No research, no risk? An empirical study of determinants of information leakage attacks

Oleh Stupak[*]

June 10, 2022

Abstract: This study provides new empirical evidence that research-intensive industries are particularly susceptible to industrial cyberespionage. Based on Eurostat aggregated data on enterprises' innovative and digital activity, I construct a tailored dataset that allows me to achieve robust statistical inference and study the relationship between information leakage attack rate, research intensity, and companies' data reliance. The study uses multivariate fractional regression analysis to distinguish two industry-specific associations: high-tech manufacturing industries are prone to experience sophisticated targeted attacks, while knowledge-intensive service industries are affected by opportunistic financially driven attacks.

Keywords: *Fixed Effects, Fractional Responses, Panel Data, Cyber Crime, Cyber Security, Hacking.*

JEL Classification: *C23, C25, K24.*

1 Introduction

This paper investigates the association between the research and development (R&D) intensity in an industry and the rate of information leakage cyberattacks. The relationship is studied under several statistical settings. The article's primary goal is to answer a simple question: are research-intensive industries more likely to become victims of information leakage attacks?

Cybersecurity has become one of the major global threats. Thousands of businesses and millions of people are becoming victims of cybercrime every year. The cost estimate of cyber-

[*] University of Oxford, Kellogg college, Oxford OX2 6PN, oleh.stupak@kellogg.ox.ac.uk

crime to the global economy surpassed \$1 trillion in 2018 and almost doubled to \$2 trillion in 2020, making it roughly 2% of global GDP. With a predicted annual growth rate of 15%, this number is expected to reach \$4-5 trillion by 2025 (Juniper Research, 2019). Recent reports suggest that more than two-thirds of global cybercrime costs were associated with data breaches: intellectual property theft, cyberespionage and financially motivated crimes (McAfee, 2020).

Around 32% of companies experienced data breaches in the United Kingdom in 2019 with total economic losses estimated around £27 billion, £16.7 billion of which was associated with industrial espionage and intellectual property theft (IBM, 2020). The average cost of the data breach incident in the UK reached £3.5 million (DCMS, 2020).

The data breaches are generally divided into two big groups: targeted and opportunistic (or untargeted) (National Cyber Security Centre, 2015). The former is closely associated with intellectual property theft and industrial espionage, while the latter is often associated with more financially driven attacks that take advantage of unidentified victims' vulnerabilities to obtain easily monetised data. However, it should not be treated as a rule: some targeted attacks could be aimed at consumer records, while some opportunistic attacks might extract trade secrets from the victims.

The study focuses on targeted attacks and theorises that they might be adopted as a means of unfair competition. An underdog in an R&D race can choose to obtain competitors' know-how or corporate secrets utilising a cyberattack. Similarly, clients databases can be targeted by cyberattacks in sales driven competition. According to Oxford Economics (2014), every fifth company in the UK experienced a data breach, with 61% and 59% of companies reporting the loss of competitive advantage due to the loss of intellectual property (IP) and commercially sensitive data, respectively. Stealing know-how or client records could be a cheap, efficient and relatively safe method of gaining a competitive edge in an R&D race (Crete-Nishihata et al., 2018).

The study employs a tailored dataset based on aggregated survey data from Eurostat to investigate this claim. The data contain cyberattack rates and plausible proxies for research intensity and customer data use.

The research estimates fixed effects fractional outcome regression models with the country and industry controls. The estimation technique is superior to alternative methods (e.g. ordinary least squares (OLS) or Tobit models), as the chosen dependent variable is the information leakage attack rate in a given country in a given industry $\in [0, 1]$. However, the coefficients estimated by the fractional outcome model are not directly interpretable. The study utilises marginal analysis to estimate average marginal effects, which can be interpreted similarly to classic OLS coefficients.

The dependent variable is estimated separately against three research intensity proxy variables that capture distinct groups of R&D personnel : (1) a share of all R&D personnel in overall employment in a given country in a given industry, (2) a share of researchers in overall employment in a given country in a given industry, and (3) a share of research support personnel in a given country in a given industry. The choice of the research intensity proxies follows conventional economic growth theory, where research personnel is often employed as an essential part of the ideas output function (e.g. Porter and Stern, 2000).

Due to the secrecy of information leakage attacks, the available data possess two major challenges. Firstly, the dataset has a considerable share of partial observations, leading to biased estimates if not treated carefully. Therefore, fractional outcome regression analysis is backed up by the multiple imputation (MI) technique, a hybrid Bayesian-frequentists approach that allows for unbiased estimates in the presence of partial observations. Secondly, the dependent variable aggregates both targeted and untargeted attacks.

Data Breach Investigations Report (DBIR) 2011 suggests that targeted attacks are generally aimed at intellectual property and trade secrets, which are the results of research activity (Verizon Business, 2011). On the contrary, untargeted (opportunistic) attacks are responsible for 90% of compromised records, customer financial data, and credentials. Thus, the relationship between the rate of information leakage attacks and research intensity cannot be studied without controlling for opportunistic attacks. To tackle the issue, I use an additional Client Resource Management (CRM) proxy variable representing the end-user data reliance in a given country in a given industry to absorb the variance associated with opportunistic attacks. The additional variable provides an opportunity to consider another (secondary) research question: are data reliant industries susceptible to information leakage attacks?

The theory is tested with univariate and multivariate fractional outcome models estimated on Eurostat data. Univariate models demonstrate a strong association between information leakage attack rate and research intensity regardless of a chosen proxy. The share of research support personnel demonstrates a more than fourfold larger marginal effect than the other two proxies. Moreover, it is the only proxy that maintains a statistically significant coefficient after controlling for opportunistic attacks. Therefore, the share of R&D support personnel might be a better proxy to capture the value of innovative ideas. The discussion on this matter can be found in Subsection 3.1.2.

The multivariate estimation reveals that the CRM proxy variable has a statistically significant coefficient. By employing the multivariate model specified with both research intensity and end-user data reliance proxies, the study distinguished two positive associations potentially linked to targeted and opportunistic attacks.

To back up the findings, the study additionally performs multivariate estimations for six subsamples of different technological environments: all manufacturing industries, high-tech manufacturing industries, low-tech manufacturing industries, all services, knowledge-intensive services, and less knowledge-intensive services. Consistent with the proposed causal mechanism, information leakage attacks in high-tech manufacturing industries are strongly associated with the research intensity proxy while demonstrating no association with the data reliance proxy. On the contrary, the information leakage attacks rate in knowledge-intensive service industries does not demonstrate any association with the research intensity proxy while showing a strong association with the data reliance proxy.

Therefore, the results suggest that research-oriented manufacturing industries are prone to fall victim to targeted information leakage attacks (potentially industrial espionage). Meanwhile, knowledge-intensive service industries are more likely to experience opportunistic financially driven attacks. However, this can be considered solely as supporting evidence favouring the proposed causal mechanism. Establishing an actual causality will require a better and more complete dataset, which is currently unavailable.

The rest of the paper is built as follows. The next Section shortly discusses related literature. Section 2 summarises main hypotheses. Detailed information on the chosen methodology and data can be found in Section 3. Section 4 reports the results and Section 5 offers conclusions and discussion.

1.1 Related literature

There have been limited studies on the economics of information leakage attacks. Arguably, the most famous series of empirical papers were written by Moore and Clayton (2007, 2008) and Moore et al. (2009). The authors studied a continuous arms race between banks and cybercriminals. Criminals created websites that impersonate banks and steal sensitive customer data, while banks' security services (or specific take-down companies hired by banks) were trying to hunt them down and minimise losses. The authors derived valuable insights and provided recommendations for defenders against phishing attacks. For example, they recognised that even though a phishing website might be known to some defending parties, it might be unknown to the party with the actual take-down contract. Therefore, the authors concluded with a recommendation for defenders to cooperate and share information. The authors also found that some banks were targeted way more frequently than others, although they did not explain this phenomenon in the original series.

In the following paper, Moore and Clayton (2011) attempted to research how cybercriminals choose their targets. To launch a phishing website, cybercriminals must compromise websites to host a phishing page, malware or other harmful content. They concluded that foes often used quite simple techniques. For instance, they utilised search engines to identify possible victims – websites with common (and usually obvious) security holes. The authors claimed that the simplicity of compromising the website is one of the main drivers of victim selection. This finding, however, is only applicable to opportunistic attacks.

Ransbotham (2010) compared the durability of open and closed source software. The industry has an ongoing debate on which approach leads to less vulnerable software. The open-source code is being developed and tested by many programmers and testers. In theory, exposure to a broader number of developers leads to discovering a more significant number of bugs, glitches, and vulnerabilities compared to a closed source with restricted access to the code. However, by analysing two years of data on intrusion detection, Ransbotham found that open source software was targeted more frequently, and exploits were being developed faster. He suggested that it happened due to the relative ease of accessing the source code. Therefore, the type of software an enterprise uses could influence its exposure to information leakage attacks.

A significant body of research is devoted to the quantification of cyber risks in a race for better cybersecurity insurance models. Such quantification is a fundamental risk management challenge that has been addressed by scholars, insurers, and businesses (see the comprehensive survey in Bardopoulos, 2020). While this body of literature is not per se directly related to the presented study, it offers some insights into the methodologies and techniques, which might be useful in analysing information leakage attacks.

An early attempt to quantify cyber risk was made by Hoo (2000). The author employed utility theory to analyse companies' decision trees based on computer security surveys. The study identified that the main factors behind the cyber risks are the company's wealth indices and probabilities of the security events associated with decision trees. It proposed to use the customer base as a main measure of exposure. However, the main factors were defined following the anecdotal evidence from a small sample survey analysis.

Similarly, Mukhopadhyay et al. (2005) also used decision trees and utility theory in an attempt to build an insurance risk premium. The main exogenous inputs were assumed to be insurance claims' frequency and severity (without further specifications). The theoretical model, however, could not be verified empirically due to unavailability of data.

Böhme (2005); Böhme and Kataria (2006) presented the first insurance models that could be supported by empirical data. Böhme (2005) argued that industrial IT infrastructure might intervene with the development of a prober cyberinsurance market. The author recognised that cyber risks of insureds are correlated due to the dominance of a few IT platforms, which, in turn, lead to correlated losses. The author has explored the conditions upon which the insurance market would be feasible. Böhme and Kataria (2006) expanded the previous work by investigating the correlation of cyber risks within and across firms based on "honeypot" data (Leita et al., 2008).

Liu et al. (2007) used a regression analysis to investigate the relation between computer virus incidents and cybersecurity measures (e.g. information security policy, human cultivation). Additionally, they have explored the correlation between the breach probability and the number of email accounts. They found statistically significant results suggesting that continuous investment in cybersecurity can significantly mitigate cyber risks related to breaches due to a virus exposure. The data, however, did not allow for causal inference.

Another empirical paper was presented by Wang and Kim (2009). The authors used time-series regression analysis to study cyber attacks on a country level. Particularly, the authors studied the spatial autocorrelation of cyberattacks and the effects of joining IT conventions. They found evidence of autocorrelated cyberattacks and concluded that physical boundaries must be considered an essential factor in designing cybersecurity regulations.

Herath and Herath (2007) presented a premium pricing framework for assessing the potential financial losses based on the observed distribution of the number of compromised computers. The empirical work was based on survey data and analysed by the means of copula. Data defects were identified as one of the main limitations of the approach (the data had only 15 observations).

Innerhofer-Oberperfler and Breu (2010) conducted a qualitative study in order to identify the main indicators for cyberinsurance. They interviewed 36 field experts and reduced their statements to the set of 94 indicators ranked according to their relative importance (subjectively chosen by experts during the post-interview questionnaire). In their study, "processing sensitive data with high confidentiality requirements" and "existence of worth-protecting know-how, patents and otherwise valuable information" are ranked third and fourth respectively – right after the indices of dependency on IT systems. The empirical model presented in this paper attempts to quantify those qualitative findings.

Overall, the main body of the literature discusses the measurement of information leakage attacks effectiveness, technological drivers of cyberattacks and espionage, and offers various ways of cyber risk quantification. However, it says little about the determinants of exposure to information leakage attacks. This study attempts to contribute to this question by researching at least one of the possible drivers of the phenomenon – the research intensity.

2 Hypotheses

I construct two hypotheses to test for the possible associations of the rate of information leakage cyberattacks[1] with (1) research intensity and (2) industries' reliance on the end-user data. The first hypothesis concerns the association of research intensity with the rate of information leakage attacks. The association is believed to be linked to targeted attacks that are predominantly aimed at IP and corporate secrets.

Hypothesis 1 *Higher research intensity in the industry (independent variable) is associated with an increase in the information leakage cyberattack rate in the industry (dependent variable).*

Simultaneously the study examines the association between information leakage attacks and end-user data reliance.

Hypothesis 2 *Higher reliance on end-user data in the industry (independent variable) is associated with an increase in the information leakage cyberattack rate in the industry (dependent variable).*

3 Empirical strategy

3.1 Data

The research uses a tailored dataset that is based on three publicly available sets from Eurostat:

1. Structural business statistics (SBS) dataset that covers the structure, conduct and performance of the industry, construction, services and trade (Eurostat, 2010c);

2. Science, technology and innovation dataset which covers statistics in the fields of science, technology and innovation (Eurostat, 2010b);

3. Digital economy and society dataset covers all aspects of the usage of Information and Communication Technologies (ICT) by individual households and businesses (Eurostat, 2010a).

[1] Definied as the rate of cyberattacks that result in the theft of information.

1 Belgium	9 Spain	17 Hungary	25 Slovakia
2 Bulgaria	10 France	18 Malta	26 Finland
3 Czechia	11 Croatia	19 Netherlands	27 Sweden
4 Denmark	12 Italy	20 Turkey	28 UK
5 Germany	13 Cyprus	21 Poland	29 Iceland
6 Estonia	14 Latvia	22 Portugal	30 Norway
7 Ireland	15 Lithuania	23 Romania	31 Macedonia
8 Greece	16 Luxembourg	24 Slovenia	32 Serbia

Table 1: List of countries included in the study

All selected datasets are survey-based, and aggregate responses from $148,000$ enterprises categorised by year, country, and economic activity according to the second revision of NACE[2] (European Commission, 2008). However, datasets have a different granularity of aggregation. For instance, the SBS dataset has detailed information on every economic activity present in the NACE classification, while the two others utilise industry aggregates that group several similar economic activities. Therefore, to make merging possible, datasets were aggregated to the highest common level, that of the Digital economy and society dataset. Countries that are included in the research are summarised in Table 1. Figure 1 illustrates the geographical distribution of the countries included in the study and the corresponding average shares of companies that experienced cyberincidents in each country. A list of industry aggregates can be found in Appendix A.1. Country and industry dummies constitute two sets of control variables.

Note that even the finalised dataset includes only variables for 2010 due to the fact that "Security Incidents and Consequences" surveys included in "Digital economy and society dataset" was conducted only for 2010 and 2019; thus, 2010 was the only year where all three datasets overlap. The partial observations constitute 36% of data, which I address by utilising a multiple imputation approach described in Subsection 3.2.

[2] NACE stands for "Nomenclature Statistique des Activités économiques dans la Communauté Européenne". It is a standard European system for classifying firms by the industrial sector.

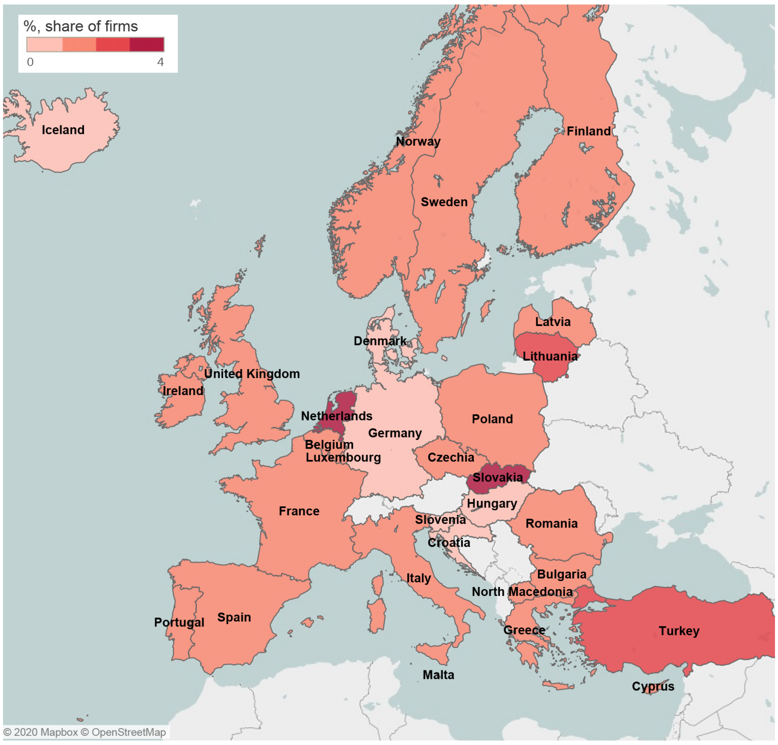

Figure 1: Geographical distribution of the countries included in the study. The countries are colour-coded according to the average share of enterprises that experienced information leakage attacks.

3.1.1 Dependent variable

The dependent variable – *the rate of information leakage attacks* (B_{ij}), represents the share of the enterprises which experienced cybersecurity incidents that resulted in the disclosure of confidential data due to intrusion, pharming or phishing attacks in a given country (i) in a given industry (j).

The variable has a right-skewed distribution, which suggests the usage of the logarithm to normalise it (See Figure 2). However, zero observations constitute 39.9% of observations, which means that log-transformation might significantly affect the results.

