MEASURING RISK IN SCIENCE

Deyun YIN [a,b], Zhao WU [a], and Sotaro SHIBAYAMA [c,d,*]

[a] Harbin Institute of Technology (Shenzhen), School of Economics and Management, Shenzhen, China
[b] World Intellectual Property Organization, Geneva, Switzerland
[c] Lund University, School of Economics and Management, Lund, Sweden
[d] The University of Tokyo, Institute for Future Initiative, Tokyo, Japan

[*] Corresponding author: sotaro.shibayama@fek.lu.se. +46 (0)46 2227812.
P.O. Box 7080, S-220 07 Lund, Sweden

ABSTRACT

Risk plays a fundamental role in scientific discoveries, and thus it is critical that the level of risk can be systematically quantified. Knowledge recombination is an important route to generating new knowledge, but it often fails. We propose a novel approach to measuring *risk* involved in this discovery process. Drawing on machine learning and natural language processing techniques, our approach converts knowledge elements in the text format into high-dimensional vector expressions and computes the probability of failing to combine a pair of knowledge elements. Testing the calculated risk indicator on survey data, we confirm that our indicator is correlated with self-assessed risk. Further, as risk and novelty have been confounded in the literature, we examine and suggest the divergence of the bibliometric novelty and risk indicators. Finally, we demonstrate that our risk indicator is negatively associated with future citation impact, suggesting that risk-taking itself may not necessarily pay off. Our approach can assist decision making of scientists and relevant parties such as policymakers, funding bodies, and R&D managers.

KEYWORDS

Risk; Uncertainty; Novelty; Recombination; Science; Word embedding; Support Vector Machine

1. INTRODUCTION

Science is a risky business by nature. Such risk and uncertainty tend to be especially high when scientists aim at novel discoveries (Bourdieu, 1975; Merton, 1973). Thus, there is a growing concern over scientists' risk-averse behavioral patterns, and science communities and policymakers emphasize that efforts should be made to facilitate high-risk-high-return research (Franzoni and Stephan, 2021; Gewin, 2012; Machado, 2021; OECD, 2021).

Despite its fundamental role, risk and uncertainty in science have been poorly understood (Althaus, 2005; Aven, 2011; Franzoni and Stephan, 2021; Hansson, 2018), which this study aims to contribute to. Specifically, we aim to develop a bibliometric approach to quantify the degree of scientific risk in a particular mode of scientific discovery process – recombination, which is an indispensable route to generate new knowledge (Fontana et al., 2020; Uzzi et al., 2013). The previous literature seems to make an assumption that novel research is risky (Machado, 2021; Reinhilde et al., 2022). While novel research may entail some risks (Franzoni and Stephan, 2021; Wang et al., 2017), risk and novelty are not equivalent. Opportunities for novel recombination may be difficult to identify but may be easily achieved once the opportunity is identified.

To quantify risk in the recombination process, we employ machine learning and natural language processing techniques. Drawing on past trajectories of science, we develop a machine learning model that predicts whether a certain pair of knowledge elements will be linked or not in the future. The developed model calculates the probability that the pair of knowledge elements is combined, or put differently, the risk in achieving or failing in recombination. This risk indicator is validated by a questionnaire survey that we carried out, in which scientists self-assessed the anticipated risk of their own projects. We further examine the relationship between risk and novelty.

The contribution of this study is two-fold. First, this study is the first to offer a validated indicator of risk in science, which contributes to the underdeveloped literature on risk in science (Franzoni and Stephan, 2021; Machado, 2021; Reinhilde et al., 2022). Second, the proposed approach offers a practical tool for scientists, policymakers, and other parties in assessing the feasibility of research plans and in developing research strategies.

This paper is structured as follows. Section 2 reviews previous studies.Section 3 describes our approach to quantify risk in recombination in science. Section 4 validates the risk indicator with the questionnaire survey. Section 5 examines the relationship between risk and novelty indicators. Section 6 summarizes the results and discusses implications.

2. LITERATURE REVIEW

2.1. Risk in Science

In general, risk is attributed to imperfect information (Marinacci, 2015). Scientific research is risky in this regard because scientists often start researching without having a clear expectation (Bourdieu, 1975; Shibayama, 2019; Whitley, 1984) and cannot perfectly know whether or what they will discover. Once a discovery is made, the consequence of the discovery might also be unpredictable (Franzoni and Stephan, 2021).

