Evaluating Treatment Effects Using Data Envelopment Analysis on Matched Samples

An Analysis of Electronic Information Sharing and Firm Performance

Bogetoft, Peter; Kromann, Lene

Evaluating treatment effects using Data Envelopment Analysis on matched samples: An analysis of electronic information sharing and firm performance

Peter Bogetoft , Lene Kromann

PII: S0377-2217(18)30222-4
DOI: 10.1016/j.ejor.2018.03.013
Reference: EOR 15032

To appear in: *European Journal of Operational Research*

Received date: 29 April 2016
Revised date: 6 March 2018
Accepted date: 8 March 2018

Please cite this article as: Peter Bogetoft , Lene Kromann , Evaluating treatment effects using Data Envelopment Analysis on matched samples: An analysis of electronic information sharing and firm performance, *European Journal of Operational Research* (2018), doi: 10.1016/j.ejor.2018.03.013

Highlights

- Estimate treatment effects on firm data
- When firms chose the treatment, the causality is unclear
- Matching before Data Envelopment Analysis can ensure sub-sample homogeneity
- Superior to standard tools like randomized sub-sampling and second stage analysis
- Applied to estimating the gains from electronic data sharing in supply chains

Evaluating treatment effects using Data Envelopment Analysis on matched samples: An analysis of electronic information sharing and firm performance.

Peter Bogetoft[a]
Lene Kromann[b]

Author affiliation
[a]*Department of Economics, Copenhagen Business School, 2000 Frederiksberg, Denmark, pb.eco@cbs.dk*
[b]*DAN Management and Organizational Studies, Faculty of Social Science, University of Western Ontario, London, Canada, lkromann@uwo.ca*

Corresponding author: Peter Bogetoft**,** Department of Economics, Copenhagen Business School, 2000 Frederiksberg, Denmark, pb.eco@cbs.dk, Phone (+45) 3815 2506.

Abstract

An intuitively obvious approach to evaluating the effects of a new business model is to compare the performance of firms using the business model (the treatment group) with the performance of a similar group of firms that do not use the business model (the control group). Data Envelopment Analysis (DEA) can be a powerful tool in such comparisons because it allows us to estimate changes in average performance as well as in frontier performance. In this paper, we suggest using matching together with DEA as a way to ensure sub-sample homogeneity. The advantages of using a matched sample compared to a random sample of non-treated firms to remove sample size bias is documented using a simulation study. A real-world application is also provided. In the application, we study how information sharing has impacted the performance of Danish manufacturing firms. We match firms that use electronic information sharing to their "twin" firms that do not on the basis of firm characteristics. Before matching, there is a considerable difference in performance between the two groups. However, after matching, we can conclude that approximately 50% of the difference is the result of selection bias.

Keywords: Data Envelopment Analysis, Bias, Matching, Propensity Score

Financial support

[1]Financial support from the Danish Industry Foundation and the Danish Council for Strategic Research is gratefully acknowledged.

1. Introduction

When firms change business practices by, for example, introducing *electronic information sharing*[1] with business partners, we are interested in knowing the impact of such changes. First, we might be interested in the average impact because it may give us an idea of the potential industry-wide benefits of such changes. Second, we might be interested in knowing whether the new practices make performances converge or diverge, i.e., whether these practices primarily help weak companies catch up or only benefit strong companies. Third, we might be interested in knowing how efficient companies are affected, because these effects not only suggest how the very best can benefit but also indicate how less-efficient firms might benefit in the years to come.

Data envelopment analysis (DEA) has been proven a useful tool in evaluating firm performance (e.g., Bogetoft, 2012; Bogetoft & Otto, 2011). By comparing the DEA-based performances of firms before and after a change or treatment or by comparing the performances of a group of firms that change business practices with a group of firms that do not, several authors – e.g., Chen and Ali (2004), Abri and Mahmoudzadeh (2015) and Giokas, Eriotis, and Dokas (2015) – have gained insight into the effects of different changes. Unfortunately, such comparisons present some difficulties, partly because of the bias of DEA models. The DEA approach makes an inner approximation of the production possibilities and compares actual performances against this approximation. The evaluated performances are therefore biased upward. Firms appear more efficient than they really are because only relative efficiency is measured. Moreover, this bias is not constant. It depends on the size of the dataset, the number of inputs and outputs, and the firms' location in production space. In the parts of the production space where the density of observations is high, the bias is smaller. In turn, when we compare the performances of two groups of firms, their apparent differences may be differences in bias rather than in performance.