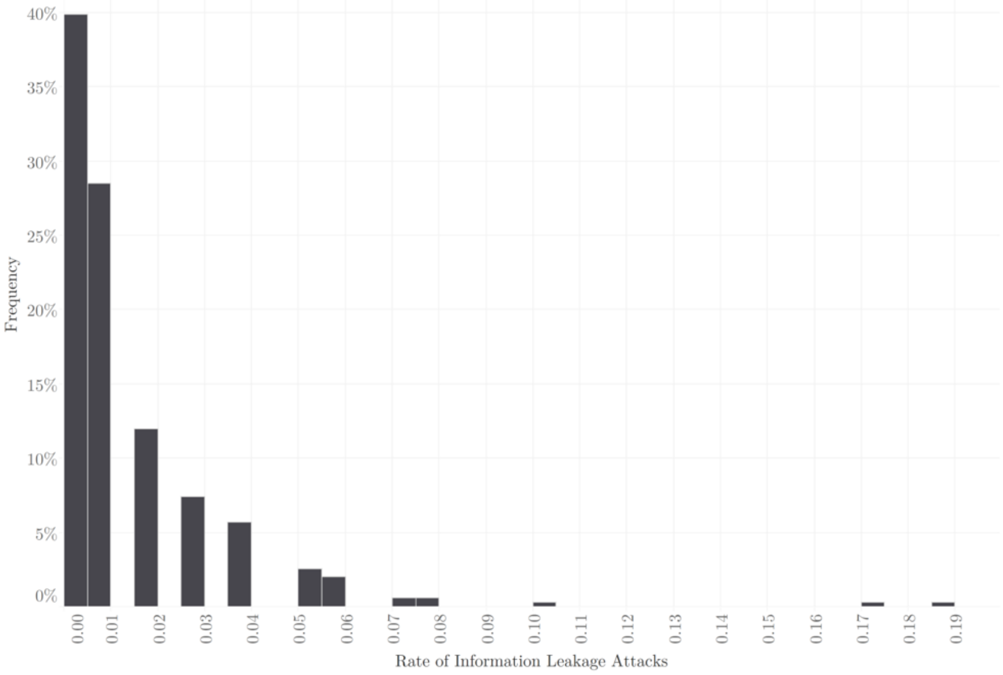

Figure 2: Dependent variable histogram

There is no reasonable way of determining the stolen type of data (e.g. personal records, financial information, blueprints). The variable aggregates all the possible cases regardless of the original motive of the breach. To shed some light on the issue, the research refers to 2011 Data Breach Investigation Report (based on The Veris Community Database (VCDB)), which analyses the majority of known data compromise incidents in 2010 (Verizon Business, 2011).

The report divided cyberincidents into two big groups: (1) targeted and (2) opportunistic attacks. Around 20% of all discovered breaches can be affiliated with targeted attacks, which, however, correspond to less than 10% of all compromised records (e.g. credentials, card details, personal data) in 2010. Those attacks are disproportionally aimed at more specific types of data that are not generally stolen in bulk: IP and corporate secrets. As later reports suggest, the primary motivation behind targeted attacks is industrial espionage - that is, the stolen data are used to achieve a competitive edge on the market (Verizon Business, 2020).

Opportunistic breaches constitute around 80% of cases. Victims of those attacks are chosen due to the exploitable vulnerabilities they exhibit (e.g. SQL Injection-vulnerable websites) rather than their business activity. Opportunistic attacks are predominantly financially motivated and are responsible for 90% of compromised records, including customers financial data, personal information, and credentials that might be quickly sold on the black market.

It should be recognised that the actual rate of intellectual property and corporate secrets theft is likely larger than reported. Financial fraud detection remains the most effective method of discovering a breach. However, IP and corporate secrets are not generally involved in financial identity theft. It is then reasonable to assume that a substantial share of targeted attacks has never been discovered.

It is worth mentioning that the research focuses on the cyberattacks performed by external actors. While there is no doubt that an insider threat is a common problem in modern competition, the external actors are the predominant source of information leakage attacks. External actors were responsible for 92% of all reported data breaches in 2010 (Verizon Business, 2011). I leave the discussion of insider threats for another occasion.

3.1.2 Independent variables

As neither the exact form of the targeting mechanism nor the actual form of the ideas output function is known, the research tests the central hypothesis using three proxy variables. Proxy variables represent shares of interlayers of organisations' R&D personnel and resemble the variables used in macroeconomic and economic growth literature to model ideas output (Porter and Stern, 2000; Jones, 2003; Coe and Helpman, 1995). The variables are derived according to the Frascati Manual[3] and are defined as follows (OECD, 2018):

1. *R&D personnel* (r_{ij}^{per}) is the share of all persons in overall employment in a given country in a given industry who engaged in R&D as well as those who provide direct support for the R&D activities (e.g. R&D managers, technicians);

2. *Researchers* (r_{ij}^{res}) is the share of professionals and researchers in overall employment in a given country in a given industry who are engaged in the conception or creation of new knowledge, conduct research and improve or develop theories, models, techniques instrumentation, software or operational methods.

3. *Technicians and other R&D supporting personnel* (r_{ij}^{sup}) represents the share of persons in overall employment in a given country in a given industry who provide direct services for the R&D activities (e.g. technicians, R&D managers, and administrators).

Note that the R&D personnel variable can be expressed as a sum of latter two variables, $r_{ij}^{per} = r_{ij}^{res} + r_{ij}^{sup}$. The industrial distribution of all three variables is demonstrated in Figure 3. The figure demonstrates that manufacturing industries (codes start with C) possess a much larger share of research personnel than service industries. As shown in the figure, the

[3] The Frascati Manual is a document setting forth the methodology for the collection of analyses of statistics about research and development.

manufacture of computer, electronic, and optical products (C26) has the largest shares of research personnel among all the industries. Services are generally less research-oriented, with only two groups demonstrating a relatively large share of R&D personnel: publishing of books, software, and media (J55-J63) and professional, scientific, and technical activities (M69-M74). See Appendix A.1 for a complete list of industries' aggregates.

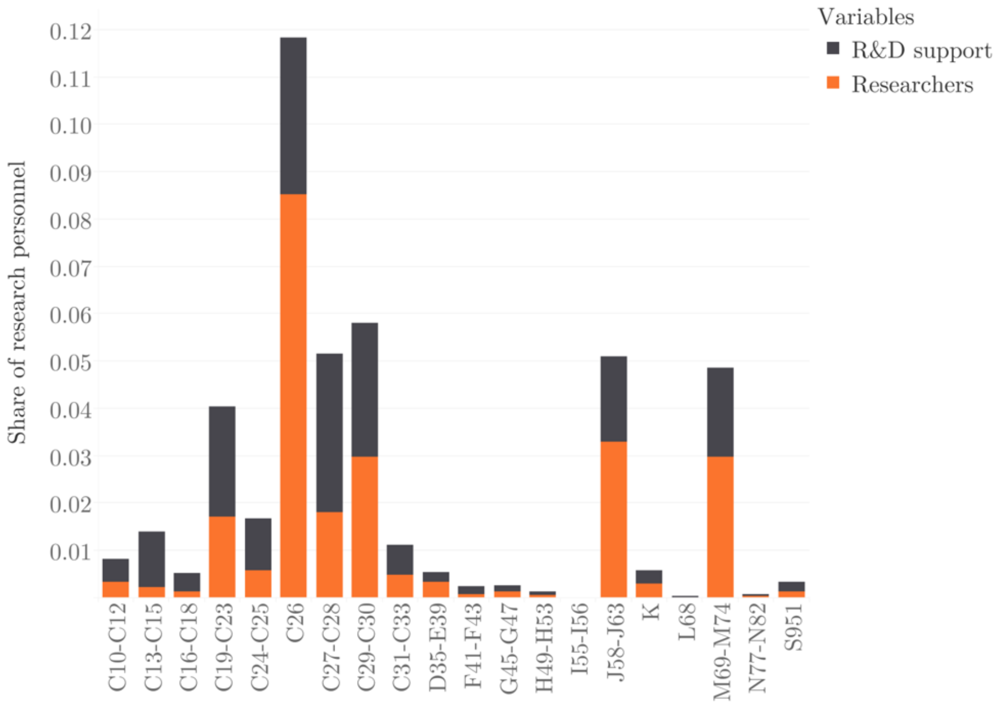

Figure 3: Industrial distribution of the research intensity proxy variables

Due to collinearity the independent variables cannot be included in the specification simultaneously. Therefore, I use three separate sets of specifications to analyse the association between each independent variable with information leakage attacks rate.

The specifications with the share of research personnel as a primary independent variable are built following the assumption that researchers and R&D support personnel cooperatively contribute to innovative ideas. It assumes that industries with the largest share of research personnel are the most research-intensive and produce the most valuable results. However, it also implies that research intensity is proxied by the sum of the researchers and research support personnel, which might not be the case (e.g. the cooperation might be synergistic, or researchers might contribute significantly more than the research support personnel).

The second set of specifications assumes that research intensity can be attributed mainly to the input of researchers and research professionals. While the assumption is plausible as the research support personnel might not contribute to the research directly, it also neglects the latter's potential effects on the output of the ideas.

Finally, the last set of tests is performed with the share of technicians, R&D managers and other support personnel as the primary independent variable. A company with highly productive researchers might hire a larger R&D support body. If researchers and research support personnel are complements in the production function of the ideas, then they would allow the researchers to focus on their work and increase their productivity.

Lewis (2017) suggested that a greater proportion of technicians and R&D managers indicates mature high-tech organisations specialising in process development. Those organisations have a more elaborate division of labour which allows them efficiently commercialise innovation and carry out abundant manufacturing in the pilot plants (Steger et al., 1975). Similarly, a significant fraction of R&D managers, administrators, and equivalent personnel in the service industries might indicate the high commercial potential of research output. In other words, industries with a large share of R&D support personnel are typically those with the most immediately marketable ideas and thus are often attractive targets for industrial espionage. Moreover, the share of research support personnel has the most significant positive correlation with the R&D expenses and overall turnover among all three proxies in our data.

On the other hand, a larger body of R&D support personnel might disproportionately increase the company's attack surface. It increases the number of people with access to sensitive data without a substantial contribution to the ideas generation. Hence, the variable might capture both research intensity and extended attack surface. However, industry control variables can rule out the attack surface effect. Additionally, I test the plausibility of the attack surface hypothesis using the extended specification (see Subsection 4.4 for details), which includes the auxiliary attack surface variable defined as the average amount of R&D support personnel in an enterprise in a given industry in a given country. Being an absolute measure, it allows for comparing the average attack surfaces among industries while neglecting the research intensity itself[4] (Terleckyj, 1960; Comanor, 1965; Gruber et al., 1966).

All research intensity proxy variables have a right-skewed distribution (See Appendix A.2). Still, as with the dependent variable, there are no solid theoretical grounds to use any transformation to normalise them.

I use an additional proxy variable to control the variance associated with opportunistic attacks and test Hypothesis 2. Note that opportunistic attacks are responsible for most compromised consumer data and credentials - the information commonly stored in Customer Re-

[4] The attack surface variable does not capture research intensity as it does not allow for comparability of R&D support bodies among industries. As the total average employment is not used in the robustness setups, it only captures the attack surface's variance.

lationship Management (CRM) systems. Therefore, the last proxy variable is defined as *CRM usage rate* (d_{ij}) in a given country in a given industry. I believe that the rate of CRM usage will be higher in industries with more valuable customer data. The industrial distribution of CRM usage rate is demonstrated in Figure 4. The histogram of this variable can be found in Appendix A.2.

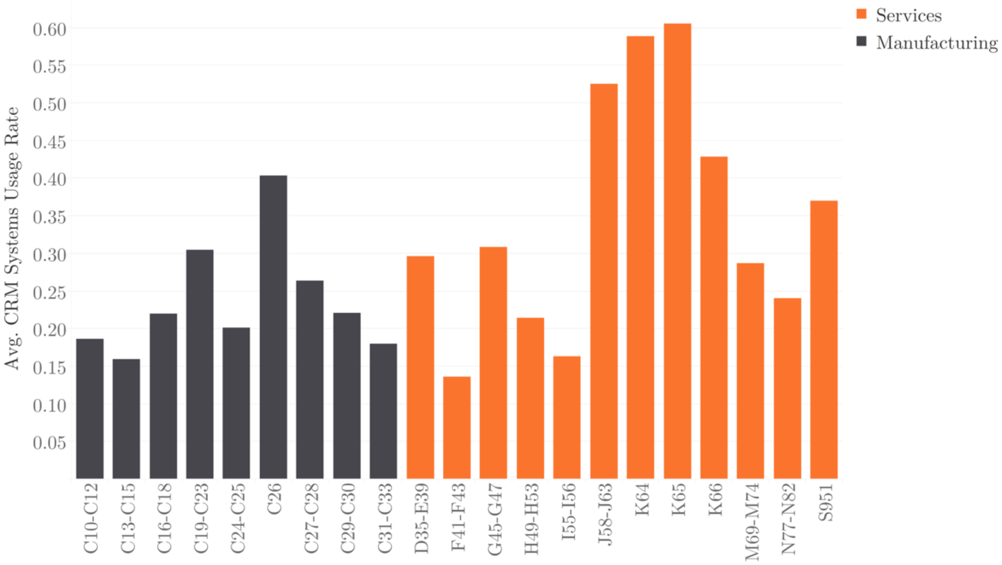

Figure 4: Industrial distribution of the CRM usage rate

The basic descriptive statistics for the dependent and independent variables can be found in Table 2.

Variable	Label	Obs.	Avg.	Std. Dev.	Min	Max
Information breaches rate	B_{ij}	351	.014	.021	0	.19
Share of R&D personnel	r_{ij}^{per}	224	.027	.059	0	.649
Share of Researchers	r_{ij}^{res}	224	.015	.042	0	.506
Share R&D support personnel	r_{ij}^{sup}	222	.012	.020	0	.143
CRM usage rate	d_{ij}	315	.285	.171	0	1

Table 2: Descriptive statistics for all main variables. Indices i and j denote a country and an industry respectively

3.1.3 Categorical variables

The study utilises two sets of categorical variables to control for heterogeneity among countries and industries. They account for unobserved differences across countries and industries by allowing individual intersects for each category.

Alongside with base sets of industry and country control variables, I use two additional sets of Eurostat indicators on high-tech industry and knowledge-intensive services (Eurostat, 2016). The manufacturing-specific variable is based on the categories of manufacturing industries, including high-technology, medium-high-technology, medium-low-technology, and low-technology. They were divided according to technological intensity. Table 13 of Appendix A.5 presents aggregated categories. Following a similar approach, the study utilises service-specific variables based on Eurostat's definitions of knowledge-intensive services (KIS) and less knowledge-intensive services (LKIS). The aggregation is demonstrated in Table 14 of Appendix A.5.

3.2 Methodology

To investigate the relationship between information leakage attacks rate and the research intensity based on the available dataset, the study adopts the Bayesian-frequentists hybrid method: multiple imputation to handle partial observations followed by fractional outcome regression analysis. The procedure is performed in three steps:

1. Multiple imputation, where M completed datasets (imputations) are simulated under the chosen imputation specification;

2. Complete-data analysis, where fractional regression analysis is performed on each imputation $m = 1, ..., M$;

3. Pooling, where the results from M completed datasets are combined.

This subsection offers detailed information on the underlying statistical methods used in each step.

3.2.1 Multiple imputation

The major challenge imposed by the data is partially missing observations: around 36% of observations are missing for research intensity variables and 10% for the CRM variable. The issue is mainly caused by the non-response in Eurostat's surveys. If not handled carefully, it could reduce statistical power, lead to biased estimates, and larger standard errors due to the smaller sample size.

There are two generally accepted approaches to handle missing data: (1) omission, by which partially incomplete observations are discarded from the analysis, and (2) imputation, by which missing observations are filled in.

The size of the dataset, the share of partially missing observations, and the lack of robustness for standard statistical inference techniques render the omission option rather inappropriate. On the other hand, imputation techniques are designed to extract all available information from partially observed data and minimise bias.

The research adopts the multiple imputation (MI) technique introduced by Rubin (1987). The multiple imputation approach is a Bayesian simulation-based statistical technique designed to handle partially observed data. The method was initially used to tackle non-response in medical surveys and handle datasets with up to 90% of observations missing (Madley-Dowd et al., 2019).

The method is superior to the single imputation counterpart (e.g. last observation carried forward, the sample mean imputation) as it accounts for the uncertainty of missing values. The single imputation approach treats imputed values as genuinely observed, which causes standard errors to be too small and causes a biased statistical inference. In reality, it is not possible to know the exact values of the missing data, and the MI technique accounts for all inherent uncertainty by injecting appropriate variability into the multiple imputed values (Sterne et al., 2009).

The research utilises two MI procedures designed for univariate and multivariate models.

The *univariate imputation* of research intensity independent variables is utilising the Gaussian normal regression model. The procedure is described below.

Consider a 351 × 1 vector $R^L = (r^L_{1j}, .., r^L_{ij}; r^L_{i1}, ..., r^L_{ij})'$, which records values for the research intensity variables: shares of research personnel ($L = per$), researchers specifically ($L = res$), or R&D support personnel ($L = sup$) in an industry i and country j. The variable then follows a Gaussian normal regression model:

$$r^L_{ij}|z_{ij} \sim N\left(z'_{ij}, \sigma^2\right), \tag{1}$$

$z_{ij} = \left(z_{ij1}, z_{ij2}, ..., z_{ijq}\right)$ is a vector of values of predictors of r^L_{ij} for observation ij, β is the $q \times 1$ vector of unknown regression coefficients, and σ^2 is the unknown scalar variance.

Consider then the division of $R^L = \left(R^{L\prime}_o, R^{L}_m{}'\right)$ into $n_0 \times 1$ and $n_1 \times 1$ vectors containing the complete and incomplete observations respectively. Similarly $Z = (Z_o, Z_m)$ is divided into $n_0 \times q$ and $n_1 \times 1$ submatrices.

The imputation then proceeds as follows to fill in R^L_m:

1. Specification (1) is fitted to the observed data (R^L_o, Z_o) to obtain estimates $\hat{\beta}$ and $\hat{\sigma}^2$ of the model parameters.

2. New parameters $\tilde{\beta}$ and $\tilde{\sigma}^2$ are simulated from their joint posterior distribution under the noninformative improper prior for a scaling parameter $\Pr(\beta, \sigma^2) \propto \frac{1}{\sigma^2}$. This is done in two steps:

$$\tilde{\sigma}^2 \sim \hat{\sigma}^2 \left(n_0 - q\right) / \chi^2_{n_0 - q},$$

$$\tilde{\beta} | \tilde{\sigma}^2 \sim N\left(\hat{\beta}, \tilde{\sigma}^2 \left(Z_o' Z_o\right)^{-1}\right);$$

3. The new set of imputed values is obtained by simulating missing values of R^L_{m1} from $N(Z_m \tilde{\beta}, \tilde{\sigma}^2 I_{n_1 \times n_1})$;

4. Steps 2 and 3 are repeated to obtain M sets of imputed values:

$$R^L_{m2}, R^L_{m3}, ..., R^L_{mM}.$$

Steps 2 and 3 follow the procedure of simulating the missing data from the posterior predictive distribution, described by Gelman et al. (2020).

