.	Std. Err.	t	P > |t|	[95% Conf. Interval]	
Const.	1.0502	0.008	134.575	0.000	1.035	1.065
Award_amount	0.0642	0.002	31.679	0.000	0.060	0.068
Career_age	0.0476	0.001	32.108	0.000	0.045	0.050
P_productivity_3y	0.0861	0.002	53.636	0.000	0.083	0.089
P_co_author_3y	-0.0285	0.002	-18.402	0.000	-0.032	-0.025
P_sjr_3y	0.5458	0.001	370.619	0.000	0.543	0.459
P_citation_3y	0.0286	0.001	19.954	0.000	0.026	0.031
prgrm_Collaborative Research and Development Grants	-0.1067	0.009	-11.662	0.000	-0.125	-0.089
prgrm_Discovery Grants Program - Individual	-0.0208	0.008	-2.524	0.012	-0.037	-0.005
prgrm_Strategic Projects - Group	-0.1487	0.010	-14.711	0.000	-0.169	-0.129

Table 10 Summary of the results of the multiple linear regression for future SJR

In the next step, to select our final model, we fine-tuned the hyperparameters for both random forest and MLP neural network. Following the hyperparameter tuning process, we conducted a comparative analysis of the results obtained from the linear regression, random forest, and MLP neural network, using the predefined evaluation metrics. The summary of the models' performances on the test set is summarized in Table 11.

	MSE	RMSE	MAE	R-Squared
Linear regression	0.48	0.69	0.50	0.42
Random forest	0.37	0.61	0.44	0.56
MLP neural network	0.40	0.63	0.46	0.51

Table 11 Summary of the models' performance for future SJR on the test set

The performance results of the models clearly show that linear regression considerably underperforms compared to the other two models. While neural network exhibits performance close to that of random forest, it still falls slightly short of matching random forest for all the metrics. Consequently, we decided to proceed with random forest as our final model for conducting our analysis.

After performing random forest on the dataset, we used SHAP analysis as a powerful tool to understand the relative importance of the variables included in our model. The resulting beeswarm plot, illustrated in Figure 15, gives a visual representation of our findings. As expected, our analysis revealed that the quality of prior research publications measured by SJR and citation counts have a significant and direct influence on the quality of future research articles by recipients of the NSERC funding. Based on the beeswarm plot, a history of consistently producing high-quality research in terms of SJR and citation counts tended to lead to higher-quality articles in the future. Conversely, a history of lower quality works would decrease the probability of producing high-quality works in the future. Specifically, we observed that the previous SJR emerged as the most critical factor in predicting the quality of journals in which researchers are likely to publish their future works. This indicates the role of an author's track record, as reflected in their previous SJR, in shaping the quality of their subsequent research. Furthermore, citation counts while not as important as previous SJR, still held substantial importance as the third most influential factor. This finding was expected as a researcher who produced high-quality articles in the past is likely that their future publications have higher SJR.

The second significant factor in predicting the quality of future research works is the history of co-authorship, implying the important role of collaboration in the quality of future works. As illustrated by the SHAP values, we found that a history of low collaboration in the past has a relatively minor negative impact on the quality of future works. On the contrary, high levels of collaboration either have a highly positive impact, boosting the quality of future research or have

a significant negative impact, potentially diminishing the quality of future work. This dynamic indicates the crucial role played by the quality and synergy of a research group or collaborators in determining the outcomes of research. The positive impact of collaboration on the quality is in line with some other studies in the literature who also found that engagement in larger groups could enhance the quality of publications (Wang & Shapira, 2015; Yan, et al., 2018; Ebadi & Schiffauerova, 2016; Zhao, 2010).

Moreover, we identified the past productivity of researchers as the fourth significant factor influencing the quality of future research works. This factor demonstrated a moderate impact on the outcomes. Specifically, we observed that high past productivity has a positive influence, contributing to the enhancement of the quality of future work. In other words, productive researchers tend to publish higher quality papers. Additionally, the award amount and career age carried a lower weight compared to the other variables in predicting future research quality. Career age, while relatively less impactful compared to other variables, indicated an interesting trend. It suggests that older researchers are more likely to produce higher-quality research compared to their younger peers. This finding shows the potential benefits of experience and accumulated knowledge in contributing to the excellence of research. The finding is partially in line with Ebadi & Schiffauerova (2016) who found that mid-career researchers are more likely to be more productive.

Considering the type of funding programs, our analysis shows that Collaborative Research and Development funding has the most substantial impact on the quality of research, indicating the importance of collaborative efforts in enhancing research quality. Following Collaborative Research and Development, we observed that Discovery Grants, Canada Research Chairs, and Strategic Projects, in that order, are important for predicting the research quality. To delve deeper into how receiving these different types of funding programs affects the quality of research, in the next section, we run our model on subsets of the dataset, each containing researchers who have received one of these four types of funding programs and compare the results.

As previously discussed, we applied the random forest model, which consistently outperformed other models, to different subsets of the dataset, each containing the data to a particular funding program. This approach allowed us to find out the significance of variables in forecasting the

quality of researchers' work within the context of the respective funding programs and make a fair comparison of programs. To do so, we used SHAP analysis.

