Economies of scope in data aggregation: evidence from health data*

Bruno Carballa-Smichowski[γ], Néstor Duch-Brown[γ], Seyit Höcük[ψ], Pradeep Kumar[ψ], Bertin Martens[φ], Joris Mulder[ψ], and Patricia Prüfer[ψ]

[γ]European Commission JRC, Sevilla, Spain
[ψ]Centerdata
[φ]TILEC, Tilburg University

March 15, 2023

Abstract

Economies of scope in data aggregation (ESDA) are attracting the attention of policymakers and researchers because of the efficiency gains they could bring about. Antitrust authorities, in turn, are concerned about their potential anti-competitive outcomes. However, the concept remains blurry and lacks empirical backing. We provide a definition: the improvement in the predictive power of a dataset resulting from adding complementary variables to it. It differs from traditional economies of scope, which are based on re-use of data or other resources. After deriving a theoretical model of ESDA, we estimate it by progressively adding explanatory variables to a dataset of health and health-related data that we use to predict health outcomes. Our three main findings confirm the existence of ESDA and lead to novel policy implications. First, in our dataset, a 1% increase in the number of predictor variables improves prediction accuracy in a range from 0.087% to 0.132%. Second, we find a positive non-linear relation between variable complementarity and ESDA. Third, in our models, ESDA are subject to increasing returns up to the third quartile of variables, and to diminishing returns thereafter. Our results support policies fostering the concentration of data in large pools with non-exclusive use rights.

JEL Classification: D24, L86, I10.

Keywords: Economies of scope, Health, Data aggregation, Predictive modelling, Machine learning

*We thank Maximilian Schäfer, Richard Haarburger and Emilio Calvano for their insightful comments on a previous version of this article. Any errors are the sole responsibility of the authors. The views and opinions expressed in this article are those of the authors and do not necessarily reflect the official policy or position of the Joint Research Centre, the European Commission or Centerdata.

1 Introduction

Over the last years, policy makers have become very much aware of the economic potential of digital data and big aggregated data pools (European Commission, 2020). At the same time, data collection by large digital platforms has triggered concerns about exclusive access to big datasets that can lead to monopolistic positions in downstream service markets (Crémer et al., 2019; Furman et al., 2019; Scott-Morton et al., 2019). More recently, high-profile antitrust cases have raised concerns about the combination of complementary datasets leading to new and/or higher-qualiy insights that can harm competition. In Apple/Shazam[1], the European Commission developed a new theory of harm according to which Apple, by acquiring of the music recognition app Shazam, could gain access to data about the preferences of music streaming customers and use it use to limit rival music streaming platforms capacity to compete. Similarly, the European Commission challenged Google's acquisition of the wearables producer Fitbit on concerns that the combination of Fitbit's databases on users' health and fitness data with Google's wide-ranging consumer databases, would create barriers to entry in online advertising. More recently, Amazon's 1.7 billion USD takeover of robot vacuum maker iRobot has triggered the Federal Trade Commission's scrutiny on similar grounds. One of the FTC's main concerns is that the data generated about consumers' homes by iRobot's robot vacuum cleaners will give Amazon an unfair advantage in the online retail market.[2] Moreover, the aggregation of datasets from firms located in different markets has been said to generate economies of scope that explain the conglomerate structure of Big Tech companies (Crémer et al., 2019).

While we acknowledge these competition-related concerns, this paper explores the benefits that come with data aggregation in large pools. We first take a more theoretical perspective on the economics of data aggregation. Scholars are increasingly referring to gains from economies of scope in data aggregation (Economides and Lianos, 2021; Antuca and Noble, 2021; Crémer et al., 2019; Krämer et al., 2020; Banalieva and Dhanaraj, 2019; Carballa-Smichowski, 2018; Mayer-Schönberger and Cukier, 2013). However, the concept remains fuzzy and sometimes conflated with economies of scope in data re-use. We present a clear distinction between these two concepts. We formalize the concept of economies of scope in data aggregation (ESDA) building on existing economic models of information and learning.

We then present an empirical application of this model in the health sector. Using an aggregated pool of health and health-related data from the Netherlands, we apply several machine-learning models to predict health outcome variables holding the number of observations constant. On that basis, we build a database relating the accuracy of these models' predictions to their characteristics and the number of variables included. We then use this database to estimate the magnitude of economies of scope in prediction scores of health outcomes, using regression analysis. Our findings confirm the existence of ESDA: a 1% increase in the number of explanatory variables increases prediction accuracy in a range from 0.087% to 0.132%. We also find evidence of a positive non-linear relation between variable complementarity (proxied by a reduction in variable correla-

[1] Apple/Shazam European Commission decision, case M.8788 (2018).

[2] See https://www.politico.com/news/2022/09/02/amazons-ftc-problem-keeps-growing-with-irobot-one-medical-probes-00054749

tion) and ESDA. Higher complementarity between variables has a positive impact on ESDA. ESDA are subject to increasing returns up to the third quartile of variables and diminishing returns thereafter.

This new interpretation of ESDA has novel policy implications. The classic concept of economies of scope in data re-use constitutes an argument in favour of lowering access barriers and wider sharing and diffusion of data. So does the existence of increasing returns to ESDA up to the third quartile of variables. At the same time, the new concept of ESDA implies efficiency gains that call for concentrating data in large pools. These two policies are not mutually exclusive. Large data pools do not necessarily imply the concentration of monopolistic control or ownership rights in the hands of a single agent. Shared use rights can bridge that gap.

The remaining of this article is structured as follows. Section 2 presents the theoretical literature referring to economies of scope in data, as well as the empirical literature on a related concept: economies of scale in data aggregation. Section 3 introduces and formalizes the novel concept of ESDA. In doing so, it traces its origins to information economics and the economics of learning. Section 4 describes the original data sources and the machine learning models deployed to obtain the final dataset. Section 5 uses the final dataset to evaluate the existence of economies of scope in the aggregation of health and health-related data. It also estimates two of its main properties: the effect of variable complementarity on ESDA and returns to scope in data aggregation. Section 6 concludes by discussing some policy implications of our results.

2 Related literature

Several recent theoretical articles have started to refer to "economies of scope in data", although in most cases without providing a precise definition of the term. In these contributions, economies of scope in data re-use (i.e., those stemming from re-using a dataset for different purposes) and ESDA (i.e., those stemming from adding more variables to a dataset while holding the number of observations constant)[3] are not distinguished. A notable exception is Calzolari et al. (2022), who are among the first to develop a theoretical model of economies of scale and scope in data aggregation. They define economies of scale as an increase in the number of observations from one source, and economies of scope as an increase in the number of different sources or data diversity, both subject to decreasing returns. They show that an aggregator prefers running a cooperative data pool with contributing firms that are not too close competitors but also not in entirely unrelated markets, with a preferred intensity of competition that diverges from the socially efficient level. The properties of their model are very similar to the model that we present in this paper. However, empirical evidence of these two types of economies of scope is scant (Tucker, 2019).

In a reference to Google/Fitbit[4], Economides and Lianos (2021) mention how Google's capacity to collect data in different dimensions generates "economies of scope in data" and

[3]We will provide a more precise definition of economies of scale in data aggregation in the next section.

[4]Google/Fitbit European Commission decision, case M.9660 (2020)

"significant learning effects, reinforcing the entrenched dominant position of the platform, both within its ecosystem, and outside". In a more general overview of the competitive effects of data, Antuca and Noble (2021) claim that "there tend to be economies of scope in data, meaning that there are increasing returns from the combination of data". Crémer et al. (2019) mention that "data conglomerates" can exploit "the economies of scope they can realize when combining their own data troves with that of another firm that is dominant in a separate market, to expand their own dominant position". Krämer et al. (2020) refer to "economies of scope", which they define as "more heterogeneous data about users or 'deep data'" and to "economies of scope in data aggregation", for which they do not provide a definition.

Banalieva and Dhanaraj (2019) employ the term "economies of scope in data analytics" to refer to situations in which a firm can "collect larger and more diverse datasets that yield sharper business insights for the company than could smaller, non-integrated datasets". The definition is ambiguous in that it combines the effect of "larger" datasets in terms of observations (economies of scale in data aggregation) with that of more "diverse" ones (economies of scope in data aggregation) in providing higher-quality insights. Carballa-Smichowski (2018) observes that the value stemming from a dataset's scope is twofold. First, "data gains value through the enhanced utility that comes from linking datasets". Second, "data-based economies of scope" arise when the investment to create a dataset "can result in a variety of uses". Finally, Mayer-Schönberger and Cukier (2013) mention that additional value can be created by combining two datasets or "recombining data".

To the best of our knowledge, no article has focused on empirically measuring ESDA. However, some contributions have taken steps in that direction. The closest contributions to ours are Schäfer and Sapi (2022)[5] and, to a lesser extent, Bajari et al. (2019). Schäfer and Sapi (2022) use a local polynomial regression on Yahoo! Search logs to show that the quality of the search results (measured as the click-through rate of the first organic search result) improve with more data on previous searches. As such, the article measures economies of scale in data aggregation. However, one of their results refers to the positive impact user history length has on the quality of search results. The authors interpret that, the longer a user history is, the more likely it is that the search engine has collected data about more of his/her observable characteristics. This could therefore be considered as an indirect measurement of ESDA, although the authors do not present it in these terms.