The univariate imputation models utilise the rate of information leakage attacks in industry i, country j as the explanatory variable and both sets of categorical variables. Specification 1 can then be represented in the linear form:

$$r^L_{ij} = \beta_1 + \beta_2 B_{ij} + a_i + c_j + \epsilon_{ij}.$$

where B_{ij} is the share of enterprises that experienced information leakage attacks in an industry i and country j, and ϵ is the error term.

The multivariate models utilise the modified version of the MI technique to allow the *imputation of multiple variables* with missing observations simultaneously: multiple imputation by chained equations (MICE) developed by Van Buuren et al. (1999). MICE is more suitable than other multivariate imputation methods as it can handle arbitrary missing-data patterns (contrary to a standard multivariate imputation, which requires missing data patterns to be monotone).

The MICE procedure imputes multiple variables iteratively following a sequence of univariate multiple imputation models (one model for one partially observed variable). The procedure can be described as follows. For imputation variables $X_1, X_2, ..., X_q$ and complete predictors,

Z, imputed values are drawn from:

$$\begin{aligned}
X_1^{(t+1)} &\sim g_1\left(X_1|X_2^{(t)},...,X_p^t,Z,\phi_1\right),\\
X_2^{(t+1)} &\sim g_2\left(X_1|X_1^{(t+1)},X_3^{(t)},...,X_p^{(t)},Z,\phi_2\right),\\
&\cdots\\
X_q^{(t+1)} &\sim g_p\left(X_1|X_1^{(t+1)},...,X_{p-1}^{(t+1)},Z,\phi_q\right).
\end{aligned}$$

for iterations $t = 0, 1, ..., T$ until the imputed values converge at doomsday $t = T$. Here, ϕ_q, is the vector of corresponding model parameters with a uniform prior and, g_k is the corresponding univariate imputation model appropriate for imputing corresponding independent variables, X_k, where $k = (1, ..., q)$. As for specification models, the research utilises multivariate normal Gaussian regressions' specifications:

$$\begin{aligned}
&X_1 = \beta_0 + \beta_2 X_2 + ... + \beta_q X_q + \beta_{q+1} B_{ij} + c_j + a_i + \epsilon_{ij}\\
&X_2 = \beta_0 + \beta_1 X_1 + \beta_3 X_3 + ... + \beta_q X_q + \beta_{q+1} B_{ij} + c_j + a_i + \epsilon_{ij}\\
&\cdots\\
&X_q = \beta_0 + \beta_1 X_1 + ... + \beta_{q-1} X_{q-1} + \beta_{q+1} B_{ij} + c_j + a_i + \epsilon_{ij},
\end{aligned}$$

For the described procedure to lead to valid statistical inference, four more condition's must be satisfied: (1) the chosen imputation models should be *appropriate* for the further statistical analysis; (2) MI procedure must be *proper*; (3) MI procedure requires the missing data mechanism to be ignorable; and (4) a sufficient number of imputations, M, must be chosen (Rubin, 1987). These conditions are discussed in more detail in Appendix A.3.

Thus, the research adopts the multiple imputation procedure to handle the relatively big share of partially observed data. The procedure is designed to account for the uncertainty embedded in the analysis of partially observed data instead of simply generating a dataset with no missing data. Given that the above-mentioned conditions are satisfied, MI procedure should lead to to the unbiased statistical analysis described in the following subsection.

3.2.2 Complete-data analysis

I use a fractional outcome regression developed by Wedderburn (1974). The method belongs to the generalised linear models (GLM) class and is designed to fit outcome variables $\in [0, 1]$.

Fractional outcome regression is specified as GLM of the binomial family with robust standard errors and a link function $G(\cdot)$ which satisfies $G(z) \in [0, 1]$ for all $z \in \mathbb{R}$:

$$E(Y_i|X_i) = G(X_i\beta). \tag{2}$$

Cumulative distribution functions (CDF) are often utilised as the link function for fractional regression, with two most common options being the logistic function $G(z) = \Lambda(z) = \frac{\exp(z)}{1-\exp(z)}$, and the probit function $G(z) = \Phi(z)$, where $\Phi(z)$ is the CDF of a standard normal distribution. However, it can be essentially any function bounded between 0 and 1, not necessarily a CDF.

Additionally, GLM class models require fewer assumptions to be satisfied to achieve unbiased results (in comparison with the OLS regression) – see Appendix A.4 for the discussion. Given that data cases are independently distributed, it is merely enough to specify the conditional mean of the variable of interest correctly to achieve an unbiased statistical inference.

The research utilises two linear specifications that will be estimated on the imputed data: S^U - univariate specification and S^M - multivariate specification:

Univariate specifications:

$$E(B_{ij}|r_{ij}^L, c_j, a_i) = G(\beta_0 + \beta_1 r_{ij}^L + c_j + a_i) = G(S^U), \tag{3}$$

which is used to assess individual effects of the research intensity variables on the rate of information leakage attacks.

Multivariate specifications:

$$E(B_{ij}|r_{ij}^L, d_{ij}, c_j, a_j) = G(\beta_0 + \beta_1 r_{ij}^L + \beta_2 d_{ij} + ... + c_j) = G(S^M), \tag{4}$$

which is used to assess the effects of the research intensity variables on the rate of information leakage attacks in the presence of an additional CRM usage rate variable (in the base multivariate specification). The variable is added to the model to control for the variance induced by opportunistic attacks in which the leakage of personal records of clients is highly likely to predominate. Note that the multivariate specification lacks the industry controls. They were removed from the specification due to collinearity: the share of enterprises that use CRM systems is industry-dependent. Appendix A.6 provides the variance inflation factor (VIF) and estimated within-industry variance of the CRM variable.

Following good practise procedures proposed by Villadsen and Wulff (2019) it is now necessary to (1) verify the conditional mean specifications and (2) choose an appropriate link function.

The conditional mean specification can be tested using the Regression Equation Specification Error Test (RESET) developed by Ramsey (1969). It tests whether non-linear combinations of the fitted values (e.g. $\hat{B}_{ij}^P$, where P is the maximum power of the non-linear combination) are useful to explain the outcome variable. If non-linear combinations have any influence on the outcome variable, then the tests suggest that the model is misspecified and it may be better approximated via some non-linear function.

Tests results for both univariate and multivariate specifications under different link functions are provided in Appendix A.7. All four tests fail to reject the null hypothesis that the fractional outcome model is specified correctly.

I follow the canonical approach and discriminate between *logit* and *probit* link functions. To do so, the research utilises the test of econometric specification in the presence of alternative specification (Davidson and Mackinnon, 1981). Test results for univariate and multivariate specifications are provided in Appendix A.8. Both performed tests fail to reject the null hypothesis implying the usage of *probit* link function. However, both link functions lead to extremely close estimates, and the choice of link function would not influence the results.

The binomial GLM with probit link function then fits coefficients via maximising the following quasi-likelihood functions for every imputed dataset:

$$\log L = \sum_{i=1}^{N^{in}} \sum_{j=1}^{N^{c}} \left(B_{ij} \log\{\Phi(S)\} + (1 - B_{ij}) \log\{1 - \Phi(S)\} \right), \tag{5}$$

where $N^{in/c}$ is the number of industries and countries represented in the research, S is the linear specification with $S = S^U$ for univariate specification and $S = S^M$ for multivariate specification, and Φ is a CDF of standard normal distribution. The maximisation is performed numerically following the Newton–Raphson iterative method (McMullen, 1987).

The statistical significance of a single coefficient is determined based on the standard score (z) statistic. However, the coefficients of fractional regression outcome with probit link function have limited interpretability. They can be interpreted as the difference in the standard scores associated with a unit change in the regressor. Therefore, raw models' results can only reveal statistically significant effects and signs.

To achieve interpretable and comparable results, I estimate the marginal effects of each variable. For univariate-specification models, the marginal effect estimation demonstrates how the outcome (the conditional mean of the rate of information leakage attacks) variable changes when a specific explanatory variable changes, $\frac{\partial Y}{\partial X}$. For models with a probit link function, the marginal effect is estimated as follows:

$$\frac{\partial Y}{\partial X_i} = \beta_i \phi(S), \tag{6}$$

where $\phi(\cdot)$ is a standard normal probability density function.

The procedures described above are performed for each imputed dataset, and then the results are combined in the following pooling step.

3.2.3 Pooling step

Given that multiple imputations are proper, unbiased estimates can be achieved by a simple averaging of the estimates derived by fitting fractional outcome regression models on each of the 40 imputed datasets separately. Pooled estimates are then interpreted similarly as the coefficients of a classic dataset model. All results reported in the next section are based on the pooled coefficients.

3.3 Methodology discussion

This subsection provides a discussion on the chosen methodology.

3.3.1 Controlling for unobserved effects

The study utilises the conventional fixed effects method to account for unobserved heterogeneity of industries and countries. Alternatively, it is possible to model heterogeneity using the mixed-effects method (also called multi-level or hierarchical models). This family of models is especially attractive for research designs in which individuals are organised hierarchically at more than one level (e.g. at the country and industry level in this study).

Unfortunately, the data adopted for this study do not allow for hierarchical mixed-effects models. This happens due to the absence of replicated measurements at the industry-country level (i, j). The hierarchical model is not identified for singleton pattern data because the random intercepts estimated for the lower level (which is the ”industry” level in the case of this research) are confounded with the overall error terms (Stata Corporation, 2013).

3.3.2 Handling missing data

The unbiasedness of the results depends on the validity of the imputation model. The imputation method might not lead to an unbiased result if not handled correctly. Several pitfalls might lead to a biased statistical inference.

Misspecification of regressors mainly exists due to the exclusion of the dependent variable from a linear specification of the imputation equation. An outcome variable might possess additional information about the missing values of the explanatory variables. Failure to include the outcome variable into the imputation model weakens the association between the dependent variable and imputed regressors (von Hippel, 2013). The problem was avoided in the proposed MI procedure, and every imputation model includes the dependent variable.

Imputation of non-normally distributed variables is another concern that is discussed in the literature. The normality assumption of Gaussian regression, in general, applies to the error terms. The assumption is satisfied, as demonstrated in Appendix A.9, so it is not an issue in this particular research.

More debatable is whether the assumed imputation model is appropriate to handle the *imputation of bounded data* (all imputed variables in this study are $\in [0, 1]$). The main concern with the chosen specification is that the predicted values may fall outside the range of the imputed variable. It is essential to remember that the MI procedure is not utilised to recreate missing observations. Instead, it achieves an unbiased statistical inference by accounting for uncertainty embedded in the missing observations. Imputed values are, therefore, not required to be bounded.

Moreover, the study by Rodwell et al. (2014), which investigated the efficiency of imputation methods of bounded variables based on 1000 incomplete simulated datasets, suggested that the Gaussian MI procedure is superior to any conventional alternatives. Post-imputation rounding, truncated normal regression, zero-skewness log-transformation, and predictive mean matching are highly sensitive to the skewness of the imputed variables and prone to reducing the variance of imputed values.

To conclude, the potential bias may arise from misspecification of MI procedure or imputing data with a non-random missing pattern. Following Rubin's recommendations, the proposed MI procedure is carefully crafted to achieve proper multiple imputation. The Gaussian specification is chosen based on a comprehensive empirical study of alternative imputation methods of bounded variables. The inference based on complete cases is reported in Appendices.

3.3.3 Analytical model

Four main analytical models are used in statistic/econometrics literature to model fractions (from the most popular to the least popular): (1) OLS regression, (2) OLS regression with transformed dependent variable, (3) Tobit models, and (4) fractional regression (Villadsen and Wulff, 2019). Additionally, the excess share of zero observations might suggest using zero-inflated models. However, the four alternative methods have theoretical and practical drawbacks discussed below.

The main problem with *OLS regressions* is that it ignores the boundaries of the outcome variable. Modelling an outcome variable $\in [0, 1]$ with Gaussian normal regression will most certainly require a violation of the linearity assumption. Another problem is the error term's heteroskedasticity, which requires special treatment (e.g. estimation of robust standard errors). Furthermore, OLS is highly sensitive to skewed variables (which is not the case with fractional outcome regressions with non-linear link functions). However, this model can still be used to investigate the signs and sizes of the relationships, though the procedure must be performed carefully.

The most popular *transformation of the fractional dependent variable* is a *log-transformation*, which might normalise the data and solve the problem with outcome boundaries. The main problem with this method is that it does not perform well with highly skewed dependent variables with a substantial share of zero observations. It requires some ad-hoc solution to make the log-transformation of zero numbers possible (e.g. addition of some small negligible number), which may lead to a biased inference. Moreover, the model requires substantial efforts to verify the underlying assumption of the OLS regression.

The Tobit model was initially developed to model censored outcomes (e.g. outcomes which values are only partially known) (McDonald and Moffitt, 1980). Fractional outcomes are not censored in any way but instead defined only on the interval of $\in [0, 1]$. Censored outcomes and corner solution responses (like fractional outcomes) are often confused. The application of the Tobit model to the fractional outcomes implies that there exist some observations of the dependent variable outside the unit range, which is not the case (Wooldridge, 2010). This leads to difficulties in the interpretation of the Tobit model coefficients. The marginal effect is only reflected on the coefficients of the latent (unobserved) variable, which is non-existent. Moreover, the model assumes homoskedasticity and normality of error terms - conditions likely to be violated when dealing with fractional outcomes.

Zero-inflated model might be another reasonable choice given the share of zero observations. However, those models are generally harder to set due to the necessity of setting the predictor for zero observations appearance, which, conditional on the nature of the data, might not be feasible. Moreover, those models are generally used to tackle the problem of overdispersion, which is not the case for a given data (see Table 2: the zero-inflated Poisson model would be appropriate if mean of the dependent variable is much smaller than the variance). Additionally, they cannot be applied to fractional variables directly and would require an additional estimation step (e.g. the two-part composite model in Liu and Xin, 2014), which would further complicate the interpretation and tracking of assumptions (Williams, 2019).

Fractional outcome regression is most suitable for our data. It performs extraordinarily well in the presence of heteroskedasticity, provides valid statistical inference for the bounded and skewed outcomes, and produces interpretable and adequate estimates. Moreover, it is agnostic about the moments of the outcome and requires only the correct specification of the conditional mean.

4 Results

This section presents the results of the study. The study adopts the univariate and multivariate specifications described in the section above. Both sets of specifications are tested with three proposed research intensity proxy variables.

4.1 Fractional regressions results

Table 3 presents results of MI univariate fractional outcome regressions with all three research intensity proxy variables. Column (I) of the table reports base regression results with no control variables, columns (II) and (III) report regression results with the country and industry controls included respectively, and column (IV) reports the results with both control variables sets. Complete case regression analysis tables can be found in Appendices A.10, A.12, and A.14.

The upper section of Table 3 contains results of univariate regression models with the share of research personnel as the main explanatory variable. Research personnel's coefficients are positive in all four specifications and statistically significant at 5% level in three out four specifications (with the the specification with the country only controls as an exception).

The middle section of Table 3 reports the results with an alternative measure of research intensity: the share of researchers. As previously, models report positive coefficients next to the primary explanatory variable. However, only coefficients for the specifications with industry only controls, or both controls are statistically significant at 5% level.

Finally, the bottom section reports the results of models with the share of research support personnel as the main explanatory variables. The coefficients are positive and statistically significant in all four specifications at 1% level.

It is evident that regardless of the chosen research intensity variables, models specified with both country and industry controls display positive and statistically significant (at 1% level) coefficients.

Raw fractional probit models' coefficients, however, are not further interpretable. Results only indicated the positive sign of the correlation between the alternative research intensity proxy variables and the information leakage attacks rate. The interpretation of results is included in the following analysis of coefficients' marginal effects.

Table 3: Regression results: fractional probit regressions with share of research personnel as an independent variable

	I	**II**	**III**	**IV**
	Base	Country	Industry	All
	model	controls	controls	controls
Research personnel MI models				
r^{per}	1.377**	1.011*	1.820**	1.285**
	(2.23)	(1.95)	(2.44)	(2.43)
Constant	-2.233***	-2.100***	-2.252***	-2.110***
	(-67.86)	(-14.56)	(-24.76)	(-13.69)
Researchers MI models				
r^{res}	1.366*	1.052	1.747**	1.315**
	(1.68)	(1.53)	(1.97)	(2.07)
Constant	-2.212***	-2.086***	-2.239***	-2.101***
	(-70.84)	(-14.27)	(-24.96)	(-13.38)
R&D support personnel MI models				
r^{sup}	6.225***	3.838***	9.106***	5.521***
	(3.81)	(2.80)	(4.38)	(3.67)
Constant	-2.273***	-2.110***	-2.279***	-2.118***
	(-68.83)	(-15.78)	(-23.31)	(-14.35)
Observations	351	351	351	351

z statistics in parentheses; Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

C10-C12 and BE are chosen as base levels; See Appendices A.11, A.13, A.15 for complete tables

4.2 Marginal effects analysis

Table 4 presents average marginal effects estimates (as shown by equation (6)) for all three alternative research intensity proxy variables estimated for models with all controls included. As expected, all presented coefficients are positive and statistically significant at 5% level (or even at 1% level for the share of research support personnel). It is now possible to interpret the coefficients in the same fashion as coefficients of a classic OLS regression.

Table 4: Average marginal effects of research intensity proxy variables in models with all controls

	$\partial B/\partial r^{per}$	$\partial B/\partial r^{res}$	$\partial B/\partial r^{sup}$
Marginal effect	0.0432**	0.0442 **	0.185***
	(2.38)	(2.04)	(3.55)

z statistics in parentheses; Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

In the model with the share of research personnel as the main independent variable, a 10 percentage points increase in the share of research personnel is associated with a 0.432 percentage point increase in the information leakage attacks rate. This marginal effect is relatively similar to that demonstrated by the model with the share of researchers as the primary independent variable: a 10 percentage points increase in the share of researchers is associated with a 0.442 percentage points increase in the information leakage attacks rate.

The model with the share of research support personnel delivers much more prominent results. A 10 percentage points increase in the share of research support personnel is associated with a 1.85 percentage points increase in the information leakage attacks rate.