To unequivocally identify the impact of a new business practice, we would ideally conduct a randomized experiment in which some firms are treated with the new business practice and others are not. If the treatment and non-treatment samples are large enough, the differences in the performance mean, performance spread and performance frontier of the two groups can then be linked to the treatment. The randomization of firms into different groups guarantees that, on average, no systematic differences in observed or unobserved covariates will be found between the units assigned to the different groups. Unfortunately, randomized experiments are rare in social sciences, particularly when analyzing firms and when the treatments are potentially important. Instead,

[1] Electronic information sharing concerns the information that is provided to the supplier and the customer upon the delivery of goods/services, e.g., coordinating inventory, demand, production, distribution, and delivery status.

non-randomized studies are available. The treated firms are selected using a non-random mechanism, and firms have even been self-selected into the treatment group. In such cases, direct comparisons of the outcomes across the two groups may be misleading because of selection bias. The performance differences may be linked to the factors that determine treatment in the first place rather than to the treatment itself.

The difficulty of selection bias may be partially avoided if characteristics of the groups are incorporated into the study through matched sampling. Matching creates a "quasi-randomized" experiment by matching the two groups based on observable information. For instance, matching firms from the same industry is a simple way to perform matched sampling. Hence, the fact that treatment is more common in some industries than in others, which may cause us to mistake an industry effect for a treatment effect, can be counteracted by the choice of non-treated firms from the same industries. Matching several variables is more complex, and a propensity score is used instead. The propensity score is calculated for each firm in the two groups, and the best match is found. Of course, this approach is not as good as pure randomization, as we can still only match on observed characteristics. However, if many of the characteristics related to the treatment assignment are observed, one can be fairly confident that comparisons of the outcomes across the two groups will reflect treatment effects. Matching has already been used extensively—for example, in accounting, finance, and marketing research—to compare variables of interest across two samples. The performance of the matched control sample serves as a benchmark and helps remove the confounding effects of extraneous variables and market forces that may influence firm performance. For instance, Bharadwaj (2000) used matching in his analysis of information technology and firm performance.

However, to the best of our knowledge, no study combines matching with DEA. We will do so in this paper. We will show how this combination allows us to balance the comparisons, such that the bias in DEA is not confounded with the effects of treatment. To illustrate our approach, we evaluate how electronic information sharing with business partners has an impact on Danish manufacturing firms' performances. The link between the information system (IS) and firm performance has received mixed results (Chen et al, 2010), with two recent studies documenting substantial gains (Prajogo & Olhager, 2012; Leidner et al, 2012).

This paper is organized as follows. In Section 2, we briefly discuss DEA-based performance evaluations, including the associated bias problem. We also discuss the estimation of a treatment effect using a Malquist-like procedure, and we introduce the idea of matching. In Section 3, we undertake a simulation study showing the advantages of combining DEA and matching. In Section 4, we use matching and DEA to evaluate the impact information sharing has on the performance of Danish manufacturing. Conclusions are provided in Section 5.

2. Method

In this section, we explain the DEA approach with an emphasis on the bias problem. Also, we discuss how to measure a treatment effect, and we introduce matching and discuss how this can be used to separate the treatment effects and the bias effect.

2.1 DEA-based performance evaluations

The DEA approach was originally proposed by Charnes, Cooper, and Rhodes (1978). This seminal contribution was subsequently refined and applied in a rapidly increasing number of papers. In his 1992 bibliography, Seiford (1994) lists no fewer than 472 relevant published articles and Ph.D. theses. In his 2002 bibliography, Tavaras (2002) includes more than 3,000 contributions. DEA is now a well-established set of methods that is also described in several textbooks, including Bogetoft (2012), Bogetoft and Otto (2011), Coelli, Rao, and Battese (1998), and Cooper, Seiford, and Tone (2007).