2.2. Risk in Recombination

Among various types of risk in science, this study focuses on risk entailed in a particular mode of discovery process – recombination, which is an important route to generate new knowledge (Fontana et al., 2020; Uzzi et al., 2013). Previous literature tends to consider attempts for more novel recombinations to be riskier. Indeed, Franzoni et al. (2018) conducted a survey and found that the respondents' assessment of risk is correlated with a recombinant novelty indicator. Wang et al. (2017) showed that publications with higher novel recombination scores have higher variance in their citation impact.

2.3. Measuring Risk in Recombination

Recombinant novelty indicator. The majority of novelty indicators drew on the rarity of or the distance between a combination of knowledge elements. Yet another approach draws on text information. Shibayama et al. (2021) assign a high dimensional vector to all relevant words based on the previous co-occurrences of those words, position all documents in the high-dimensional space based on the document text, and finally calculate the distance between a cited reference pair.

In these operationalizations, novelty may be associated with some challengingness. In fact, the previous studies proposing to use the novelty indicator as a proxy of risk are based on mixed reasonings in terms of what they mean by risk (Machado, 2021; Reinhilde et al., 2022). Thus, the two concepts are confounded (Franzoni et al., 2018; Franzoni and Stephan, 2021). Therefore, this study aims to measure risk in the process of recombination more directly.

Prediction of knowledge evolution. We need a technique to predict a type of knowledge evolution. While some studies aim to understand the generic mechanisms and laws behind knowledge evolution (Mazzolini et al., 2018; Tria et al., 2018), others focus on micro mechanisms with which knowledge elements are combined (Butun and Kaya, 2020; Sebastian et al., 2015). The latter is often reduced to link prediction problems in a network of scientific documents.

These link prediction algorithms draw on either node attributes or network topology. Earlier studies tend to rely on node attributes, while more recent studies draw on topological network features, contending that topological information allows more precise prediction (Butun and Kaya, 2020).

3. PREDICTION OF RECOMBINATION RISK

We similarly use link prediction algorithms to assess whether a pair of knowledge elements will be linked or not. Specifically, we draw on the word embedding technique to capture the semantic information of knowledge elements (Mikolov et al., 2013). We assign a vector expression to each knowledge element and predict the likelihood of a pair of elements being linked based on the corresponding vector pair.

3.1. Predicting Future Co-citation

To build a link prediction model, we draw on a published paper as a knowledge element and consider that two knowledge elements are combined if a pair of papers are cited together by at least one paper. With classification algorithms we compute the probability of two papers to be co-cited. By subtracting this probability from 1, we obtain the risk of failing to combine the knowledge elements.

***Data*.** We sampled papers in the field of biomedicine from the Web of Science (WoS). We prepared a sample of 120,000 paper pairs, a half linked and the other half non-linked, as the training data. We repeated the same sampling process to prepare the test data of the same size.

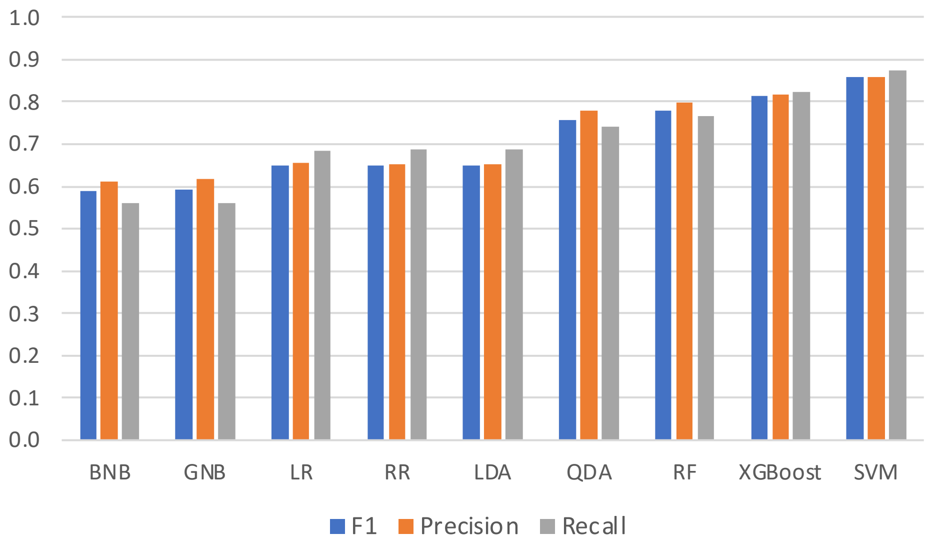
***Word embedding*.** We drew on the word embedding model that is trained with publication data up to 2010 in WoS. The model provides 300-dimensional vector representations for 1.7 million unique words. For each paper listed in the training and test data, we extracted its title and the abstract and assigned a word vector to each word included. Finally, we averaged all word vectors to generate a document vector for each paper.