Bajari et al. (2019) come close to disentangling economies of scale and scope in data aggregation when they find that product sales forecasts do not become more accurate when historical data from several product markets are aggregated. In their case study on Amazon data, weak complementarity between product markets results in separable datasets. As we explain below, this fits with the predictions of our model of ESDA and the importance of complementary data. As such, Bajari et al. (2019) could be considered as a negative proof of ESDA[6] when complementarity is weak. As for the effects of scale,

[5]This work-in-progress paper constitutes the latest version of Schäfer et al. (2018).

[6]Tucker (2019) considers that Bajari et al. (2019)'s study "explores the question of whether there are economies of scope from digital data". However, Tucker (2019) uses the traditional definition of economies of scope in re-use: "an economy of scope occurs when there is a proportionate savings in costs

their linear regression design shows no effect or a positive effect of data aggregation on the accuracy of sales forecast, depending on the case. Finally, with a different focus and objective, one of the empirical results from Schouten (2018) is that adding additional variables to survey data used for causal inference should, under certain assumptions, decrease the proportion of unexplained variance. This occurs because the additional variables help reducing the selection bias generated by missing data for the already-included variables. Our results complement this finding by providing evidence of another reason for ESDA to emerge. With our approach, ESDA occur because additional variables provide new information about previously-unobserved features of the individuals observed.

Other related empirical studies have tried to estimate the impact of the size of datasets on the performance of predictive modelling. Hence, these studies' focus is the measurement of economies of scale in data aggregation. Lee and Wright (2021) develop a Bayesian model of a recommender system that learns the correlation structure and makes customized predictions based on the target user's history. Using a dataset containing over four million anonymous joke ratings, they test to which extent feeding the model with more data increases consumer welfare by providing customized recommendations. Neumann et al. (2018) use field experiments to investigate the impact of using black box data profiles on audience targeting accuracy. Their results indicate that, although this impact is positive, given the high costs and relative inaccuracy of targeting solutions, recurring to third-party audiences is usually economically unattractive.

Using a difference-in-differences design, Chiou and Tucker (2017) find no decrease in search engine accuracy when time series of observations of consumers' historical searches in Google, Yahoo! and Bing are shortened as a result of EU privacy regulation. Claussen et al. (2019) use a randomized experiment to show that an increase in individual user data helps algorithms outperforming human news editors in providing engaging recommendations (i.e., in obtaining more clicks per recommendation), but decreasing returns to user engagement sets in rapidly. McAfee et al. (2015) find that Google Search outperforms Microsoft Bing in long-tail searches because of a higher number of users. In a similar vein, Klein et al. (2022) find that a small search engine can produce equally good search results as the largest search engine (Google) for popular queries, but not for infrequent long-tail queries. The latter two studies already hint at the fact that both economies of scale and scope in data aggregation may play a role in search engine quality: the number of searches and the variety of search terms in the long tail of data collected from its users. However, they do not attempt to disentangle these two effects. Agrawal et al. (2018) find a noticeable drop in the performance of their proposed model on a subsample of an advertising dataset, and they achieved better results by using the entire dataset. Sometimes, the aggregation of large numbers of datasets also results in sparse and high dimensional data. Junqué de Fortuny et al. (2013), for example, proposed a modified version of multivariate Naive Bayes that achieved marginal increases in performance of the model as training datasets continue to grow.

It is worth noting that many of the above-mentioned contributions that estimate economies of scale in data aggregation also find in most cases that there are diminishing (Schäfer and Sapi, 2022; Lee and Wright, 2021; Bajari et al., 2019; Claussen et al., 2019; Junqué de Fortuny et al., 2013) and, in some cases, constant (Bajari et al., 2019; Chiou

relative to an increased level of production of multiple products".

and Tucker, 2017) returns to scale in data aggregation. Two notable exceptions are Schäfer and Sapi (2022) and Lee and Wright (2021), who, in some cases, find S-shaped relations: first an interval of observations for which there are increasing returns, followed by an interval exhibiting diminishing returns. Schäfer and Sapi (2022) find a S-shaped relationship between user history length (which, as mentioned above, can be interpreted as a proxy of the number of variables about an individual) and the quality of search results. Using a dataset about joke ratings, Lee and Wright (2021), in turn, find a S-shaped relationship between the number of previous users a recommendation algorithm can learn from and the utility of the recommendations it provides.

3 A theory of economies of scope in data aggregation

In this section we distinguish between economies of scale in data aggregation and economies of scope in data re-use, on the one hand, and economies of scale in data aggregation, on the other hand. The latter being the focus of this article, we provide a theoretical definition from which we derive an equation to estimate it in the next section.

3.1 The classic concept of economies of scope (in data re-use)

The concept of economies of scope dates back to the 1980s. It originated in the literature on joint production of several goods by a single firm and the re-use of the same input to produce multiple outputs (Teece, 1980, 1982; Panzar and Willig, 1981). In this classic interpretation, economies of scope occur when a single input or asset can be re-used to produce several distinct outputs. An alternative formulation is that economies of scope occur when it is less costly to combine two or more product lines in one firm than to produce them separately. Formally, $C(n_1, n_2) < C(n_1) + C(n_2)$, where $C(.)$ is a cost function and $n_{i=1,2}$ are inputs.

Such cost savings emerge when an input or production factor that is used for one product has spare capacity that can be used to produce another product. This results in lower joint production costs compared to a situation where that production factor would be supplied separately for the production of the second good. For example, the same machine can be used to produce several goods. Economies of scope are distinct from economies of scale because cost savings are due to an expansion of the scope of the firm (the number of distinct products) rather than the scale of production (the quantity of one product). Teece (1980, 1982) argued that widening the scope of the firm to multiple products could be an indication of market failure. If markets work well, spare capacity in one resource can be sold or rented out to another firm that already has the complementary production factors required to produce the second good. Keeping production inside the firm indicates that spare capacity and complementary production factors cannot be traded easily and efficiently between firms.

While this early literature focused primarily on economies of scope in the re-use of rival physical goods, the same reasoning can be applied to non-rival immaterial products. Rival goods, such as machines, can only be used for one purpose or by one party at the same time. By contrast, non-rival immaterial products, such as designs, patents, and media products, can be used for many purposes at the same time. In this case, re-use requires copying a set of electronic files that constitute the content of the product. That

gives an additional boost to resource savings. Suffice to invest only once in the production of a non-rival good in order to use it many times for many different purposes.

That is the promise of economies of scope in the re-use of data: many parties can use the same dataset at the same time for different purposes, without functional loss to the original data collector or user (Reimsbach-Kounatze, 2016; Jones and Tonetti, 2020). Data only have to be collected once at a fixed cost in order to be used many times and for many different purposes.[7] Copying digital data files usually happens at near-zero marginal cost. As such, re-use of non-rival data results in substantial cost-savings for society. This has led to the widely held view that society would benefit if data, once collected, would be shared as widely as possible for many uses. This would be social welfare enhancing because it would increase the combined benefits of data usage for many purposes at no additional data collection costs.

Palfrey and Gasser (2012) warned that this is a biased perception. Society can benefit from wider data access, but some stakeholders may suffer losses from such a policy. For example, people do not want their personal data to be widely shared and firms want to keep their commercial data confidential. While non-rivalry implies that the original use of data is not functionally affected by re-use for another purpose, the original data collector and user may face private economic opportunity costs from re-use by other parties (Martens et al., 2020). A gap between the social and private value of data sharing – the definition of a market failure – then blocks voluntary data sharing. Regulators can overcome this market failure by making data sharing mandatory (Graef and Prüfer, 2021). They consider decide that private losses are sufficiently small compared to overall societal benefits, or they may compensate losers.[8]In extreme cases, data sharing may reduce the market value of data for the original data collector to zero, or below the cost of data collection, and thereby eliminate incentives to invest in data collection (Bergemann and Bonatti, 2019). A more efficient way to exploit economies of scope in data re-use would be to identify a more optimal balance between zero and full data sharing.

3.2 Economies of scope in data aggregation

In the previous subsection we showed how the traditional concept of economies of scope can be applied to data, resulting in a definition of economies of scope in data re-use. In this subsection, we aim to clarify the distinction between economies of scale and scope in data aggregation by introducing a new interpretation of economies of scope with respect to data aggregation. We first propose an intuitive way to distinguish between economies of scale and scope in data aggregation. We then present a formalization of the concept.

An intuitive and convenient way to distinguish economies of scale and scope is to consider a dataset as a two-dimensional spreadsheet. The number of columns represents the number of variables, and the number of rows represent the number of observations of these variables. When this two-dimensional dataset is used to make predictions, economies of

[7]While data are non-rival by nature, data collection channels may be rival. For example, consumer behaviour data collected when a person is active on one social media app or search engine, cannot be collected at the same time by a rival app, unless the original app would allow the competitor to place tracking cookies. See Krämer (2021).