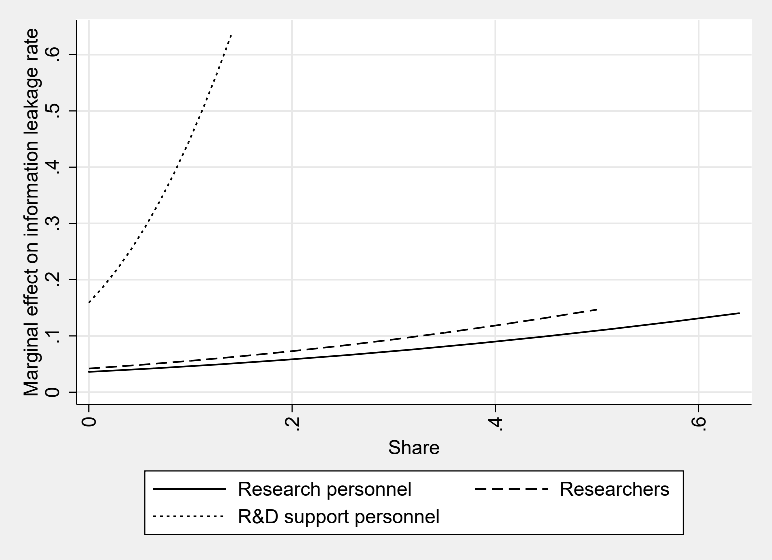

Figure 5: Point estimates of the marginal effects of the research proxy variables

The substantial difference in the magnitude is also evident from the shape of the partial derivative function. As illustrated in Figure 5, all three marginal effects follow quadratic growth as the share of the corresponding research body increases. In particular, the marginal effect of the share of research support personnel has a much steeper shape than the marginal effects of the other two proxy variables. In its peak ($r^{sup} = 14\%$), a 1 percentage point increase in the share of research support personnel is associated with a 0.634 percentage point increase in the information leakage attack rate. Meanwhile, the marginal effects of the shares of research personnel and researchers have a similar, yet more gradual incline ($\partial B/\partial r^{per} = 0.14$ at $\max r^{per} = 64\%$ and $\partial B/\partial r^{res} = 0.15$ at $\max r^{res} = 50\%$).

4.3 Targeted and opportunistic attacks

As stated above, most data breaches are affiliated with opportunistic, financially motivated breaches responsible for the majority of compromised customers' data.

Table 5: Regression results: fractional probit regressions with research poxies, CRM variable and country controls

	I	**II**	**III**
	R&D personnel	Researchers	R&D support
	model	model	model
MI models			
r^{per}	0.562		
	(1.22)		
r^{res}		0.410	
		(0.53)	
r^{sup}			3.041**
			(2.35)
d	0.715***	0.762***	0.641***
	(3.63)	(3.78)	(3.48)
Constant	-2.409***	-2.416***	-2.395***
	(-14.62)	(-14.64)	(-15.37)
Observations	351	351	351

z statistics in parentheses; Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

C10-C12 is chosen as base level; See Appendices A.16 and A.17 for complete tables

It is possible that the chosen dependent variable accounts mostly for opportunistic attacks, and the discovered relationships are spurious and are not related to the targeted information leakage attacks. To control for this possibility, the study estimates multivariate specifications, which include a proxy variable that represents a measure of how much end-user data are available to be stolen in an industry - the CRM usage rate. Results for the extended multivariate specification for all three research proxy variables (as demonstrated by equation (4)) are presented in Table 5. The table for complete case analysis can be found in Appendix A.16.

One should approach the interpretation of the coefficients estimated for the CRM variable with caution. As industry control variables were excluded from the following sets of models due to collinearity, the CRM variable might absorb the variance generated by industrial heterogeneity unrelated to the CRM usage rate.

In all three presented models, the coefficients for the CRM usage rate are positive and statistically significant at 1% level. All three models report positive signs of coefficients for the research intensity variables. Nonetheless, only the share of research support personnel has a statistically significant coefficient at 5% level.

As suggested earlier, the variable might indicate mature high-tech manufacturing or service industries with a substantial body of commercialisable research. I now consider the model presented in column (3) of Table 5 in further detail. The average marginal effects of the share of R&D support personnel and the CRM usage rate on the information leakage attacks rate are presented in Table 6.

Table 6: Average marginal effects in extended models

	$\partial B/\partial r^{sup}$	$\partial B/\partial d$
Marginal effect	0.103**	0.0217***
	(2.28)	(3.34)

z statistics in parentheses; Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

The average marginal effect of the share of R&D support personnel in the MI models is positive and statistically significant at 5% level: a 10 percentage points increase in the share of R&D support personnel is associated with a 1.03 percentage points increase in the information leakage attack rate. Likewise, the raw coefficient, the average marginal effect of the CRM usage rate on the information leakage attacks rate is positive and significant at 1% level. A 10 percentage points increase in the CRM usage rate is associated with a 0.22 percentage points increase in the information leakage attack rate.

The average marginal effect of the share of R&D support personnel is larger than that of the CRM system usage rate. Still, direct comparison of the effects is case-specific due to the industrial heterogeneity of variables' means. Point estimates of the marginal effects of both variables are demonstrated in Figure 6. As with the univariate models, the measured marginal effects of the share of R&D support personnel and the CRM usage rate exhibit quadratic growth. The marginal effect function for the research intensity proxy is much steeper than the marginal effect function of CRM usage rate. It means that more research-intensive industries experience a larger effect of an additional percentage point of R&D support personnel.

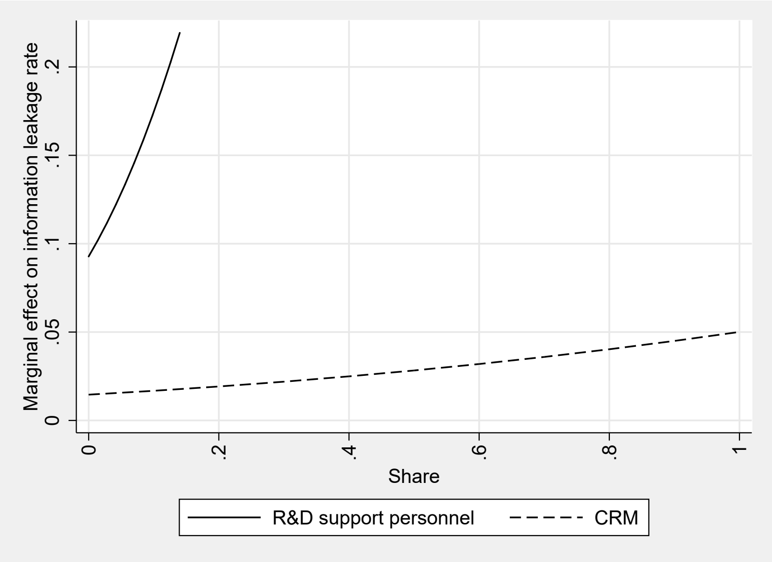

Figure 6: Marginal effects of the share of R&D support personnel and CRM usage rate point estimates

4.3.1 Technological environments

This subsection investigates the relationships between the primary dependent variable, the research intensity proxy, and the CRM usage rate in diverse technological environments. The multivariate specification is estimated separately for manufacturing, services, and corresponding technological and knowledge-intensity levels. Examining coefficients' magnitudes in different technological environments might provide supportive evidence for the underlying causal mechanism.

The manufacturing and service industries are divided according to the Eurostat indicators on high-tech industry and knowledge-intensive services provided in Appendix A.5. Note that to maintain a healthy amount of observations, the manufacturing industries are grouped into two groups instead of four: the high-tech group absorbs high-technology and medium-high-technology groups. In contrast, the low-tech group includes medium-low-technology and low-technology industries.

If the causal mechanism is plausible, two associations identified in Subsection 4.3 should exhibit different patterns depending on the technological environment. It is expected that the information leakage attacks in manufacturing industries are closely associated with research intensity, and the association would be stronger in high-tech manufacturing industries. On the other hand, relatively less R&D intensive and more end-user data-reliant service industries would experience information leakage attacks more closely related to CRM usage. Capturing this pattern might suggest that manufacturing industries are prone to experience targeted attacks, while services suffer from opportunistic, financially driven attacks.

The specifications are estimated in technological clusters. This should reduce the potential heterogeneity captured by the CRM usage rate variable in the full-sample estimation and reveal less biased coefficients. Fractional regression results for manufacturing industries are presented in Table 7.

Table 7: Regression results: fractional probit regressions on imputed data with research proxy and CRM variable for manufacturing industries

	I	**II**	**III**
	All	Low-tech	High-tech
	manufacturing	manufacturing	manufacturing
r^{sup}	4.488***	3.200	7.206***
	(2.77)	(0.66)	(3.13)
d	-0.776	-0.651	-1.337
	(-1.41)	(-0.87)	(-1.29)
Constant	-1.549***	-1.705***	-1.223**
	(-5.43)	(-4.98)	(-2.24)
Observations	152	81	71

z statistics in parentheses; Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

C10-C12 is chosen as a base level; See Appendix A.18 for the full table

Coefficients of the CRM system usage are not significant in all demonstrated specifications. It suggests that opportunistic attacks do not contribute to the information leakage attacks rate variance in manufacturing industries. Coefficients for the share of R&D support personnel are positive and significant at 1% for high-tech and all manufacturing industries on average. Neither the CRM system usage nor the research intensity proxy demonstrate statistically significant coefficients for low-tech industries. Results suggest that information leakage attacks in the manufacturing industries are more likely to be targeted, as they are strongly associated with the R&D intensity proxy while demonstrating no association with the data reliance proxy.

Table 8 presents results of the average marginal effects of the share of R&D support personnel. As expected, the marginal effect is not significant in low-tech industries. The average marginal effect of the share of R&D support personnel in the high-tech industries is substantially larger than the effect in all manufacturing industries. A 10 percentage points increase in the share of research support personnel is associated with a 1.26 percentage points increase in the information leakage attacks rate in all manufacturing industries on average. In high-tech industries, a 10 percentage points increase in the share of research support personnel is associated with a 1.92 percentage points increase in the information leakage attacks rate. The results align with the previous finding that the marginal effect function of the share of R&D support personnel is convex and steep.

Table 8: Average marginal effect in manufacturing models

	I	**II**	**III**
	All manufacturing	Low-tech manufacturing	High-tech manufacturing
$\partial B/\partial r^{sup}$	0.126***	0.091	0.192***
	(2.70)	(0.66)	(3.38)

z statistics in parentheses; Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Results of fractional outcome regressions for the subsamples of service industries are presented in Table 9. The models demonstrate the opposite outcomes to those seen in the manufacturing environment. The coefficients for the share of R&D support personnel are insignificant in all demonstrated specifications. In contrast, the CRM usage rate coefficients are significant at the 1% level for knowledge-intensive and all knowledge-based services. Table 10 offers average marginal effects estimates for the CRM variable: a 10 percentage points increase in the CRM usage rate is associated with a 0.274 percentage points increase in the information leakage attacks rate in all knowledge-based service industries. The magnitude of the average marginal effect is more pronounced in the knowledge-intensive service industries, with a 10 percentage points increase in the CRM usage rate associated with a 0.420 percentage points increase in the rate of information leakage attacks. Those results suggest that, contrary to manufacturing industries, information leakage attacks in the service industries are more likely to be opportunistic, as the dependent variable demonstrates a strong association with the data reliance proxy and no association with the research intensity proxy. Moreover, knowledge-intensive service industries (which are generally more data reliant) tend to be affected by the phenomenon more than their less knowledge-intensive counterparts.

Table 9: Regression results: fractional probit regressions on imputed data with research proxy and CRM variable for service industries

	I	**II**	**III**
	All knowledge-based services	Less knowledge-intensive services	Knowledge-intensive services
r^{sup}	3.698*	10.70*	1.862
	(1.82)	(1.89)	(1.01)
d	0.729***	0.238	0.985***
	(3.48)	(0.43)	(4.55)
Constant	-2.573***	-2.345***	-2.698***
	(-23.41)	(-8.96)	(-27.33)
Observations	199	105	94

z statistics in parentheses; Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

C10-C12 is chosen as a base level; See Appendix A.19 for the full table

Table 10: Average marginal effect in service models

	I	**II**	**III**
	All knowledge-based services	Less knowledge-intensive services	Knowledge-intensive services
d	0.0274***	0.00776	0.0420***
	(3.36)	(0.42)	(4.45)

z statistics in parentheses; Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Therefore, as suggested in the above analysis, the information leakage attacks rate in manufacturing industries is mainly attributed to targeted attacks (potentially cyberespionage), while the attacks in service industries are primarily attributed to opportunistic attacks. However, the tests only provide supporting evidence for causal relationships instead of establishing them.

4.4 Alternative explanations and robustness checks

The study focuses on the share of research support personnel as the primary research intensity proxy. The variable demonstrates statistically significant and positive coefficients in all univariate and multivariate specifications and possesses the largest average marginal effect among the alternatives. The findings align with the proposed causal mechanism, but they do not rule out other potential causal paths. In search of additional supporting evidence, I perform several robustness tests to verify the behaviour of the primary variable in the presence of auxiliary variables that stand for alternative causal mechanisms.

Table 11: Regression results: robustness checks

	I	**II**	**III**	**IV**	**V**
	A.Surface	Turnover	E-comm	Sysadmin	All
	controls	controls	controls	controls	controls
r^{sup}	3.277***	3.235***	3.783***	2.999**	3.981**
	(2.68)	(2.72)	(3.25)	(2.01)	(2.53)
d	0.585***	0.581***	0.285	0.540***	0.310
	(3.02)	(2.96)	(1.34)	(2.88)	(1.45)
A.Surface	-18E-7				703E-9
	(-0.17)				(0.07)
Turnover		0.000120			0.0000710
		(0.32)			(0.19)
E-comm			0.00768**		0.00787**
			(2.22)		(2.30)
Sysadmin				0.0822	-0.0711
				(0.33)	(-0.29)
Constant	-2.370***	-2.371***	-2.464***	-2.380***	-2.458***
	(-15.16)	(-15.08)	(-14.51)	(-14.61)	(-14.41)
Observations	351	351	351	351	351

z statistics in parentheses; Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

C10-C12 is chosen as a base level; See Appendix A.20 for the full table

Observe the R&D support personnel variable might capture two effects simultaneously: (1) the values of the research ideas, and (2) the extended attack surface. Column (I) of Table 11 shows a specification with an additional attack surface variable that represents an average quantity of the R&D support employees in the enterprise in a given industry in a given country. The additional variable hardly has any effect on the original variables' coefficients, and does not support the claim.

Another alternative scenario is that cyberattacks are generally performed on rich industries, which, in turn, can afford a larger R&D support body. It means that both information leakage attacks and research intensity might be confounded by the company's average turnover in a given industry in a given country. This alternative is tested in Column (II) of Table 11. The newly included variable does not have any effect on the results.

Column (III) of Table 11 demonstrates the results of the specification with included e-commerce variable, which shows the share of enterprises in a given country in a given industry that have any electronic sales. It tests for the alternative scenario in which the most attacked industries are the ones that participate in e-commerce. Again, the additional variable does not significantly affect the research intensity variable. However, it renders the CRM variable insignificant. This phenomenon can be easily explained: the CRM and e-commerce variables are highly correlated. The e-commerce variable should not significantly change the results for the research intensity proxy as its magnitude is mainly attributed to manufacturing industries, which have relatively less engagement in e-commerce.

It might also be the case that research intensity proxies are highly correlated with the number of system administrators in a given country in a given industry. It is a known fact that system administrators are the most common threat actor in the internal industrial espionage cases (Verizon Business, 2020). Assuming that the same might be true for attacks performed by the external actors, it could be the case that the effects of system administrators might confound independent and dependent variables. Column (IV) displays results under the specification, which utilises an additional variable, *sysadmin*, that stands for the share of enterprises with an in-house system administrator in a given country in a given industry. The coefficient next to this variable is insignificant and does not affect the original variables' coefficients.

Column (V) displays results under the specification with all additional auxiliary variables included. Again, the main result of the research intensity variable demonstrates robustness in a setting with all additional controls and even a slight increase in its magnitude.

Another potential source of endogeneity is research personnel engaged in cybersecurity research. However, cybersecurity researchers constitute only a tiny fraction of the overall employment. For example, only $43,000$ employees worked in the cybersecurity sector in the UK in 2019, which was less than 0.25% of the overall workforce. Moreover, only a fraction of them engaged in the research and development (Donaldson et al., 2020).

The additional tests suggest that the main result is robust, even when controlling for alternative scenarios. The tests provide no evidence suggesting alternative causal mechanisms. However, those possibilities cannot be ruled out entirely until the causality between the information leakage attacks rate and the research intensity proxy is established, a challenge to overcome in future research.

5 Conclusion

This paper studies the possible determinants of industrial information leakage attacks. Such attacks are often used as a means of unfair competition and are associated with industrial cyberespionage.

The study theorises that a fair share of targeted information leakage attacks happens as a part of the R&D race and, therefore, research-intensive industries are expected to be particularly susceptible to industrial cyberespionage. However, targeted attacks and industrial espionage constitute only a part of the picture.

The study implements a multi-stage approach in which the leading hypothesis is tested in a series of univariate and multivariate specifications. The multi-stage design allows assessing the performance of three different research proxy variables both solely and in the presence of auxiliary variables that control for opportunistic attacks and alternative causal mechanisms.

The results of the univariate model of Eurostat data are consistent with the proposed causal mechanism. Regardless of the chosen research intensity proxy, the models demonstrate positive and statistically significant coefficients after controlling for industry and country unobserved effects. Nevertheless, the variables differ considerably in their magnitudes. The share of research support personnel has a four times larger average marginal effect than the other two research intensity proxies. Furthermore, this is the only research intensity proxy that maintains a significant coefficient in the multivariate model.

The multivariate model reveals that user data reliance is another major contributor to the information leakage attacks rate variance. The study distinguished two significant associations, potentially connected with targeted and opportunistic attacks respectively.

Estimating the multivariate specification on the subsamples of manufacturing, services, and their respective technological levels reveals that the discovered associations are industry-specific. Manufacturing industries (that are generally more R&D oriented) experience cyberattacks associated with research intensity. The correlation is more prominent in the subsample of high-tech manufacturing industries, which can be considered additional supporting evidence favouring the proposed theory. The coefficients for user data reliance are insignificant in all models estimated for manufacturing industries. The situation is the opposite in the service industries, which experience information leakage attacks associated solely with end-user data reliance. Similarly, the effect becomes more salient in the subsample of knowledge-based services.

The association between the information leakage attack rate and research intensity in manufacturing industries suggests that a major share of attacks might be targetted (potentially cyberespionage). Moreover, the attack rate is sensitive to the manufacturing industry's research intensity and technological level. Those results provide supporting evidence in favour of the proposed causal mechanism. However, the causality is still to be established in future research, and alternative scenarios should not be ruled out completely.