To formalize, consider a case where each of *n* decision-making units (DMUs), $i \in I = \{1,2,\dots,n\}$, transforms *p* inputs *x* into *q* outputs *y*. Additionally, let *T* be the production possibility set:

$$T = \{(x,y)|x \text{ can produce } y\}$$

DEA approaches estimate *T* from the observed productions $(x^i, y^i), i \in I$, and they evaluate the efficiency of the DMUs relative to the estimated technology. The estimate of *T*, the empirical reference technology *T**, is constructed according to the so-called *Minimal Extrapolation Principle*: *T** is the smallest subset of R_0^{p+q} that contains the actual production plans $(x^i, y^i), i \in I$ and satisfies certain technological assumptions, the most common of which are explained shortly. Note that we use R_0 to denote the set of non-negative reals.

Based on a technology *T* (or an estimate thereof, *T**), the inefficiency (or relative inefficiency) of a given *DMU*, *DMU*^{i}, reflects the potential reduction of inputs and the expansion of outputs. It is most commonly measured using the so-called Farrell (1957) measures as follows:

$$E^i = \min\{E \in R_0 | (Ex^i, y^i) \in T\} \text{ or } F^i = \max\{F \in R_0 | (x^i F, y^i) \in T\}$$

where E^i is the minimum contraction of all the inputs, and F^i is the maximum expansion of all the outputs that

are feasible in *T*.

Different DEA approaches are distinguished by the technological regularities they impose on *T* and, thereby, on *T**. Classical assumptions are free disposability (A1), i.e., $(x, y) \in T$ and $x' \geq x$ and $y' \leq y \Rightarrow (x', y') \in T$; convexity (A2), i.e., T is convex; and s-returns to scale (A3(s)), i.e., $(x, y) \in T \Rightarrow k(x, y) \in T$ for $k \in K(s)$, where $s = crs, drs, vrs, or\ irs$, which correspond to constant, decreasing, varying, or increasing (returns to scale), respectively, and where $K(crs) = R_0$, $K(drs) = [0,1], K(vrs) = \{1\}$, $and\ K(irs) = [1, +\infty)$, respectively. Assuming A1, A2 and A3(s), the empirical, minimal extrapolation principle reference technology, $T^*(s)$, based on observations $(x^i, y^i), i \in I$, can be easily expressed as follows:

$$T^*(s) = \left\{(x, y) \in R_0^{p+q} | \exists \lambda \in R_0^n : x \geq \sum_i \lambda^i x^i , y \leq \sum_i \lambda^i y^i , \lambda \in \Lambda(s)\right\},$$

where $\Lambda(crs) = R_0^n$, $\Lambda(drs) = \left\{\lambda \in R_0^n \,\middle|\, \sum_i \lambda^i \leq 1\right\}, \Lambda(irs) = \left\{\lambda \in R_0^n \,\middle|\, \sum_i \lambda^i \geq 1\right\}$ and $\Lambda(vrs) = \{\lambda \in R_0^n \mid \sum_i \lambda^i = 1\}$. Assumptions A1, A2, and A3 have been relaxed, for example, in the free disposability hull (fdh) model used by Deprins, Simar, & Tulkens (1984). It invokes only A1, and *T*(fdh)* has the structure above, with $\Lambda(fdh) = \{\lambda \in R_0^n \mid \sum_i \lambda^i = 1, \lambda^i \in \{0, 1\}\ \forall i\}$.

2.2 Treatment effects

Assume that instead of one group of firms, *T,* we now have two groups of firms, *S* and *T*. The two sets may contain the same firms in an earlier period, *S*, and in a later period, *T*, or they may contain different firms, a control group, *S*, and a group that is subject to treatment, *T*. In such situations, we are interested in understanding how the frontier shifts and how the firms' performances relative to the frontier differ in the two groups. We can measure the shifts and differences in different ways, but we will introduce only a few basic measures here.