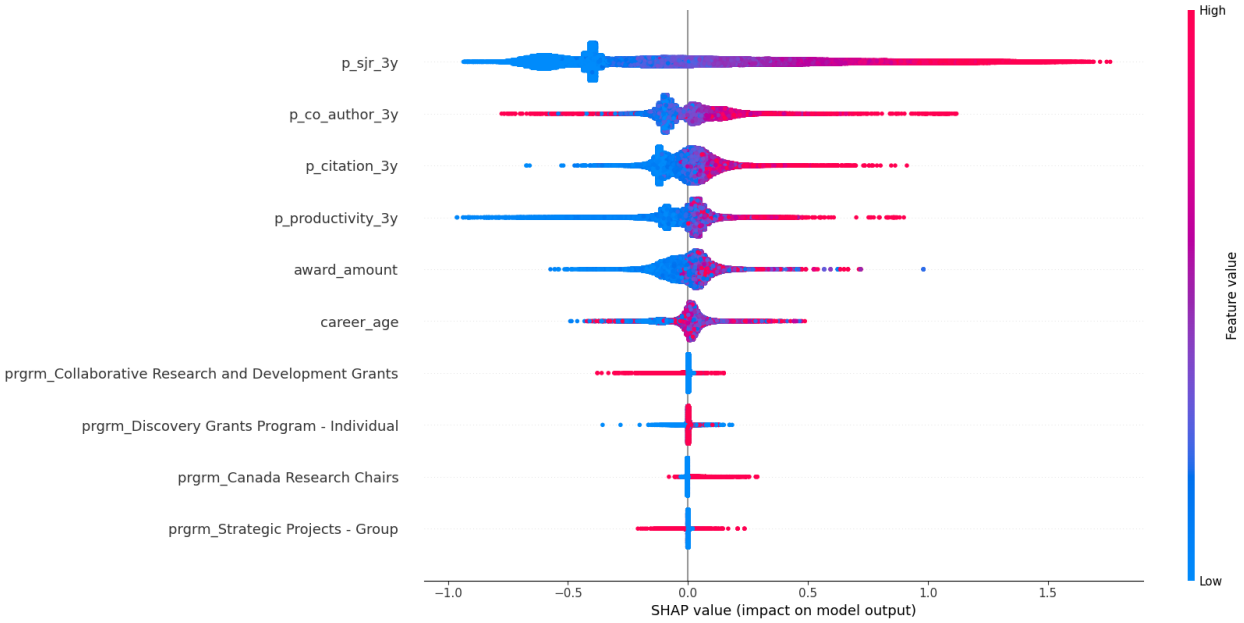


Figure 15 Beeswarm plot for future SJR – Entire dataset

As shown in Figure 16, it is evident that the average past SJR of researchers emerges as the predominant factor in forecasting future SJR across all four programs. This observation reveals that researchers who have a history of publishing articles in reputable journals have a higher tendency to continue contributing to higher-quality journals in the future.

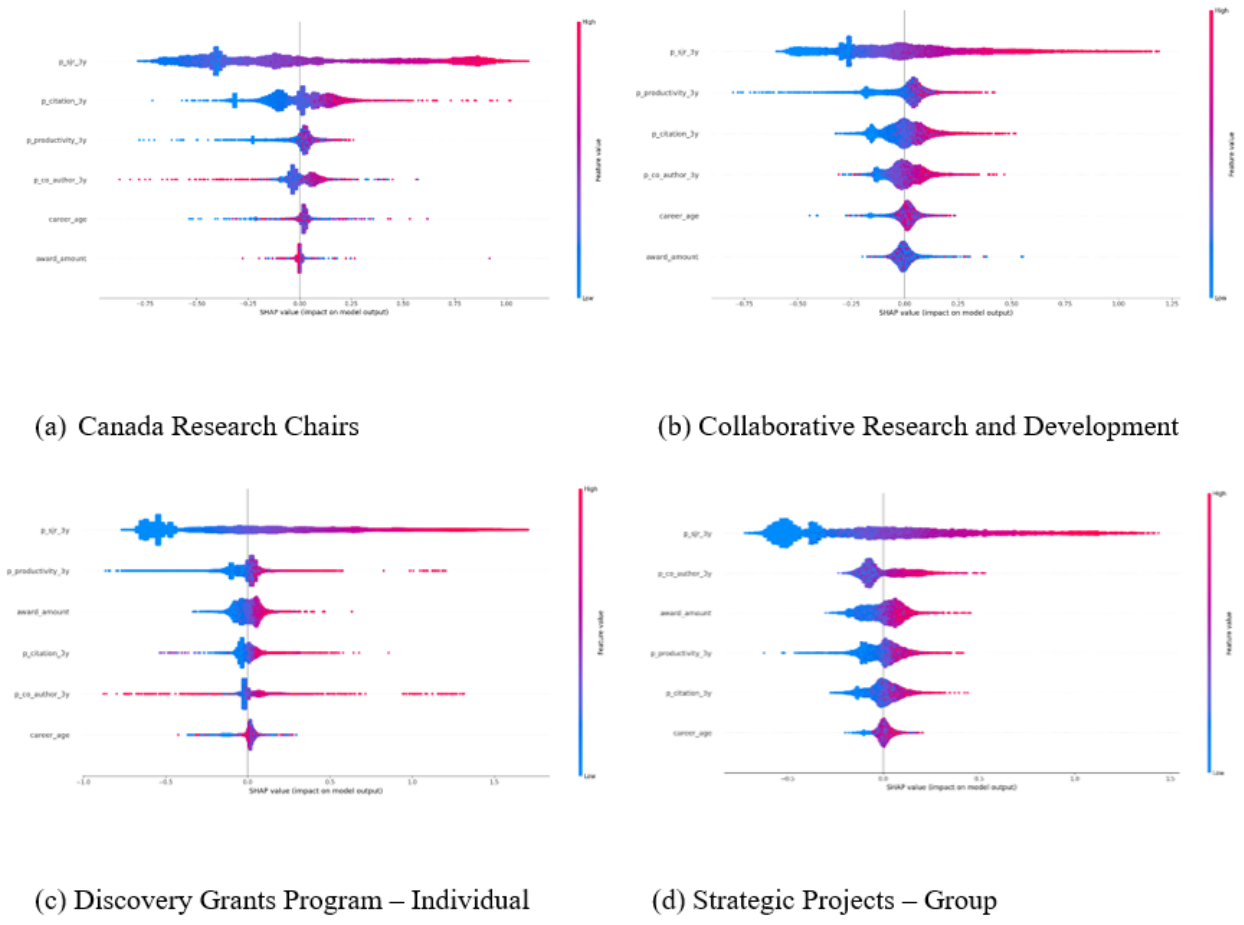
Among the analyzed programs, Canada Research Chairs and Collaborative Research and Development show a similar pattern regarding the importance of variables in predicting future SJR. The variable importance order is similar for these two programs, with a slight difference in the rankings. For Canada Research Chairs, citation counts hold the second-most influential position, followed by productivity in the third place. Conversely, for Collaborative Research and Development, productivity takes the second spot, and citation counts rank third. This suggests that for the individuals receiving funding through these programs, past productivity and quality significantly influence the prediction of future SJR. In other words, being productive and