To conclude, the study provides supporting evidence that research-intensive industries are more likely to become victims of information leakage attacks. This phenomenon could be associated with unfair competition and industrial espionage.

References

Bardopoulos, J. (2020). *Cyber-insurance pricing models.* PhD thesis, University of Cape Town.

Bland, M. (2006). *An introduction to medical statistics* .

Böhme, R. (2005). Cyber-Insurance Revisited. In *Workshop on the Economics of Information Security, Kennedy School of Government, Cambridge, MA, USA.*

Böhme, R. and Kataria, G. (2006). On the limits of cyber-insurance. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics).*

Coe, D. T. and Helpman, E. (1995). International R&D spillovers. *European Economic Review*, 39(5):859–887.

Collins, L. M., Schafer, J. L., and Kam, C. M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures.

Comanor, W. S. (1965). Research and Technical Change in the Pharmaceutical Industry. *The Review of Economics and Statistics*, 47(2):182.

Crete-Nishihata, M., Dalek, J., Maynier, E., and Scott-Railton, J. (2018). Spying on a budget.

Inside a Phishing Operation with Targets in the Tibetan Community Suggested Citation Copyright. Technical report.

Davidson, R. and Mackinnon, J. G. (1981). Several Tests for Model Specification in the Presence of Alternative Hypotheses. Technical Report 3.

DCMS (2020). Cyber Security Breaches Survey 2020. Technical report.

Donaldson, S., Navin Shah, J., Pedley, D., Crozier, D., and Furnell, S. (2020). UK Cyber Security Sectoral Analysis 2020. Technical report.

European Commission (2008). *NACE Rev. 2 – Statistical classification of economic activites in the European Community.*

Eurostat (2010a). Digital Economy and Society - Dataset.

Eurostat (2010b). Science, Technology and Innovation - Dataset.

Eurostat (2010c). Structural Business Statistics & Global Business Activities - Dataset.

Eurostat (2016). High-tech industry and knowledge-intensive services (HTEC). Technical report.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., Rubin, D. B., Carlin, J., Stern, H., Rubin, D., and Dunson, D. (2020). Bayesian Data Analysis Third edition (with errors fixed as of 13 February 2020). Technical report.

Graham, J. W., Olchowski, A. E., and Gilreath, T. D. (2007). How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prevention Science*, 8(3):206–213.

Gruber, W., Mehta, D., Vernon, R., Cremer, H., Gertz, J., Helfand, H., McBride, C., Morrow, E., Perry, L., and Teich, A. (1966). The R&D factor in international trade and international investment of United States industries.

Herath, H. and Herath, T. (2007). Cyber-Insurance: Copula Pricing Framework and Implication for Risk Management. *2007 Workshop on the Economics of Information Security (WEIS).*

Hoo, K. J. S. (2000). How Much Is Enough? A Risk-Management Approach to Computer Security.

IBM (2020). Cost of a Data Breach Study.

Innerhofer-Oberperfler, F. and Breu, R. (2010). Potential Rating Indicators for Cyberinsurance: An Exploratory Qualitative Study. In *Economics of Information Security and Privacy.*

Jones, C. I. (2003). Human capital, ideas and economic growth. In *Finance, Research, Education and Growth*, pages 51–74. Palgrave Macmillan.

Juniper Research (2019). Business Losses to Cybercrime Data Breaches to Exceed $5 trillion.

Leita, C., Pham, V. H., Thonnard, O., Ramirez-Silva, E., Pouget, F., Kirda, E., and Dacier, M. (2008). The Leurre.com Project: Collecting internet threats information using a worldwide distributed honeynet. In *Proceedings - WOMBAT Workshop on Information Security Threats Data Collection and Sharing, WISTDCS 2008.*

Lewis, P. A. (2017). Technician Roles, Skills, and Training in Industrial Biotechnology: An Analysis. *SSRN Electronic Journal.*

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data.* Wiley, 2nd edition.

Liu, W., Tanaka, H., and Matsuura, K. (2007). Empirical-Analysis Methodology for Information-Security Investment and Its Application to Reliable Survey of Japanese Firms. *IPSJ Digital Courier.*

Liu, W. and Xin, J. (2014). Modeling Fractional Outcomes with SAS .

Liu, X. (2015). *Methods and applications of longitudinal data analysis.* Elsevier.

Madley-Dowd, P., Hughes, R., Tilling, K., and Heron, J. (2019). The proportion of missing data should not be used to guide decisions on multiple imputation. *Journal of Clinical Epidemiology*, 110:63–73.

McAfee (2020). The Hidden Costs of Cybercrime. Technical report.

McDonald, J. F. and Moffitt, R. A. (1980). The Uses of Tobit Analysis. *The Review of Economics and Statistics*, 62(2):318.

McMullen, C. (1987). Families of Rational Maps and Iterative Root-Finding Algorithms. *The Annals of Mathematics*, 125(3):467.

Moore, T. and Clayton, R. (2007). Examining the impact of website take-down on phishing. *Proceedings of the anti-phishing working groups 2nd annual eCrime researchers summit on - eCrime '07*, pages 1–13.

Moore, T. and Clayton, R. (2008). The consequence of non-cooperation in the fight against phishing. *eCrime Researchers Summit, eCrime 2008.*

Moore, T. and Clayton, R. (2011). The Impact of Public Information on Phishing Attack and Defense. *Communications & Strategies*, (81):45–68.

Moore, T., Richard, C., and Ross, A. (2009). The Economics of Online Crime. *Journal of Economic Perspectives*, 23(3):3–20.

Mukhopadhyay, A., Saha, D., Mahanti, A., and Chakrabarti, B. (2005). Insurance for cyber-risk: A utility model. *Decision.*

National Cyber Security Centre (2015). How cyber attacks work.

Nielsen, S. F. (2007). Proper and Improper Multiple Imputation. *International Statistical Review*, 71(3):593–607.

OECD (2018). *Manual de Frascati 2015.*

Oxford Economics (2014). Cyber-attacks: Effects on UK Companies. Technical report.

Pinzon, E. (2016). Two faces of misspecification in maximum likelihood: Heteroskedasticity and robust standard errors.

Porter, M. E. and Stern, S. (2000). Measuring the "Ideas" Production Function: Evidence from International Patent Output. *NBER Working Papers.*

Ramsey, J. B. (1969). Tests for Specification Errors in Classical Linear Least Squares Regression Analysis. *Journal of the Royal Statistical Society, Series B*, 31(2):350–371.

Ransbotham, S. (2010). An Empirical Analysis of Exploitation Attempts Based on Vulnerabilities in Open Source Software. *Weis.*

Rodwell, L., Lee, K. J., Romaniuk, H., and Carlin, J. B. (2014). Comparison of methods for imputing limited-range variables: A simulation study. *BMC Medical Research Methodology*, 14(1):57.

Rubin, D. B. (1987). *Integrative analysis of high-throughput cancer studies with contrasted penalization.* John Wiley & Sons.

Schafer, J. L. and Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, 7(2):147–177.

Stata Corporation (2013). Introduction to multilevel mixed-effects models. Technical report.

Steger, J. A., Manners, G., Bernstein, A. J., and May, R. (1975). The Three Dimensions of the. Technical Report 3.

Sterne, J. A., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., Wood, A. M., and Carpenter, J. R. (2009). Multiple imputation for missing data in epidemiological and clinical research: Potential and pitfalls. *BMJ*, 339(7713):157–160.

Terleckyj, N. (1960). Sources of productivity advance : A pilot study of manufacturing industries, 1899-1953. *undefined*.

Van Buuren, S., Boshuizen, H. C., and Knook, D. L. (1999). Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine*, 18(6):681–694.

Verizon Business (2011). 2011 Data Breach Investigations Report (DBIR). *Trends*.

Verizon Business (2020). 2020 The Veris Community Database (VCDB).

Villadsen, A. R. and Wulff, J. N. (2019). Are you 110% sure? Modeling of fractions and proportions in strategy and management research. *Strategic Organization*, page 147612701985496.

von Hippel, P. T. (2013). Should a Normal Imputation Model be Modified to Impute Skewed Variables? *Sociological Methods & Research*, 42(1):105–138.

von Hippel, P. T. (2020). How Many Imputations Do You Need? A Two-stage Calculation Using a Quadratic Rule. *Sociological Methods & Research*, 49(3):699–718.

Wang, Q. H. and Kim, S. H. (2009). Cyber attacks: Does physical boundary matter? In *ICIS 2009 Proceedings - Thirtieth International Conference on Information Systems*.

Wedderburn, R. W. (1974). Quasi-likelihood functions, generalized linear models, and the gauss-newton method. *Biometrika*, 61(3):439–447.

White, H. (1994). *Estimation, Inference and Specification Analysis*. Cambridge University Press.

Williams, R. (2019). Analyzing Proportions: Fractional Response and Zero One Inflated Beta Models.

Wooldridge, J. M. (2010). Econometric Analysis of Cross Section and Panel Data. *MIT Press Books*, 1.

A Appendix

A.1 List of industries' aggregates

Table 12: Industries according to NACE rev.2

NACE	**Label**
C13-C15	Manufacture of textile, leather and clothes
C16-C18	Manufacture of wood and products of wood and cork, except furniture; articles of straw and plaiting materials
C19-C23	Manufacture of coke, refined petroleum, chemical and basic pharmaceutical products, rubber and plastics, other non-metallic mineral products
C24-C25	Manufacture of basic metals and fabricated metal products excluding machines and equipment
C26	Manufacture of computer, electronic and optical products
C27-C28	Manufacture of electrical equipment, machinery and equipment
C29-C30	Manufacture of motor vehicles, trailers and semi-trailers, other transport equipment
C31-C33	Manufacture of furniture and other manufacturing; repair and installatio n of machinery and equipment
D35-E39	Electricity, gas, steam, air conditioning and water supply
F41-F43	Construction and civil engineering
G45-G47	Trade of motor vehicles and motorcycles
H49-H53	Transport
I55-I56	Accommodation and Food and beverage service activities
J58-J63	Publishing of books, software and media
K64	Monetary intermediation

Industries according to NACE rev.2 (continued)

K65	Insurance, reinsurance and pension funding, except compulsory social security
K66	Activities auxiliary to financial services and insurance activities
L68	Real estate activities
M69-M74	Professional, scientific and technical activities
N77-N82	Activities for rental and leasing, employment, security and investigation, services to buildings and landscape, office administrative
S951	Repair of computers and communication equipment

A.2 Independent variables histograms

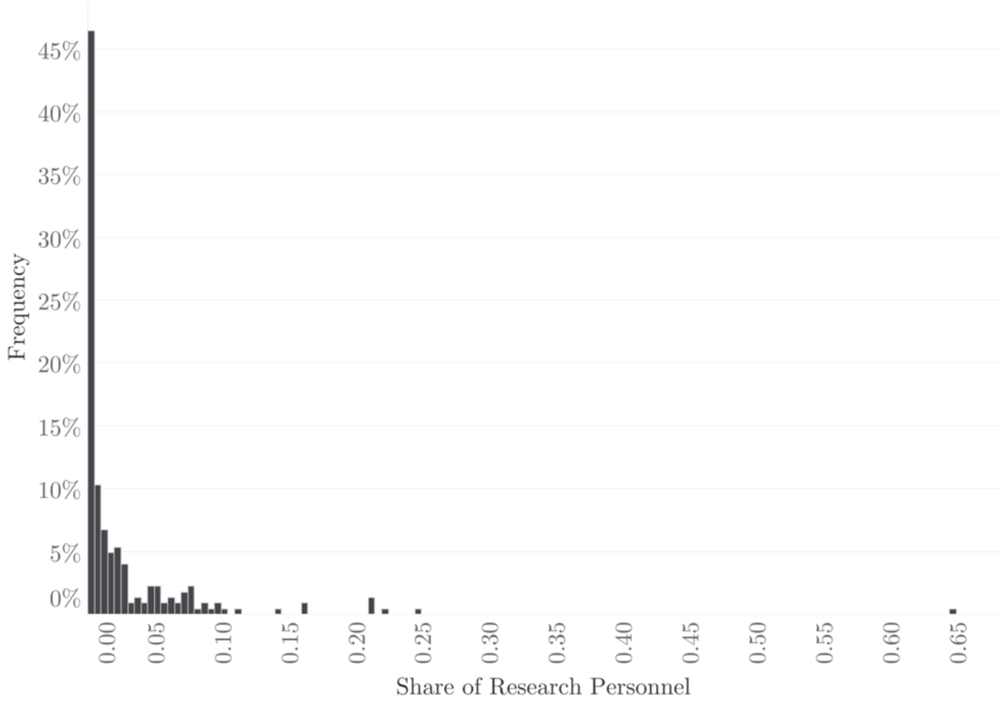

Figure 7: Share of research personnel histogram

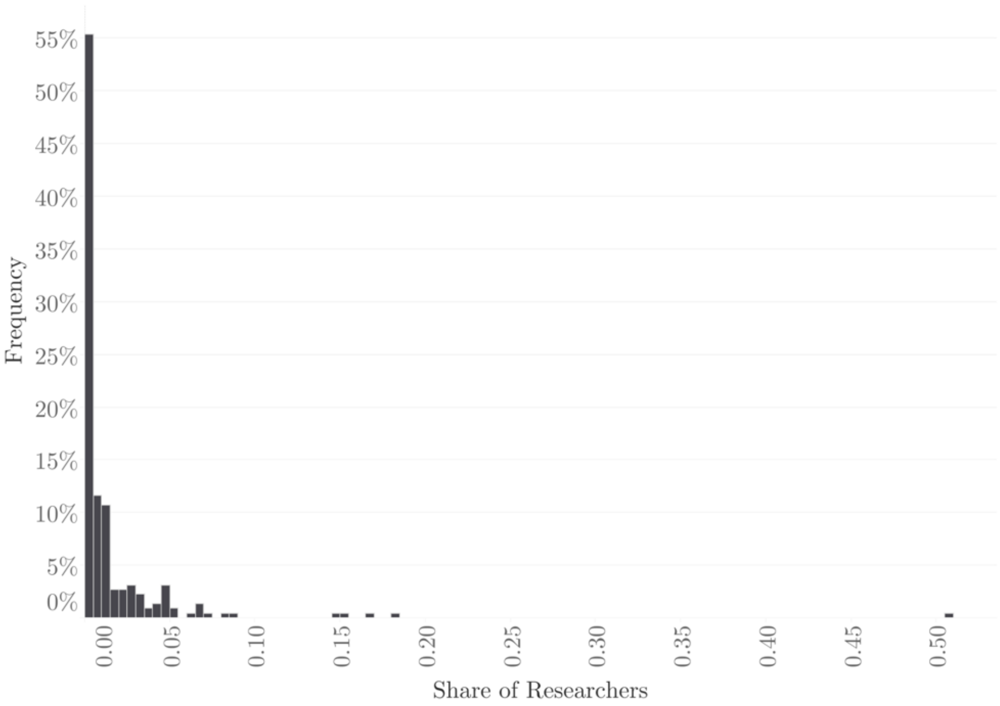

Figure 8: Share of researchers histogram

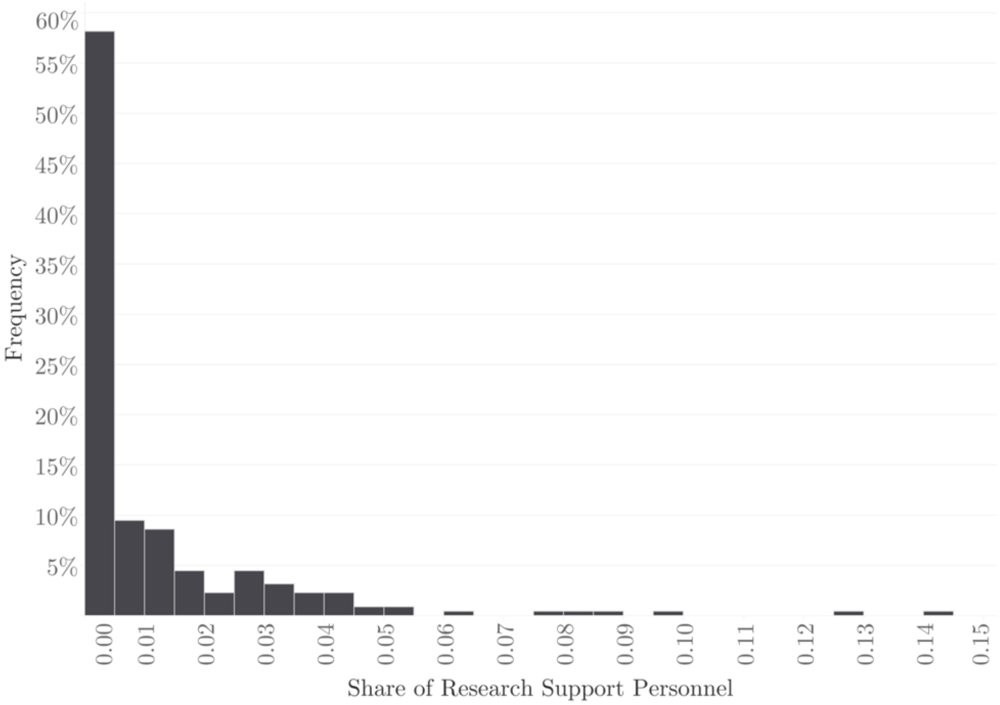

Figure 9: Share of R&D support personnel histogram

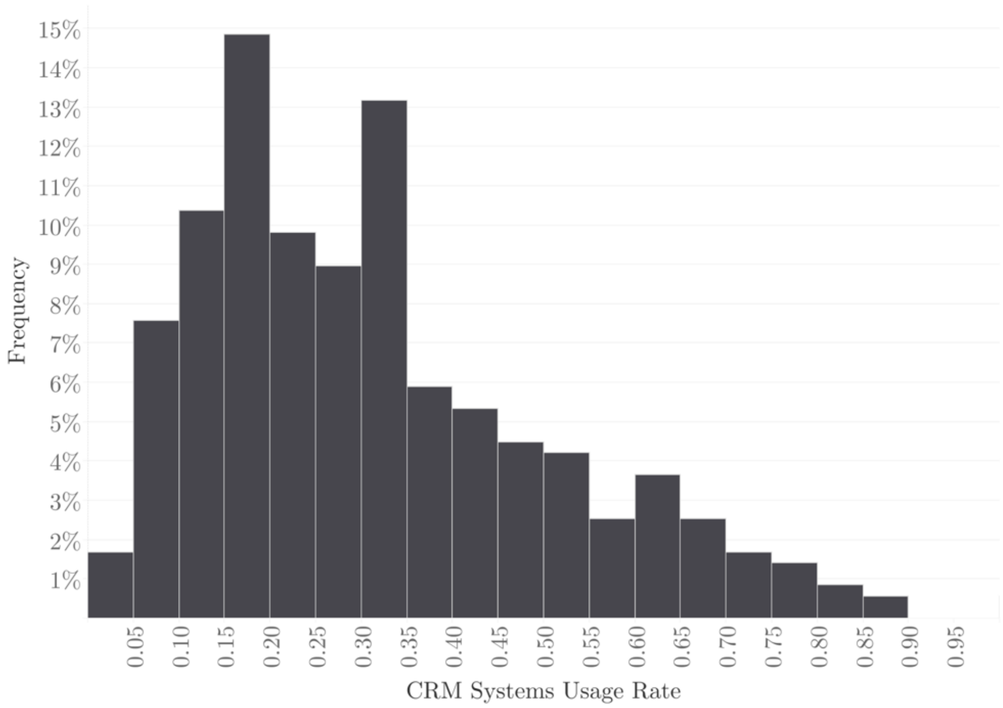

Figure 10: CRM usage rate histogram

A.3 Multiple Imputation conditions

An appropriate imputation model is the model which preserves all relationships inside the data that are relevant to the analysis. It means that imputation models should include all variables (including the dependent variable in question) utilised in the analytical models, which is ensured in this research by design. The appropriate imputation model is required for a proper MI procedure.