***Classifiers*.** The goal of the link prediction model is to classify linked and non-linked pairs. For this classification task, we drew on several classifiers that have been commonly used for text data. We ran 9 classifiers on the training data with Python's Scikit-learn package (Pedregosa et al., 2012) and developed 9 models. We fine-tuned the hyperparameters of these models using Grid search with 5-fold cross-validation.

3.2. Performance of Prediction

To assess the performance of each trained model, we applied the models to the test data. Fig.1 presents the precision, recall, and F1 scores. The figure shows that SVM has both the highest precision (0.857) and the highest recall (0.875). Overall, SVM demonstrates the most desirable performance among the tested classifiers. Then we applied the trained model based on SVM to the test data and computed the probability of each paper pair to be linked. Results suggest that our link prediction model based on word embeddings can compute the risk in recombining a pair of knowledge elements.

Fig.1 Performance of Link Prediction Classifiers



Note. BNB: Bernoulli Naïve Bayes, GNB: Gaussian Naïve Bayes, LR: logistic regressions, RR: ridge regressions, LDA: linear discriminant analysis, QDA: quadratic discriminant analysis, RF: random forest, XGBoost: eXtreme Gradient Boosting, and SVM: support vector machine.

4. VALIDATING RECOMBINATION RISK INDICATOR WITH SURVEY

To further validate our recombination risk indicator, we carried out a questionnaire survey and asked the respondents to self-assess the risk of their past project.

4.1. Risk indicator

To bibliometrically compute the risk of a project, which is operationalized as a paper, we draw on the references cited by the focal paper as knowledge elements. As a paper usually has multiple references, we form all possible combinations from these references (for example, 10 references make 45 pairs). For each reference pair, we compute the risk score based on the SVM model developed in Section 3.

Suppose a focal paper has *N* references. Let $r_{ij} \in [0,1]$ be the risk score in combining reference i and reference j $(i, j \in \{1, \dots, N\}, i \neq j)$. The focal paper is characterized by a series of risk scores for the recombination of *N* elements. We prepared a series of risk indicators by taking various percentile values:

$$Risk_p = p\ percentile\ value\ of\ r_{ij} \tag{1}$$

where $Risk_0$ is the minimum and $Risk_{100}$ is the maximum.

4.2. Questionnaire Survey

Sample. We randomly sampled 4,625 authors. After three rounds of requests, 397 were bounced back and 378 responses were collected (response rate = 8.9%).

Questionnaire. We developed a questionnaire survey on various qualities of scientific papers based on interviews of scientists and tested it with a small-scale pilot survey. Of the survey questions, this study draws on two items concerning risk, which assess how the respondents perceived the risk of their project in two aspects.

4.3. Validation

To examine the correlation between the bibliometric indicators (*p* = 0, 10, …, 100) and the survey scores, we first regressed the bibliometric indicators on the survey scores.

Table 1 shows the result of the regression analyses. Comparing the two survey scores, the result suggests that our bibliometric risk indicators are correlated with overall risk (Table 1A) but not with technical risk (Table 1B). Regarding the overall risk, the result also shows that overall risk = 2 is significantly positively correlated with the bibliometric indicators but overall risk = 1 is not, which is as expected.

Table 1 Regression Analysis

(A) Overall Risk

	$Risk_{0}$	$Risk_{10}$	$Risk_{20}$	$Risk_{30}$	$Risk_{40}$	$Risk_{50}$	$Risk_{60}$	$Risk_{70}$	$Risk_{80}$	$Risk_{90}$	$Risk_{100}$
Overall risk = 0 (base)											
Overall risk = 1	-.039	-.107	-.104	-.080	-.085	-.080	-.044	.046	-.053	-.064	-.109
	(.232)	(.135)	(.120)	(.115)	(.108)	(.105)	(.103)	(.100)	(.099)	(.105)	(.166)
Overall risk = 2	1.151*	.689*	.608*	.541*	.459*	.422*	.374†	.307	.277	.227	.595*
	(.516)	(.314)	(.279)	(.246)	(.226)	(.210)	(.198)	(.194)	(.180)	(.177)	(.272)
Chi squared	27.065***	15.704**	15.303**	13.902**	11.535*	10.468*	8.431†	5.384	3.784	4.067	22.948***
Log likelihood	-9.057	-32.567	-49.449	-66.953	-85.503	-105.082	-127.024	-149.233	-168.306	-175.925	-110.451
N	353	353	353	353	353	353	353	353	353	353	353