[8]For an example of mandatory business-to-government data sharing in China, see for example Martens and Zhao (2021).

scale improve prediction performance due to an increase in the number of rows (observations on variables), while ESDA improve prediction performance due to an increase in the number of columns (explanatory variables). Both economies of scale and scope in data aggregation can run into diminishing returns. ESDA will emerge only if additional variables bring additional information into the dataset. Adding highly-correlated variables would only increase the number of substitutes without adding complementary information. Adding totally unrelated variables does not increase the information content either. In machine learning models such as the ones on which our estimation of ESDA in Section 5 is based, adding unrelated variables amounts to adding noise. This can lead to a curse of dimensionality whereby adding predictor variables decreases prediction accuracy. Hence, ESDA only exist between complementary datasets.

The above example shows that the basic intuition of ESDA is simple. When two complementary datasets are merged or aggregated into a single data pool, more accurate insights or predictions can be extracted from the aggregated dataset. That increases the value of the data pool for decision making, compared to decisions made on the basis of the segmented datasets, and thus its economic value. While economies of scope in the re-use of data result from savings in data collection costs, ESDA result from additional insights or benefits that can be extracted from a merged dataset - insights that cannot be obtained from the separate datasets. In essence, the whole is greater than the sum of the parts.

ESDA are not entirely new. Its origins can be traced back to Blackwell's (1953) theorem on the comparison of information content in experiments. Crémer (1982) formulates Blackwell's theorem as follows: a dataset x_1 is more informative than x_2 if an only if x_2 is a subset of x_1. Alternatively, if two datasets x_1 and x_2 are aggregated into a common pool $x_{1,2}$, that aggregation is more informative than each separate set, unless $x_1 \cap x2 = x1 = x2$; that is, when the two subsets have identical content. Even if they are only partially overlapping, that is, when $x_1 \cap x_2 \neq \varnothing$ and $x_1 \Delta x_2 \neq \varnothing$, the aggregated dataset is still more informative. Blackwell's theorem only considers the information content of a dataset. It does not consider relationships between datasets, except for overlapping or substitute content. There may be other relationships between datasets. For example, one dataset may contain prices of goods and another dataset may have information on the quantities sold of these same goods. Aggregating the two datasets does not only allow calculation of the sales volume but may also enable estimation of price elasticities that is useful to design better pricing strategies. These insights cannot be extracted from examining the two datasets separately. Moreover, the Blackwell theorem does not consider the costs and benefits of extracting information content from a dataset. Some information may not be extracted because it is too costly to do so or the benefits of doing so do not justify the costs. We must turn to economic models of learning to bring these aspects into a model of ESDA.

Rosen (1983) presents a model of human learning and the choice to learn a wide range of skills or specialise in a narrow set. Learning[9] comes at a fixed cost that can be recuperated when skills are put to work in the labour market in return for a wage. Focusing on a narrow set of skills reduces the total costs of learning. Learning more skills will increase the costs of learning. However, revenue will not increase because working time will have

[9]We define learning as the extraction of insights from datasets that can be used for decision-making.

to be split between both skills. Widening the range of skills is not advantageous in that case. It becomes advantageous only when learning costs are not fully separable, that is when learning one skill reduces the cost of learning another. That happens when learning is applied to complementary datasets. Learning simultaneously from two complementary datasets decreases the costs, or increases the benefits, compared to learning from each set separately. As the above example of learning from sales prices and sales quantities datasets suggests, learning from each set separately does not yield the same valuable and actionable commercial strategy insights as learning from the combined or aggregated set because the two sets complement each other. As a result, there are economies of scope in learning from both sets, provided the additional insights extracted from the merged dataset are sufficiently large to overcome the increased costs of learning from the merged set (compared to learning from each dataset separately). While Rosen (1983) emphasizes cost savings in learning due to complementarity in the information content of two datasets, we can reformulate this in terms of net costs, or benefits minus costs, from learning. When the merged datasets are more informative than the separated datasets, the benefits of learning from the aggregated set exceed those from learning from the separated sets, provided additional learning costs are lower than additional benefits.

We can formalize our notion of ESDA in the following way. Let $f(x_i)$ be the production function indicating the degree of information that can be extracted from dataset x_i. We can say that $f(x_i)$ has a higher degree of information than $f(x_j)$ if $f(x_i)$ provides more pieces of (useful) information than $f(x_j)$, or if $f(x_i)$ provides the same information than $f(x_j)$ with more accuracy.

Suppose that the firm can exploit two datasets x_i and x_j containing different variables for the same number of observations (i.e., $x_i \neq x_j$ so that $x_i \cap x_j \neq \varnothing$ or $x_i \Delta x_j \neq \varnothing$) either separately or jointly as a single merged dataset. If the firm exploits them separately, it will obtain a degree of information of $f(x_i) + f(x_j)$. If, on the contrary, the firm decides to exploit the datasets jointly, the obtained degree of information will be $f(x_i + x_j)$.

If we define the degree of information as simply the number of pieces of useful information that can be extracted form a dataset, then, there will be ESDA if $f(x_i + x_j) > f(x_i) + f(x_j)$. In other words, there are ESDA if the merged dataset allows to obtain at least one more piece of useful information than the sum of the pieces of information that can be extracted by exploiting the two datasets separately.

Generalizing this condition to n possible datasets that can be used to obtain the required pieces of information, there will be ESDA in terms of new pieces of useful information if

$$f(x_i + \sum_{1}^{n-1} x_{j \neq i}) > f(x_i) + \sum_{1}^{n-1} f(x_{j \neq i}) \tag{1}$$

Alternatively, we can define the degree of information as the accuracy of the information obtained from a dataset. In that case, there will be ESDA if:

$$f(x_i + x_j) > max\{f(x_i), f(x_j)\}$$

In other words, there are ESDA if the merged dataset provides a piece of information with more accuracy than the one that can be obtained from any of the sub-datasets

that constitute it.[10] Generalizing this condition to n possible datasets that can be used to obtain the required information, there will be ESDA in terms of the accuracy of the obtained information if

$$f(x_i + \sum_{1}^{n-1} x_{j\neq i}) > max\{f(x_i), ...f(x_n)\} \tag{2}$$

Hereafter, we will focus on the measurement of economies of scope in terms of Equation 2.[11]

Re-writing the merged dataset as $x \equiv x_i + \sum_{1}^{m} x_{j\neq i}$ with $m < n$, we can obtain Equation 3 as a corollary of Equation 2.

$$\frac{\partial f(x)}{\partial m} > 0 \tag{3}$$

In other words, there are ESDA if, given an initial dataset, merging it with different datasets containing different variables for the same observations always increases the accuracy of the same piece of information that can be extracted from these datasets.

If we consider that the total existing n datasets contain the V total variables v_i that can be used to obtain the piece of information, and that the amount of variables $\gamma(m)$ increases monotonously with the number of m datasets added (i.e., adding a dataset always increases the amount of variables used), we can rewrite Equation 3 as:

$$\frac{\partial f(\frac{\gamma(m)}{V})}{\partial \frac{\gamma(m)}{V}} > 0 \tag{4}$$

In other words, ESDA exist if, when the percentage of variables $\frac{\gamma}{V}$ contained in a dataset x created by merging sub-datasets increases, the accuracy of the information extracted from x increases.

Based on Equation 4, we can obtain a measurement of the magnitude of returns to scope in data aggregation in terms of elasticity:

$$\frac{\partial f(\frac{\gamma(m)}{V})}{\partial \frac{\gamma(m)}{V}} \frac{\frac{\gamma(m)}{V}}{f(\frac{\gamma(m)}{V})} \tag{5}$$

Moreover, from Equation 4 we can obtain an equation that allows us to evaluate the nature of returns to scope in data aggregation:

$$\frac{\partial^2 f(\frac{\gamma(m)}{V})}{\partial (\frac{\gamma(m)}{V})^2} \tag{6}$$

[10] As mentioned above, this should only be the case if the sub-datasets are complementary. In a machine learning model, including non-complementary variables adds noise and can lead to the curse of dimensionality whereby model accuracy decreases as variables are added.

[11] Consequently, hereafter, we will use the terms "economies of scope" or "economies of scope in data aggregation" interchangeably to refer to an increase in the accuracy of the information obtained by exploiting a merged dataset in comparison to the accuracy that any of the sub-datasets composing it can provide for the same piece of information.

If $\frac{\partial^2 f(\frac{\gamma(m)}{V})}{\partial(\frac{\gamma(m)}{V})^2} = 0$, there are constant returns to scope in data aggregation. If $\frac{\partial^2 f(\frac{\gamma(m)}{V})}{\partial(\frac{\gamma(m)}{V})^2} > 0$, there are increasing returns to scope in data aggregation and if $\frac{\partial^2 f(\frac{\gamma(m)}{V})}{\partial(\frac{\gamma(m)}{V})^2} < 0$, there are diminishing returns to scope in data aggregation.