Let

$$E(s,t) = \text{The Farell input efficiency of firm } j \text{ in group } S \text{ against the frontier of group } T.$$

More specifically, using s-returns to scale, we have

$$E^j(s,t) = \min\left\{E \in R_0 | \exists \lambda \in R_0^n : x^{sj} \geq \sum_{i \in I} \lambda^i x^{ti}, y^{sj} \leq \sum_{i \in I} \lambda^i y^{ti}, \lambda \in \Lambda(s)\right\},$$

where x^{sj} is the input of firm *j* in period *s* or group *s*, y^{sj} is the output of firm *j* in period *s* or group *s,* x^{ti} is the input of firm *i* in period *t* or group *t*, and y^{ti} is the output of firm *i* in period *t* or group *t.*

Likewise, let $E^i(t,s)$ be a measure of the performance of firm *i* in group *T* against the frontier of firms in group *S*.

The frontier change, *FC,* from group *S* to group *T* can then be measured as follows:

$$FC^i(s) = \frac{E^i(s,s)}{E^i(s,t)} \text{ or } FC^i(t) = \frac{E^i(t,s)}{E^i(t,t)}, \quad (1)$$

depending on where we measure the frontier shift—in the directions where the *S* sample is located or in the directions where the *T* sample is located. To consolidate the two measures, we might take geometric averages, as in the Malmquist index (cf. Caves, Christensen, & Diewert, 1982; Färe, Grosskopf, Lindgren, & Ross, 1994; Malmquist, 1953). We can also measure improvement against the best practice frontier from group S to group T as the catch-up or efficiency change:

$$CU^i = \frac{E^i(t,t)}{E^i(s,s)}.$$

However, catch-up makes the best conceptual sense when the same firm is measured in the two samples. In all cases, ratios above 1 suggest improvement—either extending the feasible production set ($FC^i(s) > 1$ and $FC^i(t) > 1$) or catching up to the best practice ($CU^i > 1$).

2.3 The bias problem in DEA

By using the minimal extrapolation principle and assuming no noise in the data, DEA provides an inner approximation of the underlying production possibility set. Therefore, the potential input savings and output expansions are underestimated. That is, the input efficiency estimate E^* of E is biased upwards, $E^* \geq E$; similarly, the output efficiency estimate F^* of F is biased downward, $F^* \leq F$, cf. Bogetoft and Otto (2011).

Unfortunately, there is only limited theoretical insight into the size of the bias. Except for the one input one output case analyzed by Simar and Wilson (2000), the extent of the bias must be evaluated using numerical methods. Still, some insight is available. The larger the sample size, the more points we have to span the frontier and the more likely it is that there will be high-performance observations in our sample, i.e., the bias is smaller. With more input and output dimensions, the observations are most likely less comparable, and therefore, the bias is larger. Indeed, the rate of convergence of the efficiency estimates is inversely related to the number of parameters, cf. Kneip et al. (1998), and simulations suggest that we need to double the number of observations for each extra parameter to maintain the bias, cf. Pedraja-Chaparro et al. (1999). Moreover, the more curved the frontier is, the less powerful the linear approximation is between observations and the more biased the frontier estimates are, cf. Simar and Wilson (2000). Lastly, these factors should ideally be looked at locally. Therefore, the higher the density, i.e., the more observations we have, in a given part of the production space, the smaller the bias is here. The role of some of these factors has also been discussed and illustrated in Kittelsen (1995), Gijbels et al. (1999), Cubbin (2004), and Chumpitaz (2010).