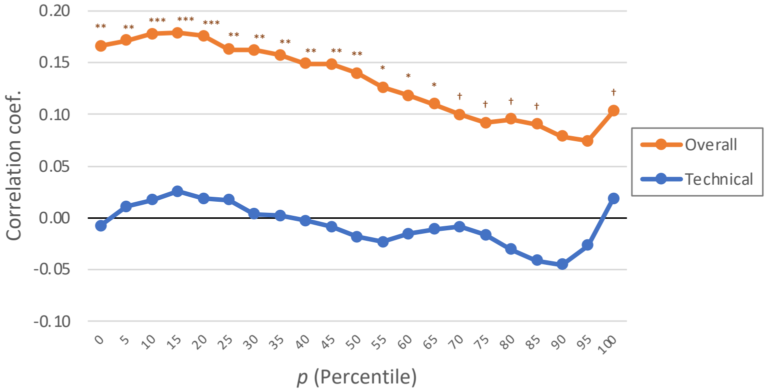
(B) Technical Risk

	$Risk_{0}$	$Risk_{10}$	$Risk_{20}$	$Risk_{30}$	$Risk_{40}$	$Risk_{50}$	$Risk_{60}$	$Risk_{70}$	$Risk_{80}$	$Risk_{90}$	$Risk_{100}$
Technical risk = 0 (base)											
Technical risk = 1	.000	.034	.030	.040	.045	.058	.090	.131	.133	.205†	.216
	(.336)	(.170)	(.144)	(.130)	(.119)	(.114)	(.110)	(.107)	(.106)	(.112)	(.174)
Technical risk = 2	-.057	.082	.0176	.033	.019	-.013	.014	.053	.003	.000	.182
	(.358)	(.213)	(.190)	(.174)	(.162)	(.154)	(.148)	(.142)	(.137)	(.136)	(.216)
Chi squared	3.937	5.352	5.737	5.833	4.864	4.661	4.413	3.419	2.191	5.245	18.682***
Log likelihood	-9.194	-32.767	-49.672	-67.143	-85.606	-105.021	-126.730	-148.766	-167.824	-175.229	-110.536
N	352	352	352	352	352	352	352	352	352	352	352

Note. Generalized linear model with a logit link and the binomial family. Unstandardized coefficients (robust errors in parentheses). Two-tailed test. †p<0.1. *p<0.05. **p<0.01.***p<0.001. The sub-fields within biomedicine are controlled for.