To calculate the magnitude of returns to scope in data aggregation, let us consider an increase in the percentage of variables used (i.e., an increase in the scope of the dataset) from $(\frac{\gamma(m)}{V})_1$ to $(\frac{\gamma(m)}{V})_2$ with $(\frac{\gamma(m)}{V})_2 > (\frac{\gamma(m)}{V})_1$. Then, the change in the returns of adding a $(\frac{\gamma(m)}{V})_2 - (\frac{\gamma(m)}{V})_1$ percent of total variables V to the dataset is equal to the change in the returns to scope:

$$\frac{\partial^2 f(\frac{\gamma(m)}{V})}{\partial(\frac{\gamma(m)}{V})^2}_2 - \frac{\partial^2 f(\frac{\gamma(m)}{V})}{\partial(\frac{\gamma(m)}{V})^2}_1 \tag{7}$$

In the following section, we estimate Equation 5 to measure ESDA and Equation 6 to evaluate the nature of returns to ESDA using health and health-related data.

4 Setting and data

We propose an empirical test of ESDA regarding the predictive performance of health outcomes using a wide variety of personal health and non-health data. This choice is grounded on prior contributions that have used health and non-health data to predict health outcomes through machine learning, although with different goals to ours. Larrabee (2008) showed that aggregating multiple indicators improved the detection of malingering, invoking health reasons to stay away from work. Colbaugh and Glass (2017) proposed a novel machine learning methodology that facilitates accurate individual-level prediction models from aggregated data. De Giorgi et al. (2021) apply machine learning to a seemingly unrelated combination of mortality data and credit card data to produce accurate predictions of death rates. In our case, the prediction of health outcomes through machine learning constitutes a first step in order to build a database. We then use this database in Section 5 to estimate ESDA and some of its properties through regression analysis.

4.1 Data sources

Our empirical test combines data from two sources in the Netherlands, the Longitudinal Internet studies for the Social Sciences (LISS) socio-economic survey dataset[12], collected by Centerdata since 2007, and individual-level microdata for the entire population, collected by the Netherlands Central Bureau of Statistics (CBS). LISS panel members have a unique ID that enables matching with CBS population data. Matching was carried out in a secure CBS computing environment. Only LISS respondents who gave consent to CBS data coupling were considered. The coupled data constitutes a substantial dataset with many observations on a large number of variables. The LISS panel covers about 5,000 Dutch households, comprising approximately 7,500 individuals who complete monthly online surveys since 2007. The multiple surveys within the LISS panel cover a broad

[12]For more details, see the "About the panel" section in `lissdata.nl`

range of topics, such as health, personality, income and assets, spending leisure time, politics, and religion. The LISS panel composition evolves over time but is relatively stable and has an average individual response rate of 79%. Participants are randomly selected from the Dutch population register.

The LISS health dataset comprises 127 health-related survey questions. These include somewhat subjective questions, such as "How good is your health?", as well as more objectively verifiable replies, such as "Do you currently smoke?".

In addition, several survey topics from the LISS panel were selected to augment the dataset with background or health-related data for this research. The selected topics are:

- family and household,
- work and schooling,
- personal and income situation,
- and demographic information on panel members.

The other data source, the CBS microdata, contains registry data of individuals on a wide range of subjects, starting in the year 2006. For the current research, we used prescribed medicine registry data from the theme "Health and wellbeing" for the entire Dutch population. The CBS medicine data exist in long format that was transformed to World Health Organization's "Anatomical Therapeutic Chemical" (ATC) code categories at level 4[13], resulting in 187 new predictor variables. The overlapping share between individuals registered in the CBS medicine dataset and the LISS panel participants reaches 92%. Other CBS datasets that were originally selected for this study, including mental healthcare treatments, received social assistance (Dutch WMO legislation) and received disability benefits, were eventually excluded because of insufficiently-low population overlaps with the LISS panel participants.

We did not use the full range of available time periods of the two datasets for our research. Targeting more years increases the number of observations on variables, but not the number of participants. This is because the list of selected participants is fixed by the selection of the target variable for the target year 2019. The further we go back in time, the more data we have per participant, but no new participants. Moreover, considering the occasional dropout in participants in the LISS panel, there is a few percent attrition of panel members yearly. A period of five years captured most of the data of the participants.

Another consideration was that the CBS data for 2020 were not yet published at the time of this research. Moreover, LISS data and CBS data may not be very representative for the year 2020 because of the COVID-19 pandemic. Hence, we eliminated year 2020. The combined dataset is thus restricted to the period 2015-2019 (5 years), but we select two time periods, 2018-2019 and 2015-2019, to assess the model (in)dependence on the period selection.

[13]For more details see Anatomical Therapeutic Chemical (ATC) Classification in who.int

We started with 2,007 LISS panel variables and a total of 89,611 observations. Cleaning eliminated 1,497 variables, leaving 512. This seems like a drastic decrease, but a lot of the data were excluded because they contained unrelated or unusable information. The cleaning steps involved removing textual data, excluding years beyond the selected range, discarding unfilled or uniformly filled (e.g., all 'yes') answers, eliminating sparsely filled variables (i.e., more than 70% missing), and disposing of research-unrelated variables (e.g., the round of the questionnaire). For the CBS dataset, the 187 variables did not require further cleaning since they are ready-to-use registry data that are extensively checked and are thus already cleaned.

The final processing step involved computing the pairwise correlation matrix between all variables in the dataset. We discard one of every pair of strongly correlated variables from the dataset to improve the performance of our prediction model.[14] We use Pearson's r or bivariate correlation that ranges from -1 to 1, but we take the absolute value for the decision. We select three arbitrarily chosen threshold levels (0.5, 0.7, and 0.9) to study the impact of removing correlated or anti-correlated variables from the analysis. The number of (anti)correlated variables removed for the 0.9 threshold is, for example, 22. Lower correlation thresholds result in the removal of more variables.

While removing correlated variables is important to avoid bias in modeling, correlation threshold levels are used here as a proxy for complementarity between the predictor variables. As explained in the previous section, ESDA increase when two merged datasets are more complementary. Pure substitute datasets add no additional information and do not generate economies of scope. We test this hypothesis by varying the correlation threshold level in our models.

The two datasets together contained 699 variables. This included 35 categorical variables that were later transformed into 143 binary variables using one hot encoding, which is more suited for modeling purposes.[15] In the end, the cleaned and transformed merged dataset comprises 807 variables (columns) and 22,792 observations (rows).

4.2 The machine learning models

The choice of a health outcome variable as the predicted variable may have an impact on results. Some health outcome variables may be more susceptible to prediction through data aggregation, others less so. Moreover, some health outcome variables can be more subjective in nature, while others are more objective. The former are better suited for machine learning model prediction. While objective variables can be accurately predicted using some key predictor variables, subjective variables require a vast amount of predictor variables about individuals' health and socio-economic characteristics to be accurately predicted. This makes subjective variables more suited to machine learning predictions than objective variables, which can be accurately predicted using regression analysis. We therefore choose the two following subjective health outcome variables in the dataset as our target variables:

[14] Since they are highly correlated, it does not matter which variable is dropped.

[15] For example, 'Marital status' has 5 categories, namely: married, divorced, economically separated and legally married, widow/widower and never married.

A. **Perceived health**, in the LISS health questionnaire: "How would you describe your health, generally speaking?" on a scale from 1 to 5. The answers were regrouped into three categories. This variable is very subjective.

B. **Functional disability**, involving three related survey questions that are reconstructed into a single reply: "To what extent did your physical health or emotional problems hinder your activities or work over the past month?", on a scale from 1 to 5. This variable is still rather subjective.

The remaining survey questions are considered as predictor variables. Note that the purpose of our prediction model is not to determine causal effects. Our only objective is to estimate economies of scope in prediction accuracy with regard to data aggregation.

In order to corroborate the pertinence of restricting our choice to subjective variables, we re-estimated economies of scope in models 1-3 (Equations 8-10 in Section 4.4.1 below) for two objectively-measurable health outcome variables: "Chronic lung disease" and "Prescribed medicine for respiratory diseases". The results show that machine learning models predicting objective variables suffer from insurmountable heteroscedasticity problems.