maintaining a high level of quality in previous work would likely lead to future publications in journals with higher SJR. Co-authorship emerges as the next influential variable for both programs, but its impact varies. In the case of Canada Research Chairs, a history of collaboration with larger research groups could have a notably negative impact on future SJR, but for the Collaborative Research and Development program, this impact is positive. Lastly, for both programs, career age and award amount are the least important factors, ranking fifth and sixth in significance. These findings underscore that within the context of these programs factors such as career age and the amount of funding received play a relatively minor role in predicting future SJR compared to other variables.

For the Discovery Grants program, past productivity is the second most influential variable highlighting that being productive in the past would significantly impact the future SJR for recipients of this funding. In contrast, for the Strategic Projects program, past productivity holds a moderate impact, ranking fourth in terms of variable importance. This suggests that the quality of future work by recipients of this funding is less dependent on their past productivity. This difference could be attributed to the nature of the research conducted under the Strategic Projects program, which may focus on specific areas of interest defined by NSERC. Thus, the impact of past productivity on future SJR may be mitigated by the program's specific objectives. Interestingly, for both the Discovery Grants and Strategic Projects programs, the award amount is identified as the third most influential factor. This indicates that higher levels of funding provided through these two programs play a crucial role in facilitating the production of better-quality work in the future. In other words, a higher funding allocation can positively influence SJR.

Considering the Discovery Grants program, citation counts are identified as the fourth in the variable importance ranking affecting future SJR. This suggests that the extent to which researchers have gained attention from their peers in the past has a moderate impact on their ability to publish future works in higher-quality journals. While citations are a relevant metric, they are not as critical as other factors in predicting SJR for recipients of this program. Co-authorship and career age are ranked fifth and sixth in importance, indicating that they have the least impact among our variables on future SJR outcomes within the Discovery Grants program. Collaborating with large research groups may not necessarily have a positive impact and could even reduce the quality of publications in this program, as suggested by red points on the left side of the graph. In

contrast, for the Strategic Projects program, co-authorship emerges as the second most important variable, highlighting that working with larger groups can be beneficial for recipients of this program when aiming to produce high-quality works. Citation counts and career age have the least impact on future SJR within the context of the Strategic Projects program, suggesting that other factors play a more significant role in determining the quality and impact of research outcomes for these recipients.



(a) Canada Research Chairs

(b) Collaborative Research and Development

(c) Discovery Grants Program – Individual

(d) Strategic Projects – Group

Figure 16 Beeswarm plots of each program for future SJR

7.3 Results on Future Collaboration

The last part of our research was to assess the impact of different funding programs on collaboration measured by the average number of authors contributing to articles. Like our approach for productivity and research impact, we considered a 3-year window to calculate the average number of authors who collaborated on articles with researchers who received NSERC funding through the selected programs.

As the first step, we conducted a multiple linear regression on the dataset, serving as the foundational baseline for our study. The F-statistic associated with the regression model is 8088, with a corresponding p-value of 0, highlighting the statistical significance of the model. Nevertheless, it is important to note that the R-squared obtained by the model on the training dataset is only 0.241, indicating that the linear regression model is performing quite poorly for our dataset. The performance of the model for each fold of cross-validation, as well as its performance on the test dataset, is shown in Table 12.

	Mean Squared Error (MSE)	Root Mean Square Error (RMSE)	Mean Absolute Error (MAE)	R-Squared (R^2)
Fold 1	2.17	1.47	1.16	0.24
Fold 2	2.15	1.46	1.15	0.24
Fold 3	2.14	1.46	1.16	0.24
Fold 4	2.19	1.48	1.16	0.23
Fold 5	2.15	1.47	1.15	0.23
Test set	2.15	1.47	1.15	0.23

Table 12 Performance of linear regression on evaluation set for each part of the 5-fold cross-validation and test set for future co-authorship

Table 13 provides a comprehensive summary of the outcomes stemming from our multiple linear regression analysis, with co-authorship as the dependent variable. Except for the dummy variable "prgrm_Discovery Grants Program – Individual", p-values of all the other variables are small which shows that these variables in our model are statistically significant and are different from zero. Like the other sections of our study, we selected the Canada Research Chairs program as the baseline against which we assess the impact of other funding programs as dummy variables.