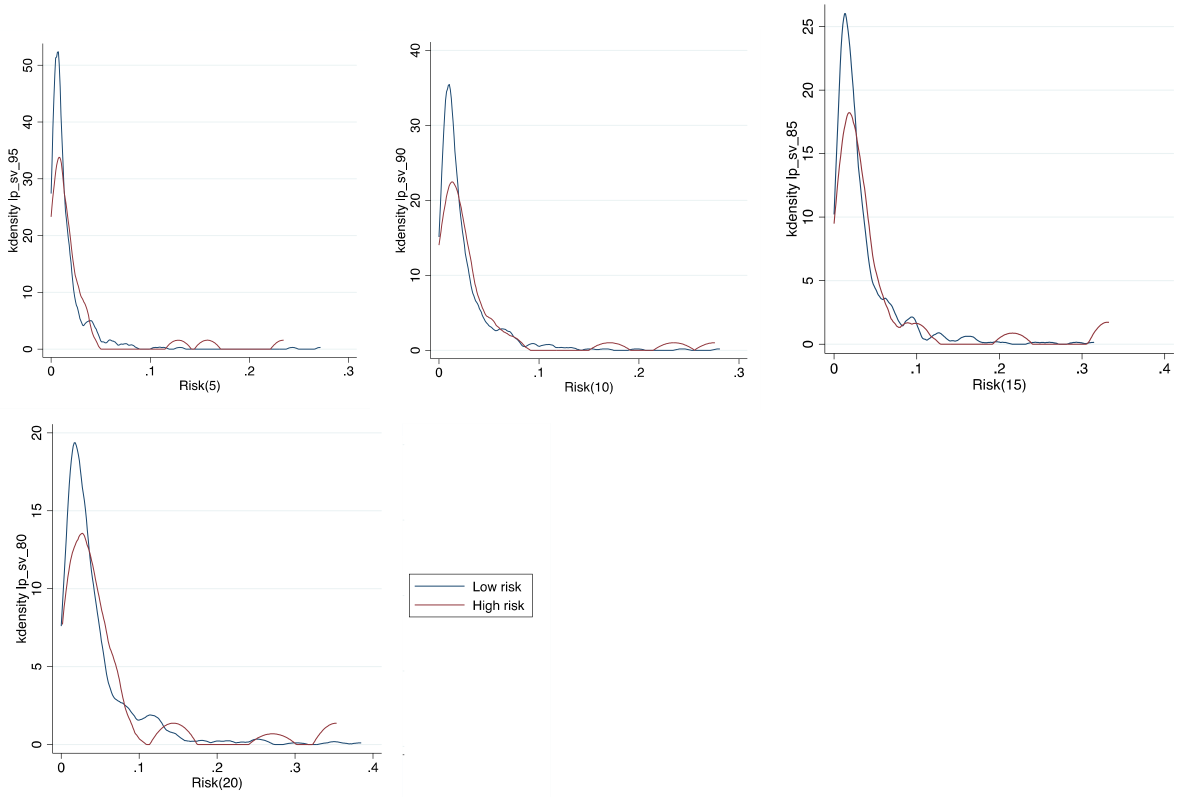
We also illustrated the Pearson's correlation coefficients between the survey scores and the bibliometric indicators, finding significant correlations for overall risk but not for technical risk (Fig.3). Concerning overall risk, the figure shows significantly positive correlations across a broad range of *p* values, but stronger correlations are observed particularly at lower *p* values. This implies that the risk of a project is determined by all risks that a project face. Fig.4 further illustrates the distribution of the risk indicators with relatively strong correlations with overall risk (*p* = {5, 10, 15, 20}). Comparing the high and low overall risk groups, it demonstrates that the high-risk group has greater risk scores.

Fig.3 Correlation between Bibliometric and Survey Risk Scores



Note. N = 353. Pearson's correlation coefficients. †p<0.1. *p<0.05. **p<0.01.***p<0.001. We dichotomized overall/technical risk by assigning 1 if overall/technical = 2 and 0 otherwise.

Fig.4 Distribution of Risk Indicators by Overall Risk



Note. High risk: overall risk = 2. Low risk: overall risk = 0 or 1.

5. RISK AND NOVELTY

5.1. Divergent Validity

Having constructed the bibliometric risk indicator, we test how it is related to novelty.

Bibliometric indicators. We use the bibliometric risk indicators based on the SVM model. As to the novelty indicator, we draw on the recombinant novelty indicator proposed by Shibayama et al. (2021) as it employs an operationalization that is consistent with that for our recombination risk indicator.

Correlation analysis. The result of the correlation analyses is summarized in Table 2. First, We find that recombinant novelty indicators do not appear to capture the risk perceived by scientists, unlike some previous studies assumed (Machado, 2021; Reinhilde et al., 2022). Second, we do not find compelling evidence showing that the novelty indicator captures the particular risk concept studied in this paper.

Table 2 Divergent Validity

		Bibliometric
		Novel
Survey	Overall risk	.070
	Technical risk	-.078
Bibliometric	$Risk_{5}$	-.098†
	$Risk_{10}$	-.058
	$Risk_{15}$	-.017
	$Risk_{20}$	.008

Note. N = 353. Pearson's correlation coefficients. †p<0.1.

5.2. Prediction of Impact

We investigate whether and how our indicator of recombination risk, together with novelty, is associated with future citation impact.

Setup of analysis. For this analysis, we use "top-1% cited" (*TC*) in the respective field as the dependent variable, coded 1 if the citation count of the paper is within top 1% and 0 otherwise, and regress it on the novelty indicator (*Novel*) and a risk indicator ($Risk_{15}$).

We randomly sampled 4,000 articles published in biomedicine in 2010 and evaluated their citation impact as of 2018.

Regression analysis. Table 3 reports the result of logistic regressions. Models 1 and 2 tests the relationship between novelty and future citation impact, finding a positive coefficient for the linear term and a negative coefficient for the quadratic term. Models 3 and 4 then examines the relationship between risk and future citation impact, finding a negative coefficient for the linear term and a positive coefficient for the quadratic term.