We train several prediction models on a progressively increasing number of predictor variables from our dataset and observe how these increases affect the predictive performance of the pre-selected health outcomes. We increase the number of the randomly selected sample of variables in incremental steps of 5 percentage points, from 5 to 100 percent of all available predictor variables. For each additional 5% of randomly-selected variables added, the variables that are randomly selected exclude those that have been previously added, if any. For each number of variables, we train a machine-learning prediction model. We used supervised machine learning for predictive modeling. Supervised machine learning models can be divided into two groups, regression and classification models. Since the target variables as well as most of the survey data are categorical in nature, we apply classification models. We employ two different types of supervised machine learning algorithms, Logistic Regression and Random Forest. Both produce transparent model outputs. This means that the models enable us to determine the contribution of each variable to model performance.[16]

Finally, we excluded the models trained with 10% or less variables in the regression analysis. Not all the predictor variables are equally important for model prediction accuracy. Hence, a random selection of smaller subsets of the predictor variables can cause significant fluctuations in model performance. This, in turn, leads to heteroscedasticity when estimating the impact of variable percentage on prediction accuracy (cf. Equations 8-10 below). This can be attributed to the inclusion and exclusion of the more important variables in the models. The same holds true for the models built for the objective predicted variables (i.e, "Chronic lung disease" and "Prescribed medicine for respiratory

[16]Note that the ratio of parameters to be estimated by the machine learning models and the observations used for the training is low. That makes the predictions robust. Maximally, we have 807 parameters and 22 792 observations, which makes this ratio 0.035. Note also that the tree-based machine learning models used are not sensitive to a high ratio. In the case of models that are, the conventional maximum ratio tolerated is 0.1.

diseases"). In this way, we could obtain heteroscedasticity-free estimations of ESDA.

4.3 Machine learning models evaluation and final dataset

The performance of predictive machine learning models is assessed against several evaluation metrics that quantify model performance. The choice of an evaluation metric depends on a particular machine learning task, such as classification, regression, ranking, or clustering. The evaluation metrics considered in this article are:

- Accuracy, for both the train and test set
- Precision, also known as positive predictive value (PPV)
- Recall, also known as sensitivity or true positive rate (TPR)
- The F1-score, which is the harmonic mean of precision and recall

We focus on the F1-score for our application. The F1-score is particularly well suited for unbalanced data, which is the case in this study. Since the evaluation metrics are given at the class level, we adopt the strict macro-average class score as the final model performance. Appendix A provides more detail about the evaluation metrics.

To run and evaluate a machine learning model, the data must first be split into a training subset and a model evaluation (test) subset. When the number of variables is too small, but also to create robustness in results against random fluctuations, the k-fold cross-validation technique can be applied in conjunction with model evaluation (Refaeilzadeh et al., 2009). Parameter k refers to the number of groups or folds into which a given sample is to be split, hence the name k-fold. The data is first shuffled and then split into k groups. Each group is, in turn, reserved as a test set, while the remainder of the dataset is used for training. The model runs on the training set. In this way, each group gets its turn to become part of training and part of testing. This approach allows the optimal use of the whole dataset for testing and evaluating. We choose a typical value of $k = 5$, resulting in an 80% train and 20% test sample, which is then run five times. Any chosen evaluation metric is applied together with k-fold cross-validation technique.

Machine learning models and variable sampling methods involve randomness. To make the model results reproducible, fixed seed randomness is selected for all models and data sampling methods. To eliminate any possible dependence on the selected random seed, all model combinations are run twice for two different, but fixed, random seeds.

The combination of all variations in model parameters (two machine learning algorithms, three correlation thresholds, two random seeds, two time periods, and two target variables) results in 24 different types of models that are run twice (randomness), obtaining a total of 48 models. We run each of these 48 models using 20 different percentages of the available number of explanatory variables, ranging from 14% to 100% . This results in a database containing 864 observations.[17] Each observation includes attributes that describe the machine learning algorithm and random seeds used, time series included and

[17]Recall that models with a variable percentage of 10% and below were eliminated from the final dataset (cf. Section 4.2).

the F1-score or prediction accuracy, the percentage of explanatory variables used and the correlation threshold applied. Table 1 below provides the list of variables in the dataset and summary statistics for them.

Table 1: Summary statistics of the dataset

Variable	Variable description	Values	Values description	n	Perc.	Mean	Min	Max	SD
score	F1-score of the model	-	-	864	100	48.3	35.7	56.3	4.4
VarPerc	Percentage of predictor variables included in the model	-	-	864	100	51.2	13.6	100	23.4
TargVar	Health target variable	CH4	Subjectively perceived health	424	50	-	-	-	-
		DIS	Functional disability	424	50	-	-	-	-
algo	Algorithm used	LR	Logistic regression	424	50	-	-	-	-
		RF	Random forest	424	50	-	-	-	-
random	Random seeds used	3	-	424	50	-	-	-	-
		30	-	424	50	-	-	-	-
period	Years included	L	Long period (2015-2019)	424	50	-	-	-	-
		S	Short period (2018-2019)	424	50	-	-	-	-
corr	Correlation threshold between predictors	0.5	-	272	32.08	-	-	-	-
		0.7	-	288	33.96	-	-	-	-
		0.9	-	288	33.96	-	-	-	-

5 Evaluating economies of scope in data aggregation

In this section, we use the dataset presented in Table 1 to test for the existence of three characteristics of ESDA:

(i) The magnitude of ESDA

(ii) The fact that variable complementarity increases ESDA

(iii) The nature of returns to scope in data aggregation.

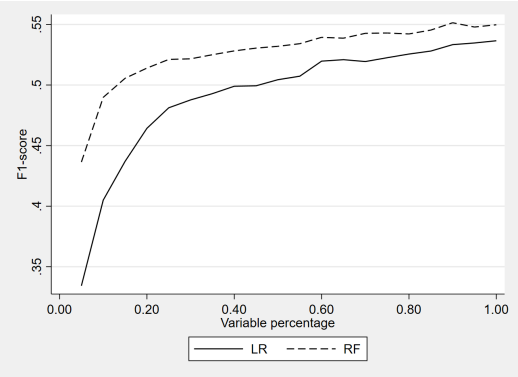
5.1 Existence and magnitude of economies of scope in data aggregation

Figures 1a and 1b below provide a first graphical representation of ESDA in our dataset. These graphs use both algorithms (RF and LR), data for the period 2018-2019 only, a correlation threshold level of 0.7 and the average result of the two random seeds. The graphs plot the predictive performance, as measured by the F1-score, against the share of the explanatory variables dataset used for the prediction. In line with the hypothesis on ESDA, we expect prediction accuracy to increase when we increase the number of explanatory variables – while keeping the number of observations on these variables constant. The two graphs visually confirm this hypothesis.

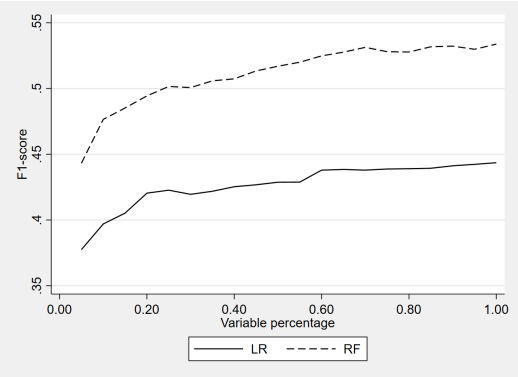
Note that prediction accuracy does not necessarily increase smoothly with the number of explanatory variables. The addition of important explanatory variables may cause a jump in model performance. Model performance may also occasionally decrease with an increasing number of variables. This may be because adding some (irrelevant) variables may induce noise in the algorithms, increase the dimensionality of the problem and

slightly reduce algorithm performance.

Figure 1: F1-scores for each target variable vs percentage of predictor variables



(a) Target variable: perceived health



(b) Target variable: functional disability

While the graphs already give a visual impression that confirms the existence of ESDA, we introduce a more rigorous quantitative measure of its magnitude. To this end, we use the dataset described in the previous section (cf. Table 1) and estimate Equation 5 with three alternative specifications (cf. Equations 8-10 below). In each specification, we include machine learning models' characteristics in different ways in order to test the robustness of the results to model specification.

$$\text{Model 1: } log(score) = \beta_0 + \beta_1 log(VarPerc) + \beta_2 RF + \beta_3 DIS + \beta_4 S + \varepsilon_i \tag{8}$$

$$\text{Model 2: } log(score) = \delta_0 + (\delta_1 RF + \delta_2 DIS + \delta_3 S) log(VarPerc) + \varepsilon_i \tag{9}$$

$$\begin{aligned}\text{Model 3: } log(score) = {} & \alpha_0 + \alpha_1 log(VarPerc) + \alpha_2 RF + \alpha_3 DIS + \alpha_4 S + \\ & (\alpha_5 RF + \alpha_6 DIS + \alpha_7 S) log(VarPerc) + \varepsilon_i\end{aligned} \tag{10}$$

Where $log(score)$ is the natural logarithm of the F1-score of the model, $log(VarPerc)$ is the natural logarithm of the percentage of variables included in it, RF is a dummy variable indicating whether the algorithm used is a random forest, DIS is a dummy variable indicating whether the predicted variable is "Functional disability", S is a dummy variable indicating whether the period of analysis is the short one (2018-2019), β_i, δ_i and α_i are estimated parameters and ε_i is the error term.

Each specification captures in a different way how the features of the machine learning models can affect the accuracy of the predictions. We distinguish between average effects on accuracy and slope effects. The slope of the graphs presented above reflects the magnitude of ESDA: the steeper the slope is, the stronger ESDA are. The level of the graphs reflects average accuracy of predictions. In model 1, each feature has an average effect on the accuracy of the prediction of the target variable. Hence, the features of the model do not affect the magnitude of ESDA, which are translated by coefficient β_1. In model 2, each dummy variable representing a feature of the model is interacted with the variable $VarPerc$. Hence, each feature affects the magnitude of economies of scope. Model 3 combines average and slope effects.