By considering Canada Research Chairs as the baseline, the coefficients associated with the remaining three funding programs are positive. This implies that researchers who secure funding through these programs are more likely to collaborate with a larger number of researchers when producing their research articles, as compared to the ones who granted Canada Research Chairs. Specifically, the coefficients for the Strategic Projects and Collaborative Research and Development Grants have relatively higher positive magnitude, suggesting that recipients of these fundings have a greater chance for collaboration. On the other hand, the p-value corresponding to

the Discovery Grants program is relatively high. This suggests that the impact of this program on collaboration is not statistically significant and different from zero.

Variable	Coef.	Std. Err.	t	P > |t|	[95% Conf. Interval]	
Const.	2.4424	0. 017	142.387	0.000	2.409	2.476
Award_amount	0.0782	0.004	17.564	0.000	0.069	0.087
Career_age	0.1790	0.003	55.086	0.000	0.173	0.185
P_productivity_3y	0.3970	0.003	117.845	0.000	0.390	0.404
P_co_author_3y	-0.1540	0.003	-49.253	0.000	-0.160	-0.148
P_sjr_3y	0.5472	0.003	170.364	0.000	0.541	0.553
P_citation_3y	0.0406	0.003	13.021	0.000	0.034	0.047
prgrm_Collaborative Research and Development Grants	0.1634	0.020	8.138	0.000	0.124	0.203
prgrm_Discovery Grants Program - Individual	0.0294	0.018	1.625	0.104	-0.006	0.065
prgrm_Strategic Projects - Group	0.1824	0.022	8.278	0.000	0.139	0.226

Table 13 Summary of the results of the multiple linear regression for future co-authorship

Among the numeric variables, the coefficient of past collaboration is negative showing that the collaboration in the past has a negative impact on the future co-authorship of the researchers. Conversely, all the other variables have positive coefficients suggesting a positive impact on the average number of authors collaborating with the recipient of the funding in the next 3 years. Past SJR and past productivity have the highest coefficients implying the importance of publishing articles in prestigious journals and maintaining high research productivity in facilitating future collaborations. Additionally, career age has a relatively high coefficient suggesting that more experienced researchers can leverage their networks to form teams and publish jointly with their peers. Award amount also shows a small positive impact, indicating that increased funding can foster collaboration. The smallest coefficient belongs to past citation counts which shows that prior

recognition from other researchers may not necessarily lead to increased collaboration in future publications.

To select the final model, a comprehensive hyperparameter tuning process was conducted for both the random forest and neural Network. After hyperparameter tuning, we compared the results of the linear regression, random forest, and MLP neural network based on the evaluation metrics that we have defined. The summary of the models' performances on the test set is summarized in the table 14.

	MSE	RMSE	MAE	R-Squared
Linear regression	2.15	1.47	1.15	0.23
Random forest	1.30	1.14	0.83	0.55
MLP neural network	1.31	1.14	0.84	0.53

Table 14 Summary of the models' performance for future co-authorship on the test set

Based on the results of the model performances, it is evident that linear regression has a poor performance for our dataset. On the other hand, the random forest and MLP neural network models demonstrated competitive performance, with a marginal difference in favor of the random forest model. Specifically, random forest outperforms MLP neural network in three out of four metrics, with the exception being RMSE (Root Mean Square Error), where both models perform equally. Therefore, based on our evaluation of the models and considering the small performance advantages observed in favor of the random forest, we selected the random forest as our final model.

Like previous sections, after deciding about the model, we analyzed the importance of variables leveraging on SHAP analysis. As depicted in Figure 17, the most influential factor in predicting future collaboration is the history of past co-authorship among researchers. Our findings indicate that being a researcher with no co-authors or only a small number of collaborators has a negative impact on their future collaboration. Interestingly, we observed a complex trend among researchers who had a history of collaborating in large groups. While, in most cases, such researchers are more likely to continue collaborating in large groups in the future, a substantial number of researchers experienced a shift in their collaboration pattern after involvement in large groups. This is an interesting result since Some researchers who investigated the impact of funding on shaping collaborations among researchers found a positive impact of funding on co-authorships of

publication (Zhao, 2010; Ebadi & Schiffauerova, 2015; Alvarez-Bornstein & Bordons, 2021). However, our findings suggest that while many NSERC funded researchers continue working in large groups, some may decide to pursue their careers in smaller groups, possibly driven by various factors like research interests, personal preferences, etc.

The next influential factor is the past productivity of researchers. As shown in the beeswarm graph, low productivity has a substantial negative impact on the future collaboration of researchers. On the other hand, being productive has a moderate effect suggesting that productivity is indeed an essential factor for forming research groups but not the sole determinant. Quality of the previous works in terms of both SJR and citation counts is the next factor influencing the collaboration. Researchers with higher-quality prior publications have a greater likelihood of participating in larger research groups in future. This shows the link between research excellence and the potential for engaging in research collaborations, likely driven by the appeal of collaborating with accomplished researchers. Lastly, the award amount and career age are identified as the least influential factors in predicting future collaboration. Higher funding amounts are associated with a positive impact on forming collaborations indicating that funding resources can facilitate collaborative efforts. Additionally, younger researchers exhibited a higher tendency to work in large research groups, potentially reflecting a desire to gain experience and build networks early in their careers. These findings are line with Ebadi & Schiffauerova (2015) who also found that while funding has a positive impact on the co-authorship, career age of researchers negatively affects their collaboration in the future.

In terms of the type of funding programs, Discovery Grants appear to be the most significant factor influencing future research collaboration. Among the selected funding programs, Discovery Grants program holds a unique position in predicting research collaboration. Specifically, when examining the blue points representing researchers who did not receive Discovery Grants, it becomes evident that the absence of this funding program does not have a substantial impact on future collaboration. However, a distinct trend emerges among researchers who received funding through other programs. The recipients of other funding programs are more likely to engage in collaborative research and form research groups.