Table 3 Prediction of Impact

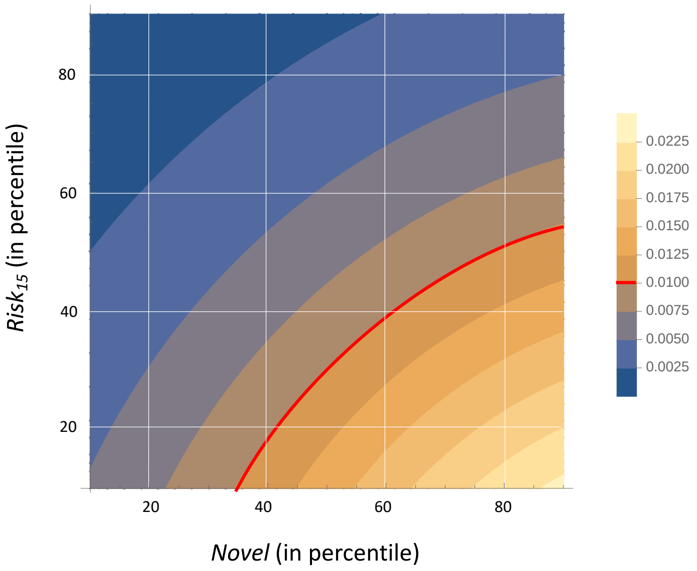
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
$Novel$	2.217***	8.183***			6.366***	6.551***	7.896***
	(.156)	(1.146)			(1.173)	(1.186)	(1.708)
$Novel^2$		-3.276***			-2.285***	-2.311***	-2.994**
		(.638)			(.656)	(.658)	(.984)
$Risk_{15}$			-14.711***	-18.375***	-18.751***	-14.101***	7.810
			(1.396)	(1.374)	(1.527)	(3.268)	(13.020)
$Risk_{15}^2$				17.581***	18.146***	17.698***	-3.513
				(1.722)	(1.866)	(1.854)	(19.653)
$Novel$						-4.977	-50.957
$\times Risk_{15}$						(3.375)	(34.117)
$Novel^2$							22.869
$\times Risk_{15}$							(21.524)
$Novel$							30.728
$\times Risk_{15}^2$							(57.463)
$Novel^2$							-7.093
$\times Risk_{15}^2$							(39.075)
Chi-squared stat	200.967***	216.578***	116.286***	192.494***	307.846***	330.828***	375.162***
Log likelihood	-109.684	-109.415	-108.769	-108.551	-106.422	-106.409	-106.382
N	3903	3903	3903	3903	3903	3903	3903

Note. Logistic regressions. Unstandardized coefficients (robust errors in parentheses). Two-tailed test. †p<0.1. *p<0.05. **p<0.01.***p<0.001. The sampling weight is incorporated in the regression analysis. The sub-fields within biomedicine are controlled for.

As risk and novelty have been confounded in the literature, Model 5 includes both the novelty and risk indicators. The magnitude of the coefficients slightly changes, but the overall relationships remain qualitatively similar. Models 6 and 7 introduce various interaction terms without finding a significant effect. Thus, novelty and risk indicators are associated with future citations through different mechanisms, which also supports our argument that these two indicators capture different concepts.

To visually illustrate the result, Fig.4 presents the contour map of the predicted citation impact with a range of novelty and risk values. it shows that higher citation impact than the average occurs only with high novelty and low risk. In particular, the result suggests that risky recombination, even if successful, cause disadvantages in attracting future citations, although high-risk research has been encouraged (OECD, Machado, 2021; 2021).

Fig.4 Prediction of Impact



Note. The contour map of *prob*.(*TC*=1) based on Model 5 in Table 3. The red curve indicates the base line (*prob*.(*TC* =1) = 0.01), below which *prob*.(*TC* =1) > 0.01. The novelty and risk indicators are scaled in their percentile values (e.g., 50 is the median of the indicators).

6. DISCUSSIONS AND CONCLUSIONS

Overall, this study makes scholarly contribution to the underdeveloped literature on risk in science (Franzoni and Stephan, 2021; Machado, 2021; Reinhilde et al., 2022) by providing the first validated indicator of a particular type of risk.

We expect that the proposed method is applicable not only to scientific papers but also to other types of scientific texts. These applications should assist decision-makers to assess the feasibility of a research project and help identify potential risks involved in a project.