Note that we excluded variable *corr* to avoid multicollinearity when estimating Models 1-3. As the correlation threshold between variables (i.e., *corr*) increases, a higher percentage of the variables available (i.e., $VarPerc$) is used to train the machine learning models. We test the effect of the correlation threshold chosen on the intensity of ESDA in Section 5.2.

Table 2 shows the F1-scores for each target variable vs percentage of predictor variableslts of the estimation of equations 8-10 (Models 1-3, respectively) using Huber-White standard errors.

Table 2: Estimation of economies of scope in data aggregation

	(1) Model 1 Intercept dummies	(2) Model 2 Slope dummies	(3) Model 3 Intercept & slope dummies
Log(VarPerc)	0.091*** (0.004)	0.087*** (0.004)	0.132*** (0.008)
RF algorithm	0.090*** (0.004)		0.244*** (0.028)
Target variable 'Functional disability'	-0.076*** (0.004)		0.048*** (0.028)
Years 2018-2019	0.019*** (0.004)		0.057*** (0.028)
RF algorithm*log(VarPerc)		0.022*** (0.001)	-0.040*** (0.007)
Target variable 'Functional disability'*log(VarPerc)		-0.020*** (0.001)	-0.033*** (0.007)
Years 2018-2019*log(VarPerc)		0.005*** (0.001)	-0.010*** (0.007)
Constant	3.511*** (0.015)	3.528*** (0.015)	3.353*** (0.030)
Observations	864	864	864
$\boldsymbol{R^2}$	0.67	0.66	0.69

Note: Standard errors in parentheses: * p<0.05 ** p<0.01 *** p<0.001

The first line in Table 2 confirms the existence of ESDA in the dataset. Depending on the specification used, increasing the number of explanatory variables by 1% rises the accuracy of the prediction (i.e., the F1-score) by 0.087% to 0.132% in average.

In models 1 and 3, the comparison of the intercept dummies with the coefficient for variable $log(VarPerc)$, on the one hand, and with the constant, on the other hand, shows that the characteristics of the machine learning models have a statistically significant yet low impact on the accuracy of the prediction. For example, in model 1, adding 30% of total variables increases the F1-score by 2.73 points, while using a RF algorithm in-

creases the F1-score by only 0.09 points in respect to using a LR algorithm. The latter value (0.09) is considerably lower than that of the constant (3.511). Hence, variable percentage is the defining feature affecting the accuracy of the prediction of health outcomes.

The signs of the coefficients for intercept dummies in models 1 and 3 are coherent with the literature and our expectations. As shown by the literature, random forest algorithms provide higher accuracy than logistic regression ones. The selection of an optimal machine learning model for a given application is an ad-hoc process that is determined by the various quality metrics, for example, accuracy, F1-score, area under curve (AUC). Random forest (RF) and Logistic regression (LR) have achieved varying performance on different datasets. However, Couronné et al. (2018) carried out a large-scale benchmarking experiment to compare the predictive power of RF and LR and found that RF models outperformed LR models. Beğenilmiş and Uskudarli (2018) also showed in their study that RF models consistently performed better than LR and support vector machine (SVM) models. Moreover, we observe that predicting the variable "Functional disability" results in a lower model accuracy than predicting the variable "Perceived health". Using a shorter time span, in turn, has a small positive impact on model accuracy. This is because, for most variables, answers to the questions do not vary significantly from one year to another. Then, taking a larger timespan generates more noise in the dataset, which slightly lowers the accuracy of predictions despite the larger number of observations it brings about.

Model characteristics have a statistically significant impact on the magnitude of ESDA, although their importance varies depending on the model characteristic considered, as well on the specification used. A comparison between the values of the coefficient for the variable $log(VarPerc)$ and those corresponding to slope dummies in models 2 and 3 illustrates it.

In model 2, using a shorter time period (years 2018-2019 instead of years 2015-2019) has a statistically significant yet small positive impact on the level of economies of scope. In model 3, using a shorter time period does not have a statistically significant impact on economies of scope. The reason of this low or lack of effect is the same as for the positive impact of using a shorter time period on the overall accuracy of the model. Predicting the variable "Functional disability", in turn, has a significant and considerable negative impact on the level of economies of scope both in models 2 and 3. In these models, predicting this variable reduces the level of economies of scope by 23% and 25%, respectively. This shows that the magnitude of ESDA can vary considerably depending on the use that is given to data.

Similarly, the algorithm used has a strong impact on the magnitude of ESDA. In model 2, using a random forest algorithm increases the level of economies of scope by 25%. In model 3, it decreases it by 30%. The discrepancy in the sign of the coefficients for the variable between models 2 and 3 is due to the fact the specification of model 3 includes the effect of the variable RF both through intercept and slope dummies. Then, in model 3, the overall effect of including a RF algorithm has to incorporate the estimated coefficients for both the intercept α_2 and the slope dummy α_5. For every value of $VarPerc$ between 0 and 100, model 3 provides a higher predicted $log(score)$ than if parameters α_2 and α_5 were equal to zero. Therefore, the impact of including a RF algorithm on the

F1-score is positive, which is consistent with the results of other models and the literature.

5.2 Effect of variable complementarity on economies of scope in data aggregation

In this section, we test whether variable complementarity has a positive impact on the magnitude of ESDA. In order to do so, we vary the admitted correlation threshold level between the predictor variables used in the machine learning models. As explained in Section 4.1, these are equal to 0.5, 0.7, and 0.9. Variables that exceed the correlation threshold are eliminated from the dataset. The higher the admitted correlation between two explanatory variables, the less complementary they are, and vice versa.

We test the hypothesis with the following simple regression (cf. Equation 11 below) run three times, once for each of the three correlation threshold values (0.5, 0.7 and 0.9). As mentioned in the previous section, we discarded including both *corr* and $VarPerc$ as independent variables to explain *score* in order to avoid multicollinearity. Hence, we opted for Equation 11 as our estimation equation. The results of this estimation using Huber-White standard errors are shown in Table 3 below. Variable description is identical to that of equations 8-10.

$$log(score) = \beta_0 + \beta_1 log(VarPerc) \tag{11}$$

Table 3: Effect of variable complementarity on economies of scope in data aggregation

	(1) Corr=0.5	(2) Corr=0.7	(3) Corr=0.9
Log(VarPerc)	0.114***	0.071***	0.069***
	(0.010)	(0.009)	(0.009)
Constant	3.407***	3.618***	3.635***
	(0.036)	(0.035)	(0.035)
Observations	272	288	288
$\boldsymbol{R^2}$	0.36	0.21	0.18

Note: Standard errors in parentheses: * p<0.05 ** p<0.01 *** p<0.001

As Table 3 shows, an increase in the correlation threshold, that is, a decrease in complementarity between added variables, diminishes the magnitude of ESDA. This corroborates the theoretical hypothesis derived in Section 3 according to which ESDA depend positively on variable complementarity. Note that the link between variable complementarity and the level of economies of scope is not linear. A decrease of 0.2 points in the correlation threshold from 0.7 to 0.5 results in a drop in the level of economies of scope of 0.043 points. The same decrease of 0.2 points in the correlation threshold from 0.9 to 0.7, in turn, results in economies of scope diminishing by 0.002 points only. This indicates decreasing returns to complementarity in the data.

5.3 Returns to scope in data aggregation

We now evaluate the nature of returns to scope in data aggregation. In Figures 1a and 1b, as variable percentage grows, the accuracy of the model increases at a decreasing rate. This suggests the existence of diminishing returns to scope in data aggregation, which echoes the majority of the empirical evidence regarding economies of *scale* in data aggregation (Schäfer and Sapi, 2022; Lee and Wright, 2021; Bajari et al., 2019; Claussen et al., 2019; Junqué de Fortuny et al., 2013). In terms of Equation 6 above, the existence of diminishing returns to scope implies that, if the percentage of variables is above a threshold value $(\frac{\gamma(m)}{K})^*$ (i.e., if $\frac{\gamma(m)}{K} > (\frac{\gamma(m)}{K})^*$), we should expect that:

$$\frac{\partial^2 f(\frac{\gamma(m)}{V})}{\partial(\frac{\gamma(m)}{V})^2} < 0 \tag{12}$$

If the sign of this second derivative is negative as in Equation 12, the coefficients for the variable $log(VarPerc)$ in models 1-3 should be lower when these models are estimated for higher values of $log(VarPerc)$. Conversely, if the sign is positive, the coefficients for the variable $log(VarPerc)$ in models 1-3 should be higher when these models are estimated for higher values of $log(VarPerc)$.