In the final step of our analysis, we applied the best-performing model to subsets of the dataset, each containing data points related to a particular funding program. This allowed us to gain insights

into how researchers who receive funding from different programs exhibit distinct behaviors when it comes to forming research groups, as proxied by their co-authorship patterns in research articles. To delve deeper into these behavioral differences, we conducted SHAP analysis for each funding program. This allowed us to understand the importance of factors that contribute to the formation of research groups among researchers in these funding programs.

Across all four funding programs, the most influential factor in co-authorship patterns is the researchers' history of co-authorship in the past three years. However, it is noteworthy that the impact varies, particularly for the Discovery Grants program. In the other three programs, working with large research groups in the past is associated with a significant positive impact on future co-authorship. This suggests that researchers who collaborated with larger groups are more likely to continue doing so in the future when they receive funding from these programs. Interestingly, for the Discovery Grants program, while a history of working with large groups generally has a significant positive impact on future co-authorship, there is a notable trend. Many researchers who previously worked in large groups started to work in smaller groups after receiving Discovery Grants. This finding suggests that the nature of research collaborations may change for some researchers within the Discovery Grants program, potentially leading to more independent research efforts despite their prior experience with larger groups.

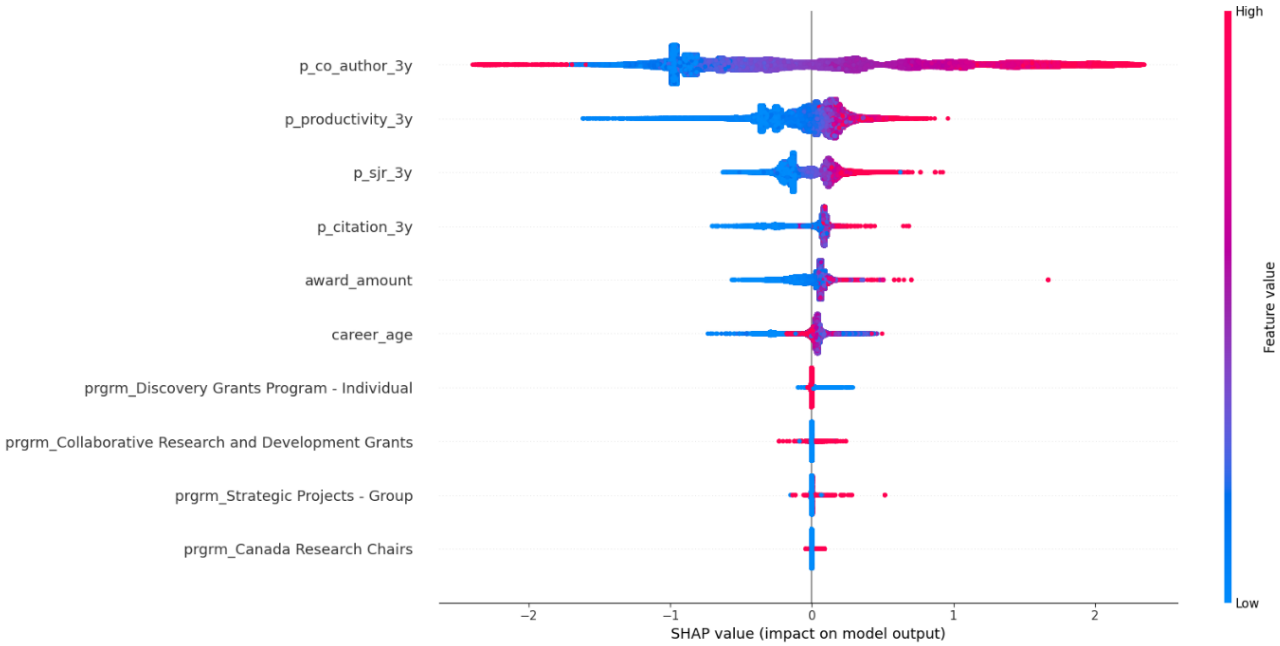
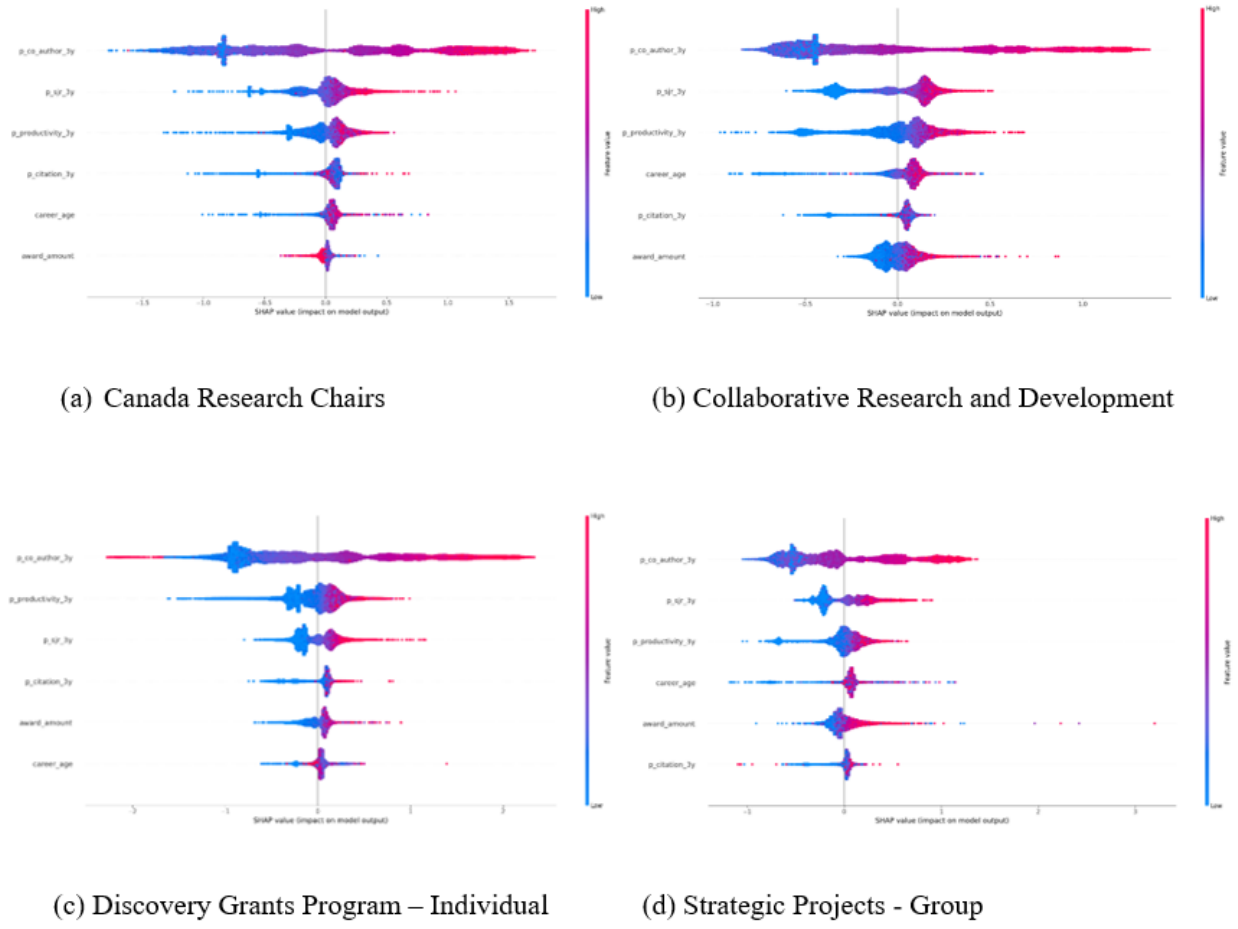