Despite all the contributions, further refinement and development of risk indicators are warranted. First, future research should develop a method to quantify risk in broader modes of scientific progress. Second, we tested our approach only in the biomedical field because of the limitation of the word embedding model. Third, there is room for improvement in extracting semantic information from documents. Fourth, not all the pairs may represent intended recombinations, which can cause errors. Finally, it is of interest to investigate the source of risk. Risk is attributed to more than novelty, but what it is remains unclear.

References

Althaus, C.E., 2005. A Disciplinary Perspective on the Epistemological Status of Risk. Risk Analysis 25, 567-588.

Aven, T., 2011. On Some Recent Definitions and Analysis Frameworks for Risk, Vulnerability, and Resilience. Risk Analysis 31, 515-522.

Boudreau, K.J., Guinan, E.C., Lakhani, K.R., Riedl, C., 2016. Looking across and Looking Beyond the Knowledge Frontier: Intellectual Distance, Novelty, and Resource Allocation in Science. Management Science 62, 2765-2783.

Bourdieu, P., 1975. The Specificity of the Scientific Field and the Social Conditions for the Progress of Reason. Social Science Information 14, 19–47.

Breiman, L., 2001. Random Forests. Machine Learning 45, 5-32.

Butun, E., Kaya, M., 2020. Predicting Citation Count of Scientists as a Link Prediction Problem. Ieee Transactions on Cybernetics 50, 4518-4529.

Chen, T., Guestrin, C., 2016. Xgboost: A Scalable Tree Boosting System, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Association for Computing Machinery, San Francisco, California, USA, pp. 785–794.

Cortes, C., Vapnik, V., 1995. Support-Vector Networks. Machine Learning 20, 273-297.

Dahlin, K.B., Behrens, D.M., 2005. When Is an Invention Really Radical? Defining and Measuring Technological Radicalness. Research Policy 34, 717-737.

Daud, A., Ahmed, W., Amjad, T., Nasir, J.A., Aljohani, N.R., Abbasi, R.A., Ahmad, I., 2017. Who Will Cite You Back? Reciprocal Link Prediction in Citation Networks. Library Hi Tech 35, 509-520.

Fleming, L., 2001. Recombinant Uncertainty in Technological Search. Management Science 47, 117-132.

Fontana, M., Iori, M., Montobbio, F., Sinatra, R., 2020. New and Atypical Combinations: An Assessment of Novelty and Interdisciplinarity. Research Policy 49, 28.

Foster, J.G., Rzhetsky, A., Evans, J.A., 2015. Tradition and Innovation in Scientists’ Research Strategies. American Sociological Review 80, 875-908.

Franzoni, C., Scellato, G., Stephan, P., 2012. Foreign-Born Scientists: Mobility Patterns for 16 Countries. Nature Biotechnology 30, 1250-1253.

Franzoni, C., Scellato, G., Stephan, P., 2018. Context Factors and the Performance of Mobile Individuals in Research Teams. Journal of Management Studies 55, 27-59.

Franzoni, C., Stephan, P., 2021. Uncertainty and Risk-Taking in Science: Meaning, Measurement and Management. National Bureau of Economic Research Working Paper Series No. 28562.

Gewin, V., 2012. Risky Research: The Sky's the Limit. Nature 487, 395-397.

Hansson, S.O., 2018. Risk, in: Zalta, E.N. (Ed.), The Stanford Encyclopedia of Philosophy. Metaphysics Research Lab, Stanford University.

Hardin, J.W., Hilbe, J.M., 2018. Generalized Linear Models and Extensions, 4th ed. Stata Press, TX, USA.

Kaplan, S., Garrick, B.J., 1981. On the Quantitative Definition of Risk. Risk Analysis 1, 11-27.

Kenter, T., Borisov, A., & de Rijke, M. (2016). Siamese CBOW: Optimizing Word Embeddings for Sentence Representations. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 941–951.

Lin, Y., Evans, J.A., Wu, L., 2022. New Directions in Science Emerge from Disconnection and Discord. Journal of Informetrics 16, 101234.

Linton, J.D., 2016. Improving the Peer Review Process: Capturing More Information and Enabling High-Risk/High-Return Research. Research Policy 45, 1936-1938.

Liu, W.Y., Nanetti, A., Cheong, S.A., 2017. Knowledge Evolution in Physics Research: An Analysis of Bibliographic Coupling Networks. Plos One 12, 19.

Machado, D., 2021. Quantitative Indicators for High-Risk/High-Reward Research, OECD Science, Technology and Industry Working Papers. OECD Publishing, Paris.