Following this logic, we test the nature of returns to scope in data aggregation by re-estimating models 1-3 for four sub-samples, one for each quartile range of the variable $log(VarPerc)$: 0-Q_1, Q_1-Q_2, Q_2-Q_3 and Q_3-100. This allows us to estimate models 1-3 for increasing variable percentages while retaining enough observations in each estimation. As a result, we can make robust estimations and smooth out occasional drops in prediction accuracy resulting from the inclusion of irrelevant variables (cf. Figures 1a and 1b). Moreover, using quartiles gives us balanced subsamples between the four estimations of each model. Tables 4-6 below show the results of these estimations for models 1-3 (Equations 8-10), respectively.

Table 4: Returns to scope in data aggregation for Model 1 (intercept dummies)

Variable percentage: x	(1) Model 1a $x \leq Q_1$	(2) Model 1b $Q_1 < x \leq Q_2$	(3) Model 1c $Q_2 < x \leq Q_3$	(4) Model 1d $x > Q_3$
Log(VarPerc)	0.112***	0.117***	0.132***	0.009***
	(0.016)	(0.028)	(0.039)	(0.027)
RF algorithm	0.123***	0.090***	0.080***	0.067***
	(0.008)	(0.008)	(0.007)	(0.006)
Target variable 'Functional disability'	-0.052***	-0.077***	-0.082***	-0.092***
	(0.008)	(0.008)	(0.007)	(0.006)
Years 2019-2019	0.029***	0.015	0.013	0.018**
	(0.008)	(0.008)	(0.007)	(0.006)
Constant	3.415***	3.415***	3.349***	3.495***
	(0.050)	(0.102)	(0.160)	(0.119)
Observations	216	208	208	216
$\boldsymbol{R^2}$	0.62	0.57	0.58	0.65

Note: Standard errors in parentheses: * p<0.05 ** p<0.01 *** p<0.001

Table 5: Returns to scope in data aggregation for Model 2 (slope dummies)

Variable percentage: x	(1) Model 2a $x \leq Q_1$	(2) Model 2b $Q_1 < x \leq Q_2$	(3) Model 2c $Q_2 < x \leq Q_3$	(4) Model 2d $x > Q_3$
Log(VarPerc)	0.096***	0.114***	0.131***	0.100***
	(0.016)	(0.028)	(0.039)	(0.027)
RF algorithm*log(VarPerc)	0.040***	0.024***	0.020***	0.015***
	(0.003)	(0.002)	(0.002)	(0.001)
Target variable 'Functional disability'*log(VarPerc)	-0.017***	-0.021***	-0.020***	-0.021***
	(0.003)	(0.002)	(0.002)	(0.001)
Years 2018-2019*log(VarPerc)	0.009***	0.004*	0.003	0.004**
	(0.003)	(0.002)	(0.002)	(0.001)
Constant	3.466***	3.429***	3.356***	3.490***
	(0.050)	(0.102)	(0.160)	(0.119)
Observations	216	208	208	216
$\boldsymbol{R^2}$	0.62	0.57	0.58	0.65

Note: Standard errors in parentheses: * p<0.05 ** p<0.01 *** p<0.001

Table 6: Returns to scope in data aggregation for Model 3 (slope and intercept dummies)

Variable percentage: x	(1) Model 3a $x \leq Q_1$	(2) Model 3b $Q_1 < x \leq Q_2$	(3) Model 3c $Q_2 < x \leq Q_3$	(4) Model 3d $x > Q_3$
Log(VarPerc)	0.193***	0.130*	0.141	0.117*
	(0.036)	(0.055)	(0.079)	(0.048)
RF algorithm	0.314**	0.089	-0.080	0.004
	(0.095)	(0.202)	(0.314)	(0.237)
Target variable 'Functional disability'	0.147	0.094	0.297	0.172
	(0.095)	(0.202)	(0.314)	(0.237)
Years 2018-2019	0.127	-0.066	-0.109	-0.038
	(0.094)	(0.205)	(0.319)	(0.262)
RF algorithm*log(VarPerc)	-0.063*	0.000	0.039	0.014
	(0.031)	(0.054)	(0.077)	(0.01)
Target variable 'Functional disability'*log(VarPerc)	-0.065*	-0.046	-0.093	-0.060
	(0.031)	(0.054)	(0.077)	(0.054)
Years 2018-2019*log(VarPerc)	-0.032	0.022	0.030	0.013
	(0.030)	(0.055)	(0.078)	(0.059)
Constant	3.167***	3.370***	3.312***	3.416***
	(0.112)	(0.203)	(0.324)	(0.211)
Observations	216	208	208	216
$\boldsymbol{R^2}$	0.62	0.57	0.58	0.65

Note: Standard errors in parentheses: * p<0.05 ** p<0.01 *** p<0.001

The values of the coefficient for the variable $log(VarPerc)$ in Tables 4-6 point to the same conclusion. There are increasing returns to ESDA up to the third quartile

(Q_3) and decreasing returns thereafter, which in our dataset is equal to 69.1% of the variables. In other words, the relationship between the percentage of variables included in the model and the accuracy of prediction is S-shaped. This result is consistent with Schäfer and Sapi's (2022) insights on the impact of user history length on the quality of a search engine's results, and with Lee and Wright's (2021) conclusions on how the number of previous users a recommendation algorithm can learn from affects the utility of the recommendations it provides.

5.4 Robustness checks

Our three empirical results rest on Models 1-3, either directly by estimating them for the entire sample (Section 5.1), in a quantile regression (cf. Section 5.3) or indirectly by estimating a simplified linear specification for different values of the variable $Corr$ (Section 5.2). Therefore, to check for the robustness of our results, we performed two robustness checks of Models 1-3. The first one aims at verifying the pertinence of a linear specification. It consists in a non-parametric kernel regression of the following form:

$$log(score) = g(log(VarPerc)) + \phi_2 RF + \phi_3 DIS + \phi_4 S + \varepsilon_i \tag{13}$$

Where $g : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ is a function and ϕ_i are estimated parameters. The detailed results of the estimation of Equation 11 are shown in Appendix B. For all the estimated coefficients ϕ_i, the results are qualitatively identical to those stemming from the estimations of Models 1-3. The estimated function $g(\cdot)$, which translates the nature and degree of economies of scope in data aggregation, is a linear function in which the coefficient of $log(VarPerc)$ is virtually identical to the corresponding estimated coefficient for Model 1, β_1. These results confirm the pertinence of using a linear specification and support the choice of Model 1 as the preferred specification.

The second robustness check aims at corroborating the robustness of the estimated coefficients translating the existence and magnitude of ESDA in our sample. The method chosen for this second robustness check is a regression averaging approach. In the database used to estimate Models 1-3, each of the machine learning models appears in 17 to 18 observations, as each machine learning model is run for 20 different percentages of total predictors but values with a less than 10% of total predictors are excluded to avoid heteroscedasticity. Therefore, the resulting dataset has 864 observations. In the second robustness test, instead of using this dataset, we create one dataset per machine learning model, each containing 17 to 18 observations (i.e., one observation per value of the variable $VarPerc$). This results in 48 datasets containing 17 to 18 observations each. For each of these 48 datasets, we estimate the following simple linear regression:

$$log(score) = \gamma_0 + \gamma_1 log(VarPerc) + \varepsilon_i \tag{14}$$

Where γ_i are estimated coefficients. Then, we average the 96 estimated coefficients γ_1 in order to obtain an estimation of ESDA in our sample. The resulting average, 0.085, is slightly below the range of estimated coefficients for $Log(VarPerc)$ in Models 1-3 (0.087 to 0.132). Note that in this second robustness check, machine learning model characteristics are only indirectly taken into account through the averaging of the 48 estimated coefficients. Moreover, the number of observations per regression is of only 17 to 18. Despite the limitations of this more rudimentary methodology, the fact that the results

fall close to the estimations made using Models 1-3 and the non-parametric kernel regression from the first robustness check corroborates the robustness of our estimation of ESDA. Descriptive statistics of the estimated parameters in the 48 models are shown in Appendix B.

6 Conclusions

There is considerable controversy and confusion about economies of scale and scope in data. While economies of scale may be intuitively clear, the interpretation of economies of scope when applied to data is ambiguous. We introduce a conceptual distinction between the traditional interpretation of economies of scope in the re-use of data and a new interpretation of ESDA. The existence of economies of scope in data re-use constitutes an argument in favour of lowering access barriers and wider sharing and diffusion of data. ESDA put a premium on concentration of data in large pools.

The existing confusion about economies of scale and scope in data in the economic literature is not only due to conceptual ambiguities but also to the absence of empirical evidence on the existence of ESDA. In the second part of this paper, we present empirical evidence in support of it. We explore the existence of economies of scope in the aggregation of health and socio-economic datasets in terms of their contribution to the prediction accuracy of health outcomes. Our findings confirm that the aggregation of larger, more complementary sets of predictor variables significantly increase the prediction accuracy of health outcomes. These findings constitute an argument in favour of opening up health data silos and merging them with socio-economic data sources into large multi-domain data pools to produce better predictive and preventive health care outcomes. Companies with access to large and diverse consumer data could contribute to this by merging consumer behaviour data with personal health data. Since health data are particularly sensitive from a personal data protection point of view, such wide-ranging data pools should be subject to strict data protection procedures.