A Proper MI procedure is defined as the one that yields consistent asymptotically normal estimators and approximately unbiased estimator of its asymptomatic variance in common regular models (Rubin's rule) (Nielsen, 2007). In practice, it means that if the multiple imputations are proper, then complete data estimates can be obtained by a simple averaging of within imputation estimates. Their variances are then obtained by combining the within- and between-imputation variances.

Missing-data mechanism should be ignorable in order for the MI technique to produce unbiased results. That means that the pattern of missing observations must not depend on unobserved data. This condition is satisfied when: (1) data are missing completely at random (MCAR) and depend neither on observed nor unobserved data, or (2) data are missing at random (MAR), which allows the missing data pattern to be dependent on the observed data.

MCAR assumption is seldom satisfied in real life. Moreover, there is no formal statistical procedure to test the MAR assumption. Therefore, the ignorability of the missing pattern is generally assumed based on the nature of the dataset and the available knowledge about it (Liu, 2015; Little and Rubin, 2002). Given that observations are likely missing due to industry-specific reasons, adopted control variables can be used to model the missingness mechanism. It then enables the research to assume that data is MAR Bland (2006).

Additionally, it was shown by Schafer and Graham (2002); Collins et al. (2001) that model-based imputation methods (e.g. maximum likelihood or multiple imputation) perform well and do not create substantial bias even when the data are not MAR.

The sufficient quantity of completed datasets is another crucial point to determine to achieve stable unbiased results. Following the recommendations by Graham et al. (2007), the research chooses 40 imputations, which correspond to up to 50% of missing data. It is more than what was initially recommended by Rubin but should allow for more consistent point estimates and standard errors (von Hippel, 2020).

A.4 Assumptions for GLM class models

1. Linearity assumption is relaxed, meaning that an outcome variable does not need to have a linear relationship with regressors. The outcome variable is assumed to be generated from a particular distribution in an exponential family,

2. The fractional outcome model assumes a correct conditional mean specification and is agnostic about other moments of the outcome. It utilises robust estimators of a variance-covariance matrix to achieve consistent estimates of the unknown standard errors, which implies the that homoscedasticity assumption is relaxed (White, 1994; Pinzon, 2016),

3. Normality of error terms is not necessary.

A.5 Eurostat indicators on high-tech industry and knowledge-intensive services

Manufacturing industires	**NACE Rev. 2**
High-technology	C21, C26
Medium-high-technology	C20, C27-C30
Medium-low-technology	C19, C22-C25, C33
Low-technology	C10-C18, C31-C32

Table 13: Aggregations of manufacturing based on NACE Rev. 2

Knowledge based services	**NACE Rev. 2**
Knowledge-intensive services (KIS)	J58-I63, K64-K66, L68, M69-M74
Less knowledge-intensive services (LKIS)	D35-E39, F41-F43, G45-G47, H49-H53, I55-I56, N77-N82, S951

Table 14: Aggregations of services based on NACE Rev. 2

A.6 Collinearity in multivariate specifications

Tables 15 and 16 provide variance inflation factors for the main independent variables in multivariate specification with and without industry controls correspondingly. Sufficiently high VIF (> 4) is conventionally considered as a sign of collinearity. The removal of industry controls mitigates the problem. Table 17 demonstrates that within industry variance of CRM usage is sufficiently small.

Variable	VIF	1/VIF
r_{ij}^L	2.15	0.464536
D_{ij}	**5.20**	0.192332

Table 15: Variance Inflation Factor for the main independent variables in multivariate specification with industry controls

Variable	VIF	1/VIF
r_{ij}^L	1.39	0.719559
D_{ij}	**1.91**	0.522411

Table 16: Variance Inflation Factor for the main independent variables in multivariate specification without industry controls

Industry NACE r2 code	Variance
C10-C12	0.003526
C13-C15	0.012441
C16-C18	0.01804
C19-C23	0.014848
C24-C25	0.010181
C26	0.058687
C27-C28	0.019917
C29-C30	0.008458
C31-C33	0.011255
D35-E39	0.014388
F41-F43	0.004358
G45-G47	0.009418
H49-H53	0.006454
I55-I56	0.006485
J58-J63	0.010973
K64	0.03126
K65	0.021616
K66	0.054944
M69-M74	0.013961
N77-N82	0.011742
S951	0.050869

Table 17: Within industry variance of CRM system usage

A.7 RESET

H0: all non-linear combinations coefficients are zero - the model is correctly specified

1. RESET results for univariate specification with probit link function

Maximum polynomial power: 2

Statistic 1.072

P-value $0.3005 > 0.05$

Result: fail to reject H0

2. RESET results for univariate specification with logit link function

Maximum polynomial power: 2

Statistic 1.353

P-value $0.2448 > 0.05$

Result: fail to reject H0

3. RESET results for base multivariate specification with probit link function

Maximum polynomial power: 2

Statistic 1.315

P-value $0.2516 > 0.05$

Result: fail to reject H0

4. RESET results for base multivariate specification with logit link function

Maximum polynomial power: 2

Statistic 0.724

P-value $0.3948 > 0.05$

Result: fail to reject H0

A.8 Alternative specifications test

Univariate fractional probit vs. logit models

H0: Univariate fractional probit model

H1: Univariate fractional logit model

Version: t-test

Statistic: -0.190

P-value: 0.8496

Result: fail to reject H0

Multivariate fractional probit vs. logit models

H0: Multivariate fractional probit model

H1: Multivariate fractional logit model

Version: t-test

Statistic: 0.915

P-value: 0.3614

Result: fail to reject H0

A.9 Kernel density plots of residuals of imputation models

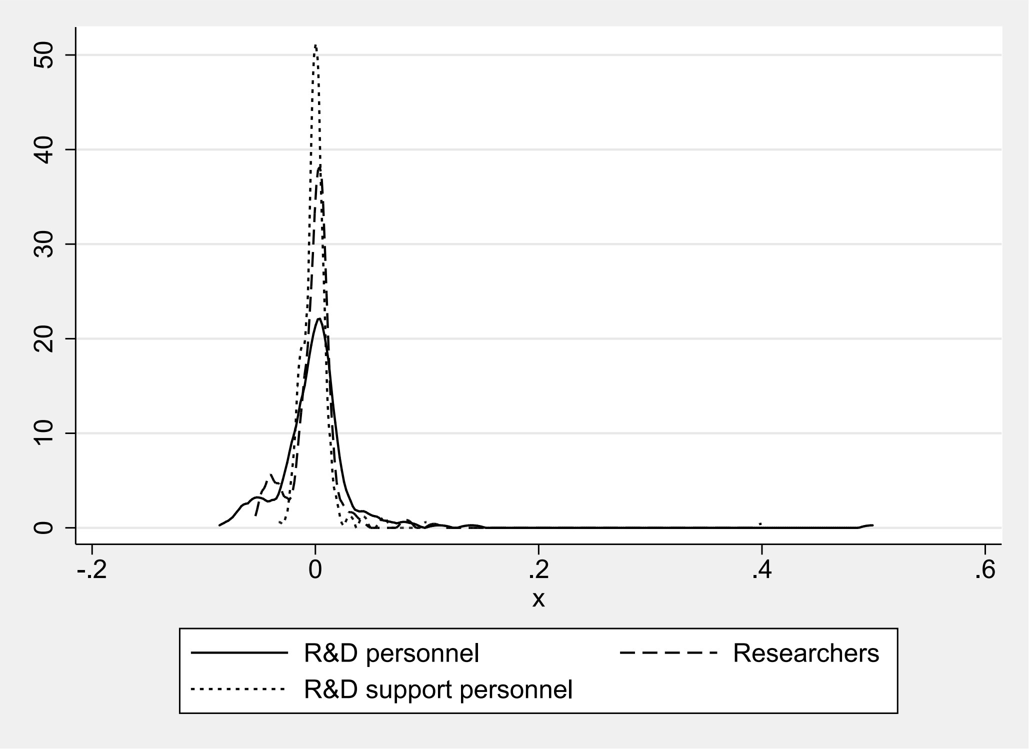

Figure 11: Residuals distributions for univariate imputation models

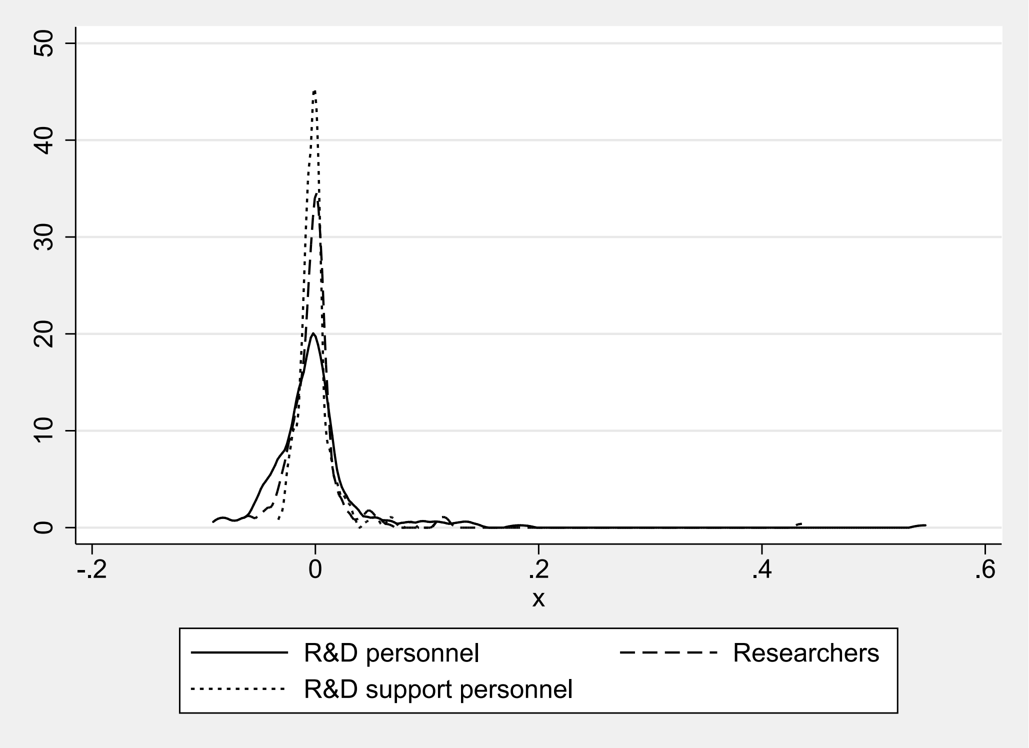

Figure 12: Residuals distributions for multivariate imputation models

A.10 Complete case regression analysis results for r^{per}

Table 18: Regression results: fractional probit regressions on complete data with share of research personnel as an independent variable

	I	**II**	**III**	**IV**
	Base	Country	Industry	All
	model	controls	controls	controls
r^{per}	0.965**	0.612**	1.217**	1.036***
Complete	(2.08)	(1.96)	(2.04)	(3.11)
BG		-0.730***		-0.760***
		(-3.01)		(-2.96)
CY		-0.820***		-0.847***
		(-3.81)		(-3.90)
CZ		-0.257*		-0.295**
		(-1.69)		(-2.01)
DK		-0.341**		-0.417**
		(-2.17)		(-2.57)
ES		-0.209		-0.240
		(-1.43)		(-1.50)
FI		-0.373**		-0.441**
		(-2.04)		(-2.37)
FR		-0.111		-0.182
		(-0.62)		(-1.15)
HR		-0.454**		-0.486**
		(-2.07)		(-2.52)
HU		-0.865***		-0.822***
		(-2.73)		(-2.78)
IE		-0.414**		-0.430**
		(-2.34)		(-2.42)
IT		-0.196		-0.207
		(-1.17)		(-1.25)
LT		0.182		0.168

Table 18: Regression results: fractional probit regressions on complete data with share of research personnel as an independent variable (continued)

	(1.20)		(1.11)
LV	-0.199		-0.198
	(-1.06)		(-1.11)
MT	-0.0241		-0.0703
	(-0.12)		(-0.39)
NL	0.268^{*}		0.244
	(1.81)		(1.61)
NO	-0.452^{**}		-0.502^{***}
	(-2.44)		(-2.70)
PT	-0.269^{*}		-0.338^{**}
	(-1.89)		(-1.98)
RO	0.0236		0.0289
	(0.15)		(0.18)
SI	-1.193^{***}		-1.213^{***}
	(-3.84)		(-3.87)
SK	0.325^{**}		0.285^{*}
	(2.01)		(1.82)
C13-C15		-0.240	-0.112
		(-1.31)	(-1.03)
C16-C18		-0.165	-0.0419
		(-1.10)	(-0.41)
C19-C23		-0.177	-0.0981
		(-0.88)	(-0.62)
C24-C25		0.0384	0.140
		(0.27)	(1.39)
C26		-0.251	-0.187
		(-1.30)	(-1.60)
C27-C28		-0.124	-0.0568
		(-0.71)	(-0.48)
C29-C30		-0.0138	0.0656

Table 18: Regression results: fractional probit regressions on complete data with share of research personnel as an independent variable (continued)

			(-0.09)	(0.64)
C31-C33			-0.106	-0.0426
			(-0.59)	(-0.41)
D35-E39			-0.121	0.0563
			(-0.81)	(0.66)
F41-F43			-0.0576	-0.00865
			(-0.46)	(-0.11)
G45-G47			-0.0333	0.0909
			(-0.29)	(1.47)
H49-H53			-0.0699	0.0448
			(-0.52)	(0.57)
I55-I56			-0.294**	0.00450
			(-2.49)	(0.03)
J58-J63			0.106	0.247***
			(0.72)	(2.97)
L68			0.0217	0.157
			(0.12)	(1.36)
M69-M74			-0.190	-0.0591
			(-1.11)	(-0.59)
N77-N82			-0.00764	0.0610
			(-0.06)	(0.90)
S951			-0.518	-0.110
			(-1.64)	(-0.42)
Constant	-2.214***	-2.080***	-2.139***	-2.093***
	(-71.13)	(-14.90)	(-23.27)	(-13.96)
Observations	224	224	224	224

z statistics in parentheses; Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

C10-C12 and BE are chosen as base levels

A.11 Multiple imputation regression analysis results for r^{per}

Table 19: Regression results: fractional probit regressions on imputed data with share of research personnel as an independent variable

	I Base model	**II** Country controls	**III** Industry controls	**IV** All controls
r^{per}	1.377**	1.011*	1.820**	1.285**
	(2.23)	(1.95)	(2.44)	(2.43)
BG		-0.651***		-0.716***
		(-3.01)		(-3.11)
CY		-0.660***		-0.749***
		(-3.13)		(-3.82)
CZ		-0.243		-0.340**
		(-1.58)		(-2.15)
DE		-3.371***		-3.568***
		(-14.00)		(-11.93)
DK		-0.341**		-0.445***
		(-2.16)		(-2.69)
EL		-0.621*		-0.648*
		(-1.81)		(-1.92)
ES		-0.131		-0.279*
		(-0.83)		(-1.77)
FI		-0.409**		-0.484**
		(-2.17)		(-2.48)
FR		-0.0809		-0.200
		(-0.46)		(-1.27)
HR		-0.493**		-0.554***
		(-2.18)		(-2.71)
HU		-0.750***		-0.839***
		(-3.04)		(-3.68)
IE		-0.404**		-0.468***

Table 19: Regression results: fractional probit regressions on imputed data with share of research personnel as an independent variable (continued)

	(-2.35)	(-2.59)
IS	-0.564**	-0.658***
	(-2.14)	(-2.85)
IT	-0.116	-0.193
	(-0.68)	(-1.18)
LT	0.262	0.201
	(1.52)	(1.21)
LU	-0.509**	-0.595***
	(-2.25)	(-2.86)
LV	-0.157	-0.221
	(-0.84)	(-1.27)
MK	-3.379***	-3.471***
	(-21.20)	(-19.04)
MT	0.0411	-0.0583
	(0.23)	(-0.34)
NL	0.292*	0.228
	(1.96)	(1.47)
NO	-0.350*	-0.451**
	(-1.74)	(-2.33)
PL	-0.241	-0.693***
	(-1.55)	(-3.68)
PT	-0.265*	-0.304*
	(-1.83)	(-1.78)
RO	-0.0278	-0.0566
	(-0.17)	(-0.33)
SE	-0.491***	-0.570***
	(-2.72)	(-3.02)
SI	-1.198***	-1.267***
	(-3.85)	(-3.95)
SK	0.395**	0.311**

Table 19: Regression results: fractional probit regressions on imputed data with share of research personnel as an independent variable (continued)