Marinacci, M., 2015. Model Uncertainty. Journal of the European Economic Association 13, 1022-1100.

Matsumoto, K., Shibayama, S., Kang, B., Igami, M., 2020. A Validation Study of Knowledge Combinatorial Novelty, NISTEP Discussion Paper. NISTEP, Tokyo.

Mazzolini, A., Colliva, A., Caselle, M., Osella, M., 2018. Heaps' Law, Statistics of Shared Components, and Temporal Patterns from a Sample-Space-Reducing Process. Physical Review E 98.

Mednick, S.A., 1962. The Associative Basis of the Creative Process. Psychological Review 69, 220-232.

Merton, R.K., 1973. Sociology of Science. University of Chicago Press, Chicago.

Mikolov, T., Chen, K., Corrado, G., Dean, J., 2013. Efficient Estimation of Word Representations in Vector Space. arXiv.

Min, C., Bu, Y., Wu, D., Ding, Y., Zhang, Y., 2021. Identifying Citation Patterns of Scientific Breakthroughs: A Perspective of Dynamic Citation Process. Information Processing & Management 58, 102428.

OECD, 2021. Effective Policies to Foster High-Risk/High-Reward Research, OECD Science, Technology and Industry Policy Papers. OECD Publishing, Paris.

Palchykov, V., Krasnytska, M., Mryglod, O., Holovatch, Y., 2021. A Mechanism for Evolution of the Physical Concepts Network. Condensed Matter Physics 24.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Müller, A., Nothman, J., Louppe, G., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, É., 2012. Scikit-Learn: Machine Learning in Python. arXiv.

Reinhilde, V., Jian, W., Paula, S., 2022. Do Funding Agencies Select and Enable Risky Research: Evidence from Erc Using Novelty as a Proxy of Risk Taking. National Bureau of Economic Research, Inc.

Sebastian, Y., Siew, E.G., Orimaye, S.O., 2015. Predicting Future Links between Disjoint Research Areas Using Heterogeneous Bibliographic Information Network, Advances in Knowledge Discovery and Data Mining, Part Ii. Springer-Verlag Berlin, Berlin, pp. 610-621.

Shibata, N., Kajikawa, Y., Sakata, I., 2012. Link Prediction in Citation Networks. Journal of the American Society For Information Science and Technology 63, 78-85.

Shibayama, S., 2019. Sustainable Development of Science and Scientists: Academic Training in Life Science Labs. Research Policy 48, 676-692.

Shibayama, S., Yin, D., Matsumoto, K., 2021. Measuring Novelty in Science with Word Embedding. Plos One 16, e0254034.

Simonton, D.K., 2003. Scientific Creativity as Constrained Stochastic Behavior the Integration of Product, Person, and Process Perspectives. Psychological Bulletin 129, 475-494.

Sun, Y., Latora, V., 2020. The Evolution of Knowledge within and across Fields in Modern Physics. Scientific Reports 10.

Trapido, D., 2015. How Novelty in Knowledge Earns Recognition: The Role of Consistent Identities. Research Policy 44, 1488-1500.

Tria, F., Loreto, V., Servedio, V.D.P., 2018. Zipf's, Heaps' and Taylor's Laws Are Determined by the Expansion into the Adjacent Possible. Entropy 20.

Tu, Y.-N., Seng, J.-L., 2012. Indices of Novelty for Emerging Topic Detection. Information Processing & Management 48, 303-325.

Uzzi, B., Mukherjee, S., Stringer, M., Jones, B., 2013. Atypical Combinations and Scientific Impact. Science 342, 468-472.

Wang, J., Veugelers, R., Stephan, P., 2017. Bias against Novelty in Science: A Cautionary Tale for Users of Bibliometric Indicators. Research Policy 46, 1416-1436.

Wang, Y., Jones, B.F., Wang, D., 2019. Early-Career Setback and Future Career Impact. Nature Communications 10, 4331.

Whitley, R., 1984. The Intellectual and Social Organization of the Sciences. Oxford University Press, New York.

Yang, J., Lu, W., Hu, J., Huang, S., 2022. A Novel Emerging Topic Detection Method: A Knowledge Ecology Perspective. Information Processing & Management 59, 102843.

Yaqub, O., 2018. Serendipity: Towards a Taxonomy and a Theory. Research Policy 47, 169-179.

Yin, D., Wu, Z., Yokota, K., Matsumoto, K., Shibayama, S., 2022. Identify Novel Elements of Knowledge with Word Embedding.