When comparing performances in two groups of firms, Zhang and Bartels (1998) suggested using a random sample with a sample size that is equal to that of the smaller group to ensure that the biases in the two groups are of similar magnitude. More specifically, their random sampling approach works as follows: They draw a random sample of firms from a large group without replacement. The size of the subsample equals the number of firms in the small group. They then analyze the subsample using DEA and repeat the procedure (for an unspecified number of times). Taking the average of these results, they obtain a sample size-adjusted mean efficiency for the larger group, which can be compared to the mean efficiency for the smaller group. Of course, as Chumpitaz (2010) notes, the gain in comparability results in lower overall precision in the estimates because only a sub-sample of the large group is utilized. Instead of sub-sampling to create similar biases, bootstrapping can be used to approximate the distribution of the efficiency score. In particular, bootstrapping can be used to correct for the bias of the efficiency estimators and to estimate confidence intervals for bias-corrected efficiency measures, as suggested in Simar and Wilson (2000).

2.4 Selection bias

When we compare firms from different groups, the bias resulting from the use of an inner approximation is not the only bias about which we should be concerned. We should also consider selection bias. The firms in the treatment group may be different from those in the non-treatment group; as such, their performances may reflect

these differences rather than the treatment effects. This selection bias makes drawing causal conclusions difficult.

To avoid selection bias, we ideally want a controlled experiment where the treated and non-treated firms only differ in terms of the treatment. Alternatively, we would like a randomized experiment where firms are randomly assigned to the experimental and control groups to minimize the variability in firm characteristics across the groups. Providing a sufficient number of firms that have been randomized, the balance in observed and unobserved characteristics between the groups enables unbiased conclusions about the treatment effect. Unfortunately, controlled or randomized experiments across firms are very rare. Instead, observational studies are often conducted where a potentially large systematic difference in observed firm characteristics across groups exists—and, in turn, the risk of selection bias.

In the econometric literature, matching is used as a remedy for selection bias. The idea is to construct a non-treatment counterfactual group that matches the treated firms in a series of covariates. Ideally, all the relevant differences in the outcomes of the control and treatment groups are captured by their observed characteristics (conditional independence). As such, conditioning on the observables, z, both observed and unobserved differences between the treatment and control groups are eliminated, thereby essentially simulating random assignments.

If the number of covariates, z, is large, matching each characteristic is difficult, and the chances of finding an exact match is reduced. A way to circumvent this curse of dimensionality is to match according to the estimated probability of treatment or the propensity score (Rosenbaum & Rubin, 1983). Letting the binary variable d indicate whether a firm is treated ($d=1$) or not ($d=0$), the propensity score, $p(z)$, is then the conditional probability:

$$p(z) = \Pr[d = 1|Z = z],$$

where, given the z variables, the d observations are assumed to be independent. We use logistic regression to model the treatment choice. Logistic regression is a good method to reduce the multiple variables in z to a single measure when there are two identifiable groups to be compared. The propensity score is calculated as follows:

$$p(z) = \frac{exp(a + b_1 z_1 + b_2 z_2 + b_3 z_3 + \cdots)}{1 + exp(a + b_1 z_1 + b_2 z_2 + b_3 z_3 + \cdots)},$$

where a is a constant, z_i is a covariate, and b_i is the regression coefficient for covariate i.

Firms in the treatment group and firms in the control group can be matched in several ways with an estimated propensity score. Here, we will use so-called nearest-neighbor matching without replacement as the matching method. For each treated unit, we match the non-treated unit that has the closest value for the propensity score. After matching, it is useful to test whether the two groups are balanced with respect to the covariates. To do so, we use an equality of means test in the treated and control groups. This test is similar to the absolute standardized difference test used by, for instance, Smith and Todd (2005) and recommended by Heinze and Jüni (2011).

The combination of one-to-one matching and DEA will ensure that the bias due to difference in sample size is removed. Moreover, it will reduce the selection bias that results from the data not being produced through a random experiment.