	(2.39)		(2.00)
TR	-0.0219		-0.0801
	(-0.14)		(-0.49)
UK	-3.354***		-3.545***
	(-14.01)		(-12.12)
C13-C15		-0.0825	-0.0919
		(-0.54)	(-0.98)
C16-C18		-0.0953	-0.0436
		(-0.68)	(-0.43)
C19-C23		-0.188	-0.127
		(-0.97)	(-0.85)
C24-C25		0.0639	0.136
		(0.46)	(1.43)
C26		-0.313	-0.210*
		(-1.60)	(-1.71)
C27-C28		-0.00476	0.0470
		(-0.03)	(0.39)
C29-C30		-0.0531	0.0384
		(-0.33)	(0.39)
C31-C33		-0.106	-0.0755
		(-0.61)	(-0.77)
D35-E39		-0.0621	0.0105
		(-0.46)	(0.13)
F41-F43		0.0210	0.0340
		(0.18)	(0.42)
G45-G47		0.0552	0.108*
		(0.49)	(1.74)
H49-H53		0.0633	0.0713
		(0.50)	(0.90)
I55-I56		-0.0986	0.0119

Table 19: Regression results: fractional probit regressions on imputed data with share of research personnel as an independent variable (continued)

			(-0.77)	(0.13)
J58-J63			0.166	0.284***
			(1.24)	(3.69)
K			0.350**	0.460***
			(2.32)	(4.42)
L68			0.0239	0.0981
			(0.14)	(0.85)
M69-M74			-0.186	-0.0305
			(-1.01)	(-0.26)
N77-N82			0.0295	0.0746
			(0.23)	(1.21)
S951			0.105	0.197
			(0.44)	(1.12)
Constant	-2.233***	-2.100***	-2.252***	-2.110***
	(-67.86)	(-14.56)	(-24.76)	(-13.69)
Observations	351	351	351	351

z statistics in parentheses; Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

C10-C12 and BE are chosen as base levels

A.12 Complete case regression analysis results for r^{res}

Table 20: Regression results: fractional probit regressions on complete data with share of researchers as an independent variable

	I	**II**	**III**	**IV**
	Base	Country	Industry	All
	model	controls	controls	controls
r^{res}	0.990*	0.669	1.106*	1.013***
Complete	(1.89)	(1.50)	(1.82)	(2.66)
BG		-0.676***		-0.706***
		(-3.17)		(-3.21)
CY		-0.827***		-0.848***
		(-3.81)		(-3.88)
CZ		-0.258*		-0.287*
		(-1.68)		(-1.94)
DK		-0.339**		-0.416**
		(-2.14)		(-2.57)
ES		-0.212		-0.238
		(-1.43)		(-1.48)
FI		-0.373**		-0.443**
		(-2.03)		(-2.40)
FR		-0.110		-0.182
		(-0.61)		(-1.14)
HR		-0.458**		-0.495**
		(-2.07)		(-2.52)
HU		-0.870***		-0.842***
		(-2.74)		(-2.82)
IE		-0.420**		-0.443**
		(-2.35)		(-2.47)
IT		-0.194		-0.205
		(-1.14)		(-1.21)
LT		0.177		0.167

Table 20: Regression results: fractional probit regressions on complete data with share of researchers as an independent variable (continued)

	(1.16)		(1.09)
LV	-0.290		-0.279
	(-1.50)		(-1.48)
MT	-0.0259		-0.0708
	(-0.13)		(-0.39)
NL	0.272*		0.261*
	(1.82)		(1.69)
NO	-0.443**		-0.481***
	(-2.40)		(-2.62)
PT	-0.264*		-0.330*
	(-1.84)		(-1.92)
RO	0.0167		0.00790
	(0.11)		(0.05)
SI	-1.194***		-1.206***
	(-3.84)		(-3.84)
SK	0.335**		0.300*
	(2.06)		(1.91)
C13-C15		-0.236	-0.102
		(-1.29)	(-0.96)
C16-C18		-0.167	-0.0328
		(-1.12)	(-0.31)
C19-C23		-0.148	-0.0695
		(-0.73)	(-0.43)
C24-C25		0.0433	0.150
		(0.30)	(1.47)
C26		-0.192	-0.135
		(-0.96)	(-1.05)
C27-C28		-0.0897	-0.0152
		(-0.50)	(-0.13)
C29-C30		-0.0219	0.0792

Table 20: Regression results: fractional probit regressions on complete data with share of researchers as an independent variable (continued)

			(-0.14)	(0.77)
C31-C33			-0.101	-0.0330
			(-0.55)	(-0.33)
D35-E39			-0.243**	0.0864
			(-2.01)	(0.80)
F41-F43			-0.0613	-0.00823
			(-0.49)	(-0.10)
G45-G47			-0.0382	0.0895
			(-0.33)	(1.45)
H49-H53			-0.0751	0.0477
			(-0.55)	(0.60)
I55-I56			-0.300**	0.00182
			(-2.53)	(0.01)
J58-J63			0.131	0.268***
			(0.88)	(3.16)
L68			-0.0836	0.0246
			(-0.39)	(0.28)
M69-M74			-0.137	-0.0195
			(-0.86)	(-0.21)
N77-N82			-0.0125	0.0576
			(-0.09)	(0.86)
S951			-0.521*	-0.115
			(-1.65)	(-0.44)
Constant	-2.209***	-2.073***	-2.133***	-2.092***
	(-71.27)	(-14.59)	(-23.12)	(-13.79)
Observations	224	224	224	224

z statistics in parentheses; Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

C10-C12 and BE are chosen as base levels

A.13 Multiple imputation regression analysis results results for r^{res}

Table 21: Regression results: fractional probit regressions on imputed data with share of researchers as an independent variable

	I	**II**	**III**	**IV**
	Base	Country	Industry	All
	model	controls	controls	controls
r^{res}	1.366*	1.052	1.747*	1.315**
	(1.68)	(1.53)	(1.96)	(2.07)
BG		-0.664***		-0.736***
		(-3.07)		(-3.17)
CY		-0.670***		-0.759***
		(-3.15)		(-3.81)
CZ		-0.248		-0.347**
		(-1.59)		(-2.16)
DE		-3.365***		-3.579***
		(-14.01)		(-12.03)
DK		-0.338**		-0.447***
		(-2.11)		(-2.69)
EL		-0.610*		-0.648*
		(-1.76)		(-1.90)
ES		-0.136		-0.284*
		(-0.85)		(-1.77)
FI		-0.409**		-0.480**
		(-2.17)		(-2.44)
FR		-0.0773		-0.201
		(-0.44)		(-1.27)
HR		-0.501**		-0.571***
		(-2.20)		(-2.73)
HU		-0.765***		-0.871***
		(-3.09)		(-3.80)
IE		-0.417**		-0.487***
		(-2.41)		(-2.65)
IS		-0.566**		-0.663***

Table 21: Regression results: fractional probit regressions on imputed data with share of researchers as an independent variable (continued)

	(-2.15)	(-2.88)
IT	-0.115	-0.191
	(-0.66)	(-1.14)
LT	0.256	0.193
	(1.45)	(1.14)
LU	-0.506**	-0.591***
	(-2.22)	(-2.83)
LV	-0.166	-0.234
	(-0.89)	(-1.32)
MK	-3.369***	-3.479***
	(-21.23)	(-18.35)
MT	0.0409	-0.0543
	(0.22)	(-0.31)
NL	0.303**	0.238
	(2.01)	(1.51)
NO	-0.330*	-0.428**
	(-1.67)	(-2.22)
PL	-0.242	-0.708***
	(-1.60)	(-3.91)
PT	-0.258*	-0.302*
	(-1.77)	(-1.74)
RO	-0.0428	-0.0831
	(-0.26)	(-0.48)
SE	-0.489***	-0.572***
	(-2.71)	(-3.00)
SI	-1.200***	-1.274***
	(-3.85)	(-3.94)
SK	0.411**	0.332**
	(2.47)	(2.12)
TR	-0.0144	-0.0729

Table 21: Regression results: fractional probit regressions on imputed data with share of researchers as an independent variable (continued)

		(-0.09)		(-0.45)
UK		-3.342***		-3.548***
		(-13.86)		(-11.96)
C13-C15			-0.0704	-0.0807
			(-0.46)	(-0.87)
C16-C18			-0.0951	-0.0441
			(-0.69)	(-0.43)
C19-C23			-0.154	-0.106
			(-0.79)	(-0.69)
C24-C25			0.0699	0.140
			(0.51)	(1.46)
C26			-0.232	-0.160
			(-1.19)	(-1.24)
C27-C28			0.0461	0.0875
			(0.30)	(0.76)
C29-C30			-0.0285	0.0545
			(-0.18)	(0.54)
C31-C33			-0.102	-0.0756
			(-0.58)	(-0.77)
D35-E39			-0.0747	-0.00105
			(-0.56)	(-0.01)
F41-F43			0.0163	0.0267
			(0.14)	(0.32)
G45-G47			0.0460	0.0984
			(0.41)	(1.56)
H49-H53			0.0507	0.0606
			(0.41)	(0.78)
I55-I56			-0.113	0.00463
			(-0.88)	(0.05)
J58-J63			0.203	0.308***

Table 21: Regression results: fractional probit regressions on imputed data with share of researchers as an independent variable (continued)

			(1.52)	(4.00)
K			0.370**	0.480***
			(2.42)	(4.64)
L68			0.0183	0.0915
			(0.11)	(0.79)
M69-M74			-0.107	0.0100
			(-0.64)	(0.09)
N77-N82			0.0212	0.0671
			(0.17)	(1.09)
S951			0.127	0.220
			(0.53)	(1.23)
Constant	-2.212***	-2.086***	-2.239***	-2.101***
	(-70.84)	(-14.27)	(-24.96)	(-13.38)
Observations	351	351	351	351

z statistics in parentheses; Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

C10-C12 and BE are chosen as base levels

A.14 Complete case regression analysis results for r^{sup}

Table 22: Regression results: fractional probit regressions on complete data with share of R&D support personnel as an independent variable

	I	**II**	**III**	**IV**
	Base	Country	Industry	All
	model	controls	controls	controls
r^{sup}	4.584***	2.324***	6.548***	4.302***
Complete	(3.71)	(2.60)	(3.52)	(3.51)
BG		-0.726***		-0.722***
		(-3.02)		(-2.92)
CY		-0.815***		-0.827***
		(-3.85)		(-3.93)
CZ		-0.262*		-0.291**
		(-1.77)		(-2.06)
DK		-0.349**		-0.415***
		(-2.28)		(-2.63)
ES		-0.213		-0.250
		(-1.50)		(-1.61)
FI		-0.380**		-0.449**
		(-2.10)		(-2.44)
FR		-0.110		-0.169
		(-0.62)		(-1.08)
HR		-0.453**		-0.480***
		(-2.11)		(-2.59)
HU		-0.860***		-0.786***
		(-2.74)		(-2.69)
IE		-0.410**		-0.407**
		(-2.38)		(-2.39)
IT		-0.214		-0.234
		(-1.33)		(-1.49)
LT		0.186		0.193

Table 22: Regression results: fractional probit regressions on complete data with share of R&D support personnel as an independent variable (continued)

	(1.28)		(1.34)
LV	-0.289		-0.261
	(-1.52)		(-1.45)
MT	-0.0287		-0.0790
	(-0.15)		(-0.45)
NL	0.237		0.211
	(1.61)		(1.43)
NO	-0.447**		-0.482***
	(-2.53)		(-2.82)
PT	-0.289**		-0.353**
	(-2.06)		(-2.14)
RO	0.0310		0.0646
	(0.21)		(0.42)
SI	-1.202***		-1.213***
	(-3.87)		(-3.91)
SK	0.300*		0.249*
	(1.90)		(1.65)
C13-C15		-0.260	-0.126
		(-1.42)	(-1.16)
C16-C18		-0.160	-0.0417
		(-1.06)	(-0.42)
C19-C23		-0.271	-0.142
		(-1.37)	(-0.96)
C24-C25		0.0207	0.132
		(0.15)	(1.34)
C26		-0.419**	-0.270**
		(-2.21)	(-2.43)
C27-C28		-0.246	-0.138
		(-1.44)	(-1.10)
C29-C30		-0.0797	0.0344

Table 22: Regression results: fractional probit regressions on complete data with share of R&D support personnel as an independent variable (continued)

			(-0.50)	(0.34)
C31-C33			-0.124	-0.0515
			(-0.72)	(-0.52)
D35-E39			-0.235*	0.0386
			(-1.89)	(0.34)
F41-F43			-0.0463	0.00867
			(-0.37)	(0.11)
G45-G47			-0.0172	0.105*
			(-0.15)	(1.74)
H49-H53			-0.0537	0.0587
			(-0.40)	(0.73)
I55-I56			-0.275**	0.0131
			(-2.33)	(0.08)
J58-J63			0.0358	0.205**
			(0.26)	(2.48)
L68			-0.0601	0.0415
			(-0.28)	(0.57)
M69-M74			-0.267	-0.0868
			(-1.48)	(-0.82)
N77-N82			0.00775	0.0680
			(0.06)	(1.01)
S951			-0.508	-0.113
			(-1.63)	(-0.43)
Constant	-2.254***	-2.087***	-2.158***	-2.100***
	(-70.11)	(-15.60)	(-23.55)	(-14.72)
Observations	222	222	222	222

z statistics in parentheses; Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

C10-C12 and BE are chosen as base levels

A.15 Multiple imputation regression analysis results results for r^{sup}

Table 23: Regression results: fractional probit regressions on imputed data with share of R&D support personnel as an independent variable

	I	**II**	**III**	**IV**
	Base	Country	Industry	All
	model	controls	controls	controls
r^{sup}	6.225***	3.838***	9.106***	5.521***
	(3.81)	(2.80)	(4.38)	(3.67)
BG		-0.642***		-0.684***
		(-3.05)		(-3.06)
CY		-0.650***		-0.727***
		(-3.18)		(-3.81)
CZ		-0.256*		-0.353**
		(-1.77)		(-2.32)
DE		-3.311***		-3.481***
		(-13.78)		(-12.05)
DK		-0.355**		-0.446***
		(-2.37)		(-2.77)
EL		-0.629*		-0.633*
		(-1.89)		(-1.89)
ES		-0.142		-0.294*
		(-0.96)		(-1.93)
FI		-0.420**		-0.498***
		(-2.29)		(-2.59)
FR		-0.0781		-0.187
		(-0.47)		(-1.22)
HR		-0.497**		-0.542***
		(-2.27)		(-2.79)
HU		-0.737***		-0.788***
		(-3.06)		(-3.53)
IE		-0.386**		-0.429**

Table 23: Regression results: fractional probit regressions on imputed data with share of R&D support personnel as an independent variable (continued)

	(-2.37)	(-2.49)
IS	-0.566**	-0.653***
	(-2.20)	(-2.88)
IT	-0.148	-0.229
	(-0.93)	(-1.48)
LT	0.268*	0.220
	(1.65)	(1.40)
LU	-0.494**	-0.573***
	(-2.23)	(-2.77)
LV	-0.157	-0.208
	(-0.88)	(-1.24)
MK	-3.370***	-3.444***
	(-22.27)	(-19.04)
MT	0.0312	-0.0685
	(0.18)	(-0.41)
NL	0.244*	0.176
	(1.69)	(1.18)
NO	-0.338*	-0.434**
	(-1.84)	(-2.44)
PL	-0.246*	-0.673***
	(-1.68)	(-3.64)
PT	-0.299**	-0.323**
	(-2.19)	(-1.96)
RO	-0.0139	-0.0135
	(-0.09)	(-0.08)
SE	-0.477***	-0.540***
	(-2.72)	(-2.92)
SI	-1.216***	-1.274***
	(-3.91)	(-4.02)
SK	0.358**	0.258*

Table 23: Regression results: fractional probit regressions on imputed data with share of R&D support personnel as an independent variable (continued)

	(2.30)		(1.70)
TR	-0.0195		-0.0673
	(-0.13)		(-0.43)
UK	-3.318***		-3.492***
	(-13.74)		(-12.00)
C13-C15		-0.112	-0.106
		(-0.71)	(-1.10)
C16-C18		-0.0898	-0.0458
		(-0.62)	(-0.47)
C19-C23		-0.307	-0.190
		(-1.57)	(-1.32)
C24-C25		0.0496	0.122
		(0.35)	(1.29)
C26		-0.531***	-0.324***
		(-2.66)	(-2.71)
C27-C28		-0.174	-0.0587
		(-1.05)	(-0.44)
C29-C30		-0.130	-0.00396
		(-0.77)	(-0.04)
C31-C33		-0.127	-0.0917
		(-0.73)	(-0.91)
D35-E39		-0.0467	0.0148
		(-0.34)	(0.17)
F41-F43		0.0544	0.0552
		(0.44)	(0.67)
G45-G47		0.0804	0.122*
		(0.68)	(1.93)
H49-H53		0.0821	0.0843
		(0.62)	(1.01)
I55-I56		-0.0645	0.0279

Table 23: Regression results: fractional probit regressions on imputed data with share of R&D support personnel as an independent variable (continued)

			(-0.46)	(0.28)
J58-J63			0.0846	0.234***
			(0.63)	(2.96)
K			0.285**	0.423***
			(1.97)	(4.15)
L68			0.0341	0.103
			(0.20)	(0.89)
M69-M74			-0.270	-0.0700
			(-1.40)	(-0.59)
N77-N82			0.0593	0.0871
			(0.45)	(1.34)
S951			0.102	0.184
			(0.44)	(1.13)
Constant	-2.273***	-2.110***	-2.279***	-2.118***
	(-68.83)	(-15.78)	(-23.31)	(-14.35)
Observations	351	351	351	351

z statistics in parentheses; Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

C10-C12 and BE are chosen as base levels

A.16 Complete case regression analysis results for models with research intensity proxies and CRM variable

Table 24: Regression results: fractional probit regressions on complete data with research poxies and CRM variable

	I	**II**	**III**
	R&D personnel model	Researchers model	R&D support model
r^{per}	0.332		
	(0.94)		
r^{res}		0.225	
		(0.37)	
r^{sup}			1.606*
			(1.71)
d	0.487**	0.524**	0.440**
	(2.29)	(2.48)	(1.98)
BG	-0.576**	-0.522**	-0.582**
	(-2.25)	(-2.32)	(-2.29)
CY	-0.696***	-0.697***	-0.700***
	(-3.10)	(-3.10)	(-3.14)
CZ	-0.144	-0.139	-0.158
	(-0.83)	(-0.81)	(-0.92)
DK	-0.285*	-0.281	-0.296*
	(-1.65)	(-1.63)	(-1.74)
ES	-0.139	-0.140	-0.147
	(-0.87)	(-0.87)	(-0.93)
FI	-0.354*	-0.354*	-0.360*
	(-1.85)	(-1.86)	(-1.89)
FR	-0.0176	-0.00982	-0.0260
	(-0.09)	(-0.05)	(-0.14)
HR	-0.346	-0.344	-0.354