Figure 17 Beeswarm plot for future co-authorship – Entire dataset

For all four funding programs, past productivity and a history of publishing in high-quality journals emerge as the next two important factors shaping future co-authorship patterns. However, there is a distinction for the Discovery Grants program. In other programs, past SJR takes the second influential position, followed by past productivity in the third place. In the Discovery Grants program, past productivity takes the second position and past SJR ranks third. This unique behavior among Discovery Grants recipients suggests that their collaborative patterns are distinct from those in the other programs. It indicates that, for researchers who receive Discovery Grants, their past productivity plays a more critical role in shaping their future co-authorship networks compared to the quality of journals in which they have previously published.

Among the independent variables considered, citation counts and career age exhibit a moderate to low impact, ranking from fourth to sixth in importance depending on the type of funding program. For Canada Research Chairs and Discovery Grants, where citation counts are the fourth most important factor, having a high number of citations has a noticeable positive impact on future co-authorship. This suggests that researchers with a strong citation record are more likely to engage in collaborative research efforts within these programs. In the case of Collaborative Research and Development, citation counts are the fifth most important factor and in the Strategic Projects program, citation counts rank as the least important factor suggesting that this variable has limited impact on future collaboration patterns within these two programs.



(a) Canada Research Chairs

(b) Collaborative Research and Development

(c) Discovery Grants Program – Individual

(d) Strategic Projects - Group

Figure 18 Beeswarm plots of each program for future co-authorship

Lastly, award amount is the least influential factor for the Canada Research Chairs and Collaborative Research and Development programs, while for the other two programs award amount ranks as the fifth important variable out of six. For all the programs except Canada Research Chairs this aligns with the conventional understanding that increased financial resources tend to facilitate and encourage collaborative research efforts among researchers. In these programs, researchers who receive larger funding allocations are more likely to engage in collaborative research endeavors, possibly due to the enhanced resources at their disposal. However, a unique behavior within the Canada Research Chairs program is observed. In this program, higher award amounts are associated with lower collaboration in the future. This finding suggests that within the Canada Research Chairs program researchers who receive larger funding allocations may have a preference for more independent research pursuits.

	Canada Research Chairs	Collaborative Research and Development	Discovery Grants Program – Individual	Strategic Projects - Group
Future Productivity	1. Past Productivity 2. Past Citation counts 3. Past SJR 4. Past Co-authorship 5. Career Age 6. Award Amount	1. Past Productivity 2. Past Citation counts 3. Past SJR 4. Career Age 5. Past Co-authorship 6. Award Amount	1. Past Productivity 2. Career Age 3. Past Citation counts 4. Past Co-authorship 5. Award Amount 6. Past SJR	1. Past Productivity 2. Past Citation counts 3. Past SJR 4. Past Co-authorship 5. Award Amount 6. Career Age
Future SJR	1. Past SJR 2. Past Citation counts 3. Past Productivity 4. Past Co-authorship 5. Career Age 6. Award Amount	1. Past SJR 2. Past Productivity 3. Past Citation counts 4. Past Co-authorship 5. Career Age 6. Award Amount	1. Past SJR 2. Past Productivity 3. Award Amount 4. Past Citation counts 5. Past Co-authorship 6. Career Age	1. Past SJR 2. Past Co-authorship 3. Award Amount 4. Past Productivity 5. Past Citation counts 6. Career Age
Future Co-authorship	1. Past Co-authorship 2. Past SJR 3. Past Productivity 4. Past Citation counts 5. Career Age 6. Award Amount	1. Past Co-authorship 2. Past SJR 3. Past Productivity 4. Career Age 5. Past Citation counts 6. Award Amount	1. Past Co-authorship 2. Past SJR 3. Past Productivity 4. Past Citation counts 5. Career Age 6. Award Amount	1. Past Co-authorship 2. Past SJR 3. Past Productivity 4. Career Age 5. Award Amount 6. Past Citation counts