More generally, the existence of economies of scope in the aggregation of complementary data constitutes an argument in favour of creating large data pools comprising a wide variety of data sources. For example, the European Data Strategy policy initiative seeks to create sectoral data pools in several industrial and services sectors (European Commission, 2020). This may be social welfare enhancing when it contributes to the extraction of more valuable insights from aggregated data pools, compared to fragmented datasets. At the same time, it triggers a debate on an equitable distribution of this additional social value between the private contributors to these data pools. When large aggregated data pools remain under exclusive private control, they may result in monopolistic behaviour, information asymmetries and equity concerns that undermine the social welfare benefits that they could potentially generate. This risk is enhanced by increasing returns to scope in data aggregation (for which we find evidence), as they can constitute a barrier to entry (Furman, 2018). However, data pooling to achieve economies of scale and scope in aggregation does not necessarily imply centralizing data control rights. Large data pools with non-exclusive use rights may provide a solution for this conundrum. As such, data policies remain a difficult balancing act between the welfare gains from large datasets

and the welfare costs of monopolistic behaviour by the operators of these pools (Cabral et al., 2021).

References

Agrawal, A., Gans, J., and Goldfarb, A. (2018). *Prediction machines: the simple economics of artificial intelligence*. Harvard Business Press.

Antuca, A. and Noble, R. (2021). Data: how it affects competitive dynamics, how to value it, and whether to provide third-party access to it. *Competition Law Journal*, 20(2):102–110.

Bajari, P., Chernozhukov, V., Hortaçsu, A., and Suzuki, J. (2019). The impact of big data on firm performance: An empirical investigation. In *AEA Papers and Proceedings*, volume 109, pages 33–37.

Banalieva, E. R. and Dhanaraj, C. (2019). Internalization theory for the digital economy. *Journal of International Business Studies*, 50(8):1372–1387.

Beğenilmiş, E. and Uskudarli, S. (2018). Organized behavior classification of tweet sets using supervised learning methods. In *Proceedings of the 8th International Conference on Web Intelligence, Mining and Semantics*, pages 1–9.

Bergemann, D. and Bonatti, A. (2019). Markets for information: An introduction. *Annual Review of Economics*, 11:85–107.

Blackwell, D. (1953). Equivalent comparisons of experiments. *The annals of mathematical statistics*, pages 265–272.

Cabral, L., Haucap, J., Parker, G., Petropoulos, G., Valletti, T. M., and Van Alstyne, M. W. (2021). The eu digital markets act: a report from a panel of economic experts. *Cabral, L., Haucap, J., Parker, G., Petropoulos, G., Valletti, T., and Van Alstyne, M., The EU Digital Markets Act, Publications Office of the European Union, Luxembourg*.

Calzolari, G., Cheysson, A., and Rovatti, R. (2022). Machine data: market and analytics. *Working Paper*.

Carballa-Smichowski, B. (2018). The value of data: an analysis of closed-urban-data-based and open-data-based business models.

Chiou, L. and Tucker, C. (2017). Search engines and data retention: Implications for privacy and antitrust. Technical report, National Bureau of Economic Research.

Claussen, J., Peukert, C., and Sen, A. (2019). The editor vs. the algorithm: Targeting, data and externalities in online news. *Data and Externalities in Online News (June 5, 2019)*.

Colbaugh, R. and Glass, K. (2017). Learning about individuals' health from aggregate data. In *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 3106–3109. IEEE.

Couronné, R., Probst, P., and Boulesteix, A.-L. (2018). Random forest versus logistic regression: a large-scale benchmark experiment. *BMC bioinformatics*, 19(1):1–14.

Crémer, J. (1982). A simple proof of blackwell's "comparison of experiments" theorem. *Journal of Economic Theory*, 27(2):439–443.

Crémer, J., de Montjoye, Y.-A., and Schweitzer, H. (2019). Competition policy for the digital era. *Report for the European Commission*.

De Giorgi, G., Harding, M., and Vasconcelos, G. F. (2021). Predicting mortality from credit reports. *Financial Planning Review*, 4(4):e1135.

Economides, N. and Lianos, I. (2021). Restrictions on privacy and exploitation in the digital economy: a market failure perspective. *Journal of Competition Law & Economics*, 17(4):765–847.

European Commission (2020). Communication from the commission to the european parliament, the council, the european economic and social committee and the committee of the regions: A european strategy for data.

Furman, J. (2018). Competition and consumer protection in the 21st century. Federal Trade Commission hearings.

Furman, J., Coyle, D., Fletcher, A., McAuley, D., and Marsden, P. (2019). Unlocking digital competition: Report of the digital competition expert panel. *UK government publication, HM Treasury*, 27.

Graef, I. and Prüfer, J. (2021). Governance of data sharing: A law & economics proposal. *Research Policy*, 50(9):104330.

Jones, C. I. and Tonetti, C. (2020). Nonrivalry and the economics of data. *American Economic Review*, 110(9):2819–58.

Junqué de Fortuny, E., Martens, D., and Provost, F. (2013). Predictive modeling with big data: is bigger really better? *Big data*, 1(4):215–226.

Klein, T. J., Kurmangaliyeva, M., Prüfer, J., Prüfer, P., and Park, N. N. (2022). How important are user-generated data for search result quality? experimental evidence. Technical report, TILEC Discussion Paper No.

Krämer, J. (2021). Personal data portability in the platform economy: Economic implications and policy recommendations. *Journal of Competition Law & Economics*, 17(2):263–308.

Krämer, J., Schnurr, D., and Micova, S. B. (2020). *The role of data for digital markets contestability: case studies and data access remedies*. Centre on Regulation in Europe asbl (CERRE).

Larrabee, G. J. (2008). Aggregation across multiple indicators improves the detection of malingering: Relationship to likelihood ratios. *The Clinical Neuropsychologist*, 22(4):666–679.

Lee, G. and Wright, J. (2021). Recommender systems and the value of user data. *National University of Singapore Working Paper, https://836d83c2-c629-40a4-9ca7-b1bf324d720d. filesusr. com/ugd/c6ffe6e391febde31845a08bc1ef6969182398. pdf (June2021)*.

Martens, B., De Streel, A., Graef, I., Tombal, T., and Duch-Brown, N. (2020). Business-to-business data sharing: An economic and legal analysis. *EU Science Hub*.

Martens, B. and Zhao, B. (2021). Data access and regime competition: A case study of car data sharing in china. *Big Data & Society*, 8(2):20539517211046374.

Mayer-Schönberger, V. and Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think*. Houghton Mifflin Harcourt.

McAfee, P., Rao, J., Kannan, A., He, D., Qin, T., and Liu, T. (2015). Measuring scale economies in search. *Microsoft slides*.

Neumann, N., Tucker, C. E., and Whitfield, T. (2018). How effective is black-box digital consumer profiling and audience delivery?: Evidence from field studies. *Social Science Research Network Working Paper Series*.

Palfrey, J. and Gasser, U. (2012). *Interop: The promise and perils of highly interconnected systems*. Basic Books.

Panzar, J. C. and Willig, R. D. (1981). Economies of scope. *The American Economic Review*, 71(2):268–272.

Refaeilzadeh, P., Tang, L., and Liu, H. (2009). Cross-validation. *Encyclopedia of database systems*, 5:532–538.

Reimsbach-Kounatze, C. (2016). Maximizing the economic and social value of data, understanding the benefits and challenges of enhanced data access. Technical report, OECD, Directorate for Science and Technology, Committee on Digital Economic Policy.

Rosen, S. (1983). Specialisation and human capital. *Journal of Labor Economics*, 1:43–49.

Schäfer, M. and Sapi, G. (2022). Complementarities in learning from data: Insights from general search. Technical report, Available at: https://drive.google.com/file/d/1RRxhTW560PwtMGLEN-0wHikW7oVS9CEn/view.

Schäfer, M., Sapi, G., and Lorincz, S. (2018). The effect of big data on recommendation quality: The example of internet search.

Schouten, B. (2018). Statistical inference based on randomly generated auxiliary variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(1):33–56.

Scott-Morton, F., Bouvier, P., Ezrachi, A., Jullien, B., Katz, R., Kimmelman, G., Douglas-Melamed, A., and Morgenstern, J. (2019). Report of the committee for the study of digital platforms. market structure and antitrust subcommittee. Technical report, George J. Stigler Center for the Study of the Economy and the State.

Teece, D. J. (1980). Economies of scope and the scope of the enterprise. *Journal of economic behavior & organization*, 1(3):223–247.

Teece, D. J. (1982). Towards an economic theory of the multiproduct firm. *Journal of Economic Behavior & Organization*, 3(1):39–63.

Tucker, C. (2019). Digital data, platforms and the usual [antitrust] suspects: Network effects, switching costs, essential facility. *Review of Industrial Organization*, 54(4):683–694.