Table 24: Regression results: fractional probit regressions on complete data with research poxies and CRM variable (continued)

	(-1.55)	(-1.55)	(-1.60)
HU	-0.692**	-0.684**	-0.704**
	(-2.12)	(-2.10)	(-2.16)
IE	-0.368*	-0.372**	-0.367**
	(-1.95)	(-1.97)	(-1.98)
IT	-0.0950	-0.0905	-0.116
	(-0.53)	(-0.50)	(-0.65)
LT	0.315*	0.319*	0.305*
	(1.86)	(1.89)	(1.82)
LV	-0.141	-0.136	-0.155
	(-0.68)	(-0.66)	(-0.75)
MT	0.0759	0.0787	0.0649
	(0.38)	(0.39)	(0.33)
NL	0.372**	0.381**	0.340**
	(2.30)	(2.36)	(2.07)
NO	-0.427**	-0.416**	-0.432**
	(-2.15)	(-2.11)	(-2.24)
RO	0.128	0.128	0.126
	(0.76)	(0.75)	(0.76)
SI	-1.023***	-1.016***	-1.043***
	(-3.19)	(-3.18)	(-3.24)
SK	0.404**	0.415**	0.378**
	(2.31)	(2.39)	(2.18)
Constant	-2.293***	-2.301***	-2.280***
	(-13.69)	(-13.80)	(-13.65)
Observations	214	215	213

z statistics in parentheses; Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

C10-C12 is chosen as a base level

A.17 Multiple imputation regression analysis results results for models with research intensity proxies and CRM variable

Table 25: Regression results: fractional probit regressions on imputed data with research poxies and CRM variable

	I R&D personnel model	II Researchers model	III R&D support model
r^{per}	0.562		
	(1.22)		
r^{res}		0.410	
		(0.53)	
r^{sup}			3.041**
			(2.35)
d	0.715***	0.762***	0.641***
	(3.63)	(3.78)	(3.48)
BG	-0.485**	-0.485**	-0.487**
	(-2.15)	(-2.16)	(-2.20)
CY	-0.545***	-0.550***	-0.543***
	(-2.70)	(-2.72)	(-2.74)
CZ	-0.104	-0.103	-0.122
	(-0.63)	(-0.62)	(-0.77)
DE	-3.436***	-3.450***	-3.417***
	(-13.59)	(-13.70)	(-13.06)
DK	-0.265	-0.258	-0.285*
	(-1.55)	(-1.52)	(-1.74)
EL	-0.615*	-0.625*	-0.604*
	(-1.66)	(-1.67)	(-1.69)
ES	-0.0930	-0.0939	-0.0949
	(-0.61)	(-0.61)	(-0.64)
FI	-0.380**	-0.381**	-0.391**

Table 25: Regression results: fractional probit regressions on imputed data with research poxies and CRM variable (continued)

	(-2.02)	(-2.04)	(-2.11)
FR	0.0259	0.0346	0.0159
	(0.15)	(0.20)	(0.10)
HR	-0.372	-0.368	-0.381*
	(-1.60)	(-1.60)	(-1.68)
HU	-0.528**	-0.527**	-0.532**
	(-2.14)	(-2.14)	(-2.20)
IE	-0.333*	-0.342*	-0.319*
	(-1.85)	(-1.90)	(-1.85)
IS	-0.519**	-0.523**	-0.521**
	(-1.99)	(-2.01)	(-2.03)
IT	-0.00988	-0.00298	-0.0410
	(-0.06)	(-0.02)	(-0.25)
LT	0.453**	0.456**	0.442**
	(2.42)	(2.43)	(2.51)
LU	-0.426*	-0.428*	-0.423*
	(-1.91)	(-1.92)	(-1.93)
LV	0.0182	0.0198	0.00470
	(0.09)	(0.10)	(0.03)
MK	-3.218***	-3.216***	-3.229***
	(-18.38)	(-18.34)	(-19.55)
MT	0.170	0.173	0.157
	(0.95)	(0.96)	(0.90)
NL	0.431***	0.442***	0.379**
	(2.77)	(2.87)	(2.49)
NO	-0.315	-0.300	-0.327*
	(-1.58)	(-1.52)	(-1.74)
PL	-0.305*	-0.313**	-0.272*
	(-1.95)	(-2.00)	(-1.80)
PT	-0.274*	-0.272*	-0.285*

Table 25: Regression results: fractional probit regressions on imputed data with research poxies and CRM variable (continued)

	(-1.69)	(-1.66)	(-1.83)
RO	0.120	0.115	0.129
	(0.69)	(0.66)	(0.78)
SE	-0.457**	-0.461**	-0.451**
	(-2.37)	(-2.39)	(-2.40)
SI	-1.038***	-1.033***	-1.064***
	(-3.21)	(-3.19)	(-3.31)
SK	0.524***	0.538***	0.467***
	(3.06)	(3.17)	(2.87)
TR	0.0441	0.0433	0.0375
	(0.27)	(0.27)	(0.24)
UK	-3.444***	-3.459***	-3.411***
	(-13.99)	(-14.10)	(-14.18)
Constant	-2.409***	-2.416***	-2.395***
	(-14.62)	(-14.64)	(-15.37)
Observations	351	351	351

z statistics in parentheses; Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

C10-C12 is chosen as a base level

A.18 Multiple imputation regression analysis results results for models with research intensity proxies and CRM variable for manufacturing industries

Table 26: Regression results: fractional probit regressions on imputed data with research poxies and CRM variable for manufacturing industries

	I	**II**	**III**
	All manufacturing	Low-tech manufacturing	High-tech manufacturing
r^{sup}	4.488***	3.200	7.206***
	(2.77)	(0.66)	(3.13)
d	-0.776	-0.651	-1.337
	(-1.41)	(-0.87)	(-1.29)
BG	-1.222***	-0.794**	-4.560***
	(-3.50)	(-2.29)	(-10.74)
CY	-4.216***	-4.037***	-4.612***
	(-16.02)	(-13.17)	(-9.70)
CZ	-0.682***	-0.408	-1.115***
	(-2.71)	(-1.49)	(-2.77)
DK	-0.643***	-0.468***	-0.927***
	(-3.25)	(-2.78)	(-3.03)
EL	-0.821**	-3.930***	-0.893*
	(-2.07)	(-16.98)	(-1.78)
ES	-0.650***		-0.870***
	(-3.36)		(-2.83)
FI	-4.134***	-3.900***	-4.573***
	(-18.11)	(-19.28)	(-13.66)
FR	-0.643***	-0.557**	-0.875**
	(-2.82)	(-2.46)	(-2.49)
HR	-4.119***	-3.992***	-4.426***
	(-18.30)	(-13.51)	(-13.67)

Table 26: Regression results: fractional probit regressions on imputed data with research poxies and CRM variable for manufacturing industries (continued)

HU	-1.287***	-4.039***	-1.418***
	(-3.38)	(-13.27)	(-2.60)
IE	-0.771***	-0.535**	-1.168***
	(-3.31)	(-2.20)	(-3.02)
IS	-4.146***	-4.013***	-4.376***
	(-17.01)	(-13.60)	(-13.05)
IT	-0.607***	-0.624***	-0.758**
	(-2.58)	(-3.25)	(-1.98)
LT	-0.253	-0.0903	-0.550
	(-1.00)	(-0.33)	(-1.23)
LU	-4.079***	-3.962***	-4.238***
	(-16.99)	(-13.54)	(-14.04)
LV	-0.880***	-0.480	-4.656***
	(-2.81)	(-1.39)	(-10.96)
MK	-4.188***	-3.993***	-4.621***
	(-16.57)	(-12.98)	(-10.64)
MT	-0.515	-0.223	-4.995***
	(-1.46)	(-0.83)	(-8.86)
NL	-0.277	-0.0948	-0.681*
	(-1.17)	(-0.49)	(-1.67)
NO	-0.994***	-0.559*	-4.406***
	(-2.85)	(-1.86)	(-15.56)
PT	-0.524***		-0.655**
	(-2.76)		(-2.41)
RO	-0.319	-0.171	-0.532
	(-1.42)	(-0.79)	(-1.42)
SE	-1.203***	-0.902***	-4.348***
	(-3.55)	(-2.77)	(-15.02)
SI	-4.274***	-4.027***	-4.851***
	(-16.59)	(-15.59)	(-11.11)

Table 26: Regression results: fractional probit regressions on imputed data with research poxies and CRM variable for manufacturing industries (continued)

SK	-0.188	-0.0713	-0.420
	(-0.86)	(-0.47)	(-1.14)
TR	-0.292	-0.205	-0.392
	(-1.50)	(-1.40)	(-1.47)
Constant	-1.549***	-1.705***	-1.223**
	(-5.43)	(-4.98)	(-2.24)
Observations	152	81	71

z statistics in parentheses; Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

C10-C12 is chosen as a base level

A.19 Multiple imputation regression analysis results results for models with research intensity proxies and CRM variable for service industries

Table 27: Regression results: fractional probit regressions on imputed data with research poxies and CRM variable for service industries

	I All knowledge-based services	II Less knowledge-intensive services	III Knowledge-intensive services
r^{sup}	3.698*	10.70*	1.862
	(1.82)	(1.89)	(1.01)
d	0.729***	0.238	0.985***
	(3.48)	(0.43)	(4.55)
BG	-0.237	-0.401	-0.182
	(-1.09)	(-1.21)	(-0.56)
CY	-0.272*	-0.230	-0.412
	(-1.77)	(-1.19)	(-1.54)
CZ	0.0182	-0.119	0.0828
	(0.16)	(-0.55)	(0.75)
DE	-3.395***	-3.463***	
	(-10.45)	(-9.16)	
DK	-0.216	-0.285	-0.201
	(-1.49)	(-0.93)	(-1.48)
EL	-3.547***	-3.621***	-3.343***
	(-11.77)	(-9.13)	(-14.28)
ES	0.0543	0.0000345	0.0168
	(0.60)	(0.00)	(0.17)
FI	-0.170	-0.325	-0.109
	(-1.26)	(-1.41)	(-0.64)
FR	0.206*	-0.0498	0.295**
	(1.66)	(-0.32)	(2.40)

Table 27: Regression results: fractional probit regressions on imputed data with research poxies and CRM variable for service industries (continued)

HR	-0.0212	-0.175	0.0735
	(-0.15)	(-0.78)	(0.36)
HU	-0.331	-3.228***	-0.0782
	(-1.22)	(-9.32)	(-0.36)
IE	-0.157	-0.180	-0.163
	(-0.97)	(-0.95)	(-0.58)
IS	-0.128	-0.566*	0.0304
	(-0.64)	(-1.68)	(0.15)
IT	0.136	-0.0439	0.195**
	(1.35)	(-0.28)	(2.01)
LT	0.605***	0.549**	0.429***
	(3.67)	(2.03)	(3.20)
LU	-0.179	-0.279	-0.101
	(-1.06)	(-0.91)	(-0.53)
LV	0.298**	0.0346	0.527***
	(2.00)	(0.14)	(4.26)
MK	-3.232***	-3.326***	-3.442***
	(-13.50)	(-10.27)	(-18.72)
MT	0.339***	0.225	0.410***
	(3.14)	(1.36)	(3.13)
NL	0.607***	0.470**	0.640***
	(6.77)	(2.46)	(7.98)
NO	-0.0540	-0.0128	-0.0806
	(-0.35)	(-0.06)	(-0.68)
PL	-0.155		-0.146
	(-1.51)		(-1.59)
RO	0.144	-0.312	0.343**
	(0.92)	(-0.98)	(2.37)
SE	-0.150	-0.0449	-0.315*
	(-0.97)	(-0.19)	(-1.77)

Table 27: Regression results: fractional probit regressions on imputed data with research poxies and CRM variable for service industries (continued)

SI	-0.761**	-0.629*	-3.431***
	(-2.39)	(-1.84)	(-21.46)
SK	0.701***	0.499**	0.811***
	(5.90)	(2.59)	(6.04)
TR	0.181	0.0171	0.274*
	(1.51)	(0.09)	(1.70)
UK	-3.401***	-3.449***	
	(-12.05)	(-10.40)	
Constant	-2.573***	-2.345***	-2.698***
	(-23.41)	(-8.96)	(-27.33)
Observations	199	105	94

z statistics in parentheses; Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

C10-C12 is chosen as a base level

A.20 Robustness checks

Table 28: Regression results: robustness checks

	I	**II**	**III**	**IV**	**V**
	A.Surface	Turnover	E-comm	Sysadmin	All
	controls	controls	controls	controls	controls
r^{sup}	3.277***	3.235***	3.783***	2.999**	3.981**
	(2.68)	(2.72)	(3.25)	(2.01)	(2.53)
d	0.585***	0.581***	0.285	0.540***	0.310
	(3.02)	(2.96)	(1.34)	(2.88)	(1.45)
A.Surface	-18E-7				703E-9
	(-0.17)				(0.07)
Turnover		0.000120			0.0000710
		(0.32)			(0.19)
E-comm			0.00768**		0.00787**
			(2.22)		(2.30)
Sysadmin				0.0822	-0.0711
				(0.33)	(-0.29)
BG	-0.504**	-0.502**	-0.391*	-0.500**	-0.390*
	(-2.30)	(-2.29)	(-1.69)	(-2.26)	(-1.70)
CY	-0.553***	-0.552***	-0.465**	-0.549***	-0.466**
	(-2.79)	(-2.79)	(-2.29)	(-2.76)	(-2.31)
CZ	-0.136	-0.134	-0.182	-0.146	-0.173
	(-0.85)	(-0.85)	(-1.11)	(-0.90)	(-1.05)
DE	-3.365***	-3.367***	-3.449***	-3.388***	-3.429***
	(-13.52)	(-13.60)	(-13.85)	(-13.15)	(-13.17)
DK	-0.290*	-0.291*	-0.318*	-0.286*	-0.320*
	(-1.78)	(-1.80)	(-1.91)	(-1.74)	(-1.91)
EL	-0.639*	-0.639*	-0.459	-0.637*	-0.456
	(-1.87)	(-1.87)	(-1.31)	(-1.87)	(-1.31)
ES	-0.106	-0.110	-0.0715	-0.104	-0.0755
	(-0.70)	(-0.75)	(-0.47)	(-0.71)	(-0.49)

Table 28: Regression results: robustness checks (continued)

FI	-0.393**	-0.392**	-0.347*	-0.392**	-0.346*
	(-2.13)	(-2.13)	(-1.77)	(-2.12)	(-1.77)
FR	0.00912	-0.00145	0.0251	0.00698	0.0110
	(0.05)	(-0.01)	(0.15)	(0.04)	(0.06)
HR	-0.397*	-0.395*	-0.446*	-0.399*	-0.445*
	(-1.74)	(-1.73)	(-1.94)	(-1.73)	(-1.95)
HU	-0.558**	-0.558**	-0.536**	-0.577**	-0.520**
	(-2.33)	(-2.33)	(-2.28)	(-2.39)	(-2.18)
IE	-0.325*	-0.322*	-0.361**	-0.330*	-0.355**
	(-1.90)	(-1.89)	(-2.01)	(-1.89)	(-1.99)
IS	-0.520**	-0.522**	-0.515**	-0.522**	-0.515**
	(-2.09)	(-2.10)	(-2.06)	(-2.10)	(-2.08)
IT	-0.0501	-0.0583	0.0666	-0.0489	0.0570
	(-0.30)	(-0.36)	(0.38)	(-0.30)	(0.32)
LT	0.421**	0.424**	0.352**	0.417**	0.355**
	(2.38)	(2.36)	(2.08)	(2.41)	(2.15)
LU	-0.449**	-0.451**	-0.378*	-0.451**	-0.378*
	(-2.04)	(-2.06)	(-1.70)	(-2.04)	(-1.70)
LV	-0.0259	-0.0243	0.0495	-0.0312	0.0552
	(-0.14)	(-0.13)	(0.26)	(-0.17)	(0.29)
MK	-3.280***	-3.280***	-3.162***	-3.288***	-3.154***
	(-18.69)	(-18.92)	(-16.59)	(-18.39)	(-16.55)
MT	0.133	0.134	0.151	0.137	0.149
	(0.76)	(0.77)	(0.85)	(0.78)	(0.84)
NL	0.363**	0.361**	0.314**	0.354**	0.317**
	(2.37)	(2.37)	(1.99)	(2.32)	(2.01)
NO	-0.337*	-0.334*	-0.445**	-0.338*	-0.441**
	(-1.82)	(-1.81)	(-2.34)	(-1.81)	(-2.35)
PL	-0.313**	-0.314**	-0.347**	-0.325**	-0.340**
	(-2.00)	(-2.03)	(-2.09)	(-2.03)	(-1.99)
PT	-0.302*	-0.301*	-0.423***	-0.293*	-0.429***

Table 28: Regression results: robustness checks (continued)

	(-1.92)	(-1.91)	(-2.63)	(-1.88)	(-2.69)
RO	0.113	0.115	0.221	0.129	0.211
	(0.69)	(0.70)	(1.24)	(0.75)	(1.17)
SE	-0.432**	-0.432**	-0.442**	-0.430**	-0.443**
	(-2.31)	(-2.32)	(-2.35)	(-2.29)	(-2.37)
SI	-1.090***	-1.088***	-1.082***	-1.091***	-1.081***
	(-3.37)	(-3.37)	(-3.35)	(-3.37)	(-3.35)
SK	0.446***	0.445***	0.547***	0.456***	0.540***
	(2.80)	(2.74)	(3.23)	(2.80)	(3.15)
TR	-0.00319	-0.00292	0.113	-0.00303	0.118
	(-0.02)	(-0.02)	(0.66)	(-0.02)	(0.70)
UK	-3.419***	-3.415***	-3.337***	-3.432***	-3.322***
	(-14.03)	(-13.99)	(-13.27)	(-13.54)	(-12.71)
Constant	-2.370***	-2.371***	-2.464***	-2.380***	-2.458***
	(-15.16)	(-15.08)	(-14.51)	(-14.61)	(-14.41)
Observations	351	351	351	351	351

z statistics in parentheses; Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

C10-C12 is chosen as a base level; See Appendix A.19 for the full table