Table 15 Summary of variable importance for each of the dependent variables in different funding programs

8. Conclusion

In this study, we examined how receiving grants through different funding programs offered by NSERC can affect the outcomes of researchers in terms of their future productivity, the quality of their work, and their collaboration. Specifically, our research focused on understanding the effectiveness of NSERC funding programs, each pursuing distinct strategies. To do so, we closely investigated how different factors affect the outcomes of NSERC funding programs. By thoroughly exploring the connection between funding strategies and program outcomes, we aimed to facilitate informed decision-making and fostering the development of more efficient NSERC initiatives.

In our analysis, we selected four major programs introduced by NSERC: Discovery Grants Program – Individual, Canada Research Chairs, Collaborative Research and Development, and Strategic Projects - Group. We initially conducted 5-fold cross-validation multiple linear regression, random forest, and MLP neural network on our dataset. This step aimed to determine the most effective model for our data. After thoroughly preprocessing and cleaning the dataset and

fine-tuning the hyperparameters, we observed consistent results across all combinations of funding programs and dependent variables. The random forest model consistently outperformed the other two models. Consequently, we made the decision to utilize the random forest for all subsequent analyses.

In the next phase of our research, our focus shifted towards determining the importance of independent variables in predicting the three dependent variables for each funding program. Our research uncovers key distinctions among various funding programs.

For Canada Research Chairs recipients, the quality of their previous work, measured by SJR and citation metrics, holds greater importance in shaping research outcomes compared to other programs. This highlights the unique emphasis on research excellence within the Canada Research Chairs. In contrast, for recipients of Canada Research Chairs, career age has a diminished impact compared to other programs. While career age typically has a moderate to low impact, it surprisingly becomes influential in predicting future productivity within the Discovery Grants program. Specifically, providing funding to young researchers in this program is associated with a higher likelihood of increased article production. Moreover, researchers who have been granted Discovery Grants Program and have been involved in large group collaborations exhibit an intriguing exception. While these collaborations generally shape future co-authorship positively, within the Discovery Grants some researchers experience a negative impact on their future collaborations.

Expanding our focus across various programs, we found that the award amount holds greater importance in shaping the research outcomes of recipients involved in strategic projects. This highlights the influence of funding levels for recipients of Strategic Projects. Surprisingly, our analysis reveals a nuanced relationship between SJR and citation counts. When forecasting future SJR, citation counts exhibit a moderate impact across all programs, except for Canada Research Chairs, where they rank as the second most important factor after past SJR. This suggests that in these programs the direct link between publishing in higher journals and receiving citations is not evident.

To the best of my knowledge, this study is the first comprehensive examination of the influence of different funding programs on research outcomes. As discussed, the prediction patterns for future productivity, research quality, and collaboration differ among recipients of different funding

programs. These findings suggest the importance of tailoring the allocation of specific funding programs to researchers who align with the program's objectives and requirements.

9. Limitations and future works

In this research, we studied the patterns within four major NSERC funding programs and examined how receiving these funding programs could influence the research outcomes of recipients. The exclusive focus on the NSERC limits the generalizability of the results as different funding agencies may have distinct program structures and objectives. To expand our understanding, future research endeavors could explore the impact of receiving funding in areas other than natural science and engineering, from other funding agencies, and under programs with different objectives.

Another limitation of our research involves the matching process of the NSERC and Scopus datasets. In the Scopus dataset, the availability of only the last name and the first letter of the first name of researchers introduced a potential challenge for accurate matching. However, we attempted to improve matching accuracy by considering researchers' affiliations as well.

In examining collaboration, we focused on co-authorship. Therefore, we were not able to understand the specific contribution of each researcher to the articles. This simplification may overlook individual researchers' roles within collaborative projects, potentially influencing the interpretation of collaboration's impact on research outcomes.

Finally, considering other input variables could be helpful in predicting the research outcomes more precisely and provide us with a better understanding of the relationship between the funding programs and research productivity, quality, and collaboration.

Bibliography

Alvarez-Bornstein, B. & Bordons, M., 2021. Is funding related to higher research impact? Exploring its relationship and the mediating role of collaboration in several disciplines. *Journal of Informetrics,* 15(1).

Boyack, K. & Börner, K., 2003. Indicator-assisted evaluation and funding of research: Visualizing the influence of grants on the number and citation counts of research papers. *Journal of the American Society for Information Science and Technology,* 54(5), pp. 447 - 461.

Campbell, D. et al., 2010. Bibliometrics as a performance measurement tool for research evaluation: The case of research funded by the national cancer institute of Canada. *American Journal of Evaluation,* 31(1), pp. 66 - 83.

Chairs, C. R., 2023. *Canada Research Chairs.* [Online]
Available at: https://www.chairs-chaires.gc.ca/home-accueil-eng.aspx
[Accessed 13 11 2023].

De Solla Price, D., 1963. *Little Science, Big Science..* New York: Columbia University Press.

Ebadi, A. & Schiffauerova, A., 2013. Impact of funding on scientific output and collaboration: A survey of literature. *Journal of Information and Knowledge Management,* 12(4).

Ebadi, A. & Schiffauerova, A., 2015. How to become an important player in scientific collaboration networks?. *Journal of Informetrics,* 9(4), pp. 809 - 825.

Ebadi, A. & Schiffauerova, A., 2015. How to receive more funding for your research? get connected to the right people. *PLoS ONE,* 10(7).

Ebadi, A. & Schiffauerova, A., 2016. How to boost scientific production? A statistical analysis of research funding and other influencing factors. *Scientometrics,* 106(3), pp. 1093 - 1116.

Geuna, A. & Martin, B., 2003. University research evaluation and funding: An international comparison. *Minerva,* 41(4), pp. 277 - 304.

Godin, B., 2003. The impact of research grants on the productivity and quality of scienti¯c research. (No. 2003). INRS Working Paper..

Goh, A., 1995. Back-propagation neural networks for modeling complex systems. *Artificial Intelligence in Engineering,* 9(3), pp. 143 - 151.

Gök, A., Rigby, J. & Shapira, P., 2016. The impact of research funding on scientific outputs: Evidence from six smaller European countries. *Journal of the Association for Information Science and Technology,* 67(3), pp. 715 - 730.

Hagan, M., Demuth, H., Beale, M. & De Jesús, 2014. 2 ed. s.l.:Martin Hagan.

Hastie, T., Tibshirani, R. & Friedman, J. H., 2009. *The elements of statistical learning: data mining, inference, and prediction.* second ed. New York: Springer.

Huang, M.-H. & Huang, M.-J., 2018. An analysis of global research funding from subject field and funding agencies perspectives in the G9 countries. *Scientometrics,* 115(2), pp. 833 - 847.

James, G., Witten, D., Hastie, T. & Tibshirani, R., 2013. *An Introduction to Statistical Learning : with Applications in R.* New York: Springer.

Katz, J. S. & Martin, B. R., 1997. What is research collaboration?. *Research Policy,* 26(1), pp. 1 - 18.

Kůrková, V., 1992. Kolmogorov's theorem and multilayer neural networks. *Neural Networks,* 5(3), pp. 501 - 506.

Lewison, G. & Dawson, G., 1998. The effect of funding on the outputs of biomedical research. *Scientometrics,* 1-2(41), pp. 17 - 27.

Lundberg, S. & Lee, S.-I., 2017. *A unified approach to interpreting model predictions.* Long Beach, Advances in Neural Information Processing Systems.

Muller, A. & Guido, S., 2018. *Introduction to Machine Learning with Python: A Guide for Data Scientists.* illustrated, reprint ed. s.l.:O'Reilly Media.

Muscio, A., Quaglione, D. & Vallanti, G., 2013. Does government funding complement or substitute private research funding to universities?. *Research Policy,* 42(1), pp. 63 - 75.

NSERC, 2022. *NSERC.* [Online]
Available at: https://www.nserc-crsng.gc.ca/professors-professeurs/rpp-pp/crd-rdc_eng.asp
[Accessed 13 11 2023].

NSERC, 2022. *NSERC.* [Online]
Available at: https://www.nserc-crsng.gc.ca/professors-professeurs/rpp-pp/spg-sps_eng.asp
[Accessed 13 11 2023].

NSERC, 2023. *NSERC.* [Online]
Available at: https://www.nserc-crsng.gc.ca/index_eng.asp
[Accessed 12 07 2023].

NSERC, 2023. *NSERC.* [Online]
Available at: https://www.nserc-crsng.gc.ca/professors-professeurs/grants-subs/dgigp-psigp_eng.asp
[Accessed 13 11 2023].

Okubo, Y., 1997. "Bibliometric Indicators and Analysis of Research Systems: Methods and Examples", OECD Science, Technology and Industry Working Papers, No. 1997/01, OECD Publishing, Paris, https://doi.org/10.1787/208277770603.. In: s.l.:s.n.

Payne, A. A. & Siow, A., 2003. Does federal research funding increase university research output?. *Advances in Economic Analysis and Policy,* 3(1).

SCImago, 2023. *SJR — SCImago Journal & Country Rank.* [Online]
Available at: http://www.scimagojr.com
[Accessed 13 04 2023].

Tahmooresnejad, L., Beaudry, C. & Schiffauerova, A., 2015. The role of public funding in nanotechnology scientific production: Where Canada stands in comparison to the United States. *Scientometrics,* 102(1), pp. 753 - 787.

Tijssen, R., 2004. Is the commercialisation of scientific research affecting the production of public knowledge? Global trends in the output of corporate research articles. *Research Policy,* 33(5), pp. 709 - 733.

Ubfal, D. & Maffioli, A., 2011. The impact of funding on research collaboration: Evidence from a developing country. *Research Policy,* 40(9), pp. 1269 - 1279.

Veletanlic, E. & Sa, C., 2020. Implementing the Innovation Agenda: A Study of Change at a Research Funding Agency. *Minerva,* 58(2), pp. 261 - 283.

Vercellis, C., 2009. *Business Intelligence: Data Mining and Optimization for Decision Making.* s.l.:Wiley Publishing.

Wang, J. & Shapira, P., 2015. Is there a relationship between research sponsorship and publication impact? An analysis of funding acknowledgments in nanotechnology papers. *PLoS ONE,* 10(2).

Yan, E., Wu, C. & Song, M., 2018. The funding factor: a cross-disciplinary examination of the association between research funding and citation impact. *Scientometrics,* 115(1), pp. 369 - 384.

Zhao, D., 2010. Characteristics and impact of grant-funded research: A case study of the library and information science field. *Scientometrics,* 84(2), pp. 293 - 306.