Instead of matching to cope with selection, one can potentially apply regression-based methods. A technique that is common in the DEA literature that involves making second-stage adjustments; see, for instance, Bogetoft and Otto (2011). Initially, the differences between the groups are ignored. The observations are pooled, and performance is measured against a common frontier. Next, a second-stage analysis is undertaken to determine whether performance differences may be related to other firm characteristics. It is common to use a modified right-censored Tobit regression of the efficiency scores from the full sample on the firm characteristics. If a systematic relationship exists between efficiency E and firm characteristic z:

$$E^i = f(z^i) + \varepsilon^i,$$

we can transform the efficiencies to obtain the adjusted efficiency score $E^{Adj;i}$, which is calculated, for example, as:

$$E^{Adj;i} = \frac{E^i / f(z^i)}{\max_j E^j / f(z^i)}.$$

If z is different in the treatment and control groups, the second-stage corrections can remove the impact of z, thereby leading to a cleaner "risk-adjusted" measure of the treatment effect. We can then test whether the average

corrected efficiencies are different in the treatment and control groups. Although such corrections may be useful, this analysis only provides us with one number—the average treatment effect—whereas by correcting for selection bias using matching before running the DEA model, we obtain the entire distribution of individual firm effects. Using matching, we explicitly allow for the fact that the frontier may change when the treatment is introduced. The change in frontier may not be uniform. For some combinations of inputs, for example, the treatment may push the frontier out to a large extent, while for other input combinations, the impact may be smaller or even have a different sign, i.e., lead to a contraction of the set of feasible outputs. Moreover, the second stage approach may not remove the selection bias. The treatment variable may be correlated with the error-term in the second stage regression, since there may be omitted confounding factors. The second stage analysis should therefore ideally involve an instrument variable analysis to identify the pure treatment effect. Finding a good instrument, i.e., one that is sufficiently correlated with the treatment variable but otherwise uncorrelated with the efficiency level, is unfortunately often difficult. This also speaks in favor of the matching approach.

3. Simulation experiment

To demonstrate the potential value of the matching approach, we will now use simulations. We will show that by using a matched sample instead of a random sample in a DEA analysis with two groups of different sizes, we can reduce selection bias.

The simulation study uses real data for the input variables of our case below. There are three inputs, which are number of employees, capital and material. Output is sales revenue, but the output will be simulated to ensure we have control over the simulations, i.e., that we know the underlying truth. The treatment concerns the use of electronic information sharing with business partners. The digitalization dummy is denoted: dig_{dum}. This treatment is more likely to be applied among large firms, which we will here identify using a size dummy: $size_{dum}$, where $size_{dum}$=1 if fte >100. Also, the large firms are generally assumed to be more efficient.

We use the actual observations of inputs and outputs of all firms to determine a DEA technology T^* assuming constant returns to scale, crs. The crs assumption is the same assumption as is used in the empirical analysis below. It is here supported by data, it makes conceptual sense, and it comes with the advantage of making all linear programming problems feasible. We construct the observations we use in the simulations as follows. Firm i with inputs x^i is initially assumed to have an output of 1. We can now determine the output efficiency F^i of this firm according to the technology T^*. It follows that (x^i, F^i) will be a frontier point of the underlying simulation

technology T^*. Now, the actual output we assign to the firm with inputs x^i depends on the size of the firm and the automation assumed:

$$y^{Sim\ i} = F^i\left(1 + 0.3\ \ size_{dum}{}^i\right)\left(1 + 0.2\ dig_{dum}{}^i\right).$$

We thus assume that there is a size effect of 30% and a digitalization effect of 20%. Being a large firm using electronic information sharing boosts the output by 1-(1+0.3) (1+0.2)=56%.

In our actual data introduced in the next section, 26% of firms use electronic information sharing. If we examine the relation in more detail (see Table 1), we see that small firms use electronic information sharing with a probability of 17%, whereas large firms use electronic information sharing with a probability of 38%. It therefore seems important to control for firm size when comparing the performance of firms that use and do not use electronic information sharing. If we do not control for firm size, we will tend to exaggerate the effects of digitalization by including a size effect. In the same way, it can be important to control, e.g., for industry and export rate, but to keep things simple here, we will focus on one variable, firm size.

	Small firms (size_dum=0)	Large firms (size_dum=1)	Sum	Share of large
Non-Digitalized (dig_dum=0)	511	304	815	0.37
Digitalized (dig_dum=1)	103	184	287	0.64
Sum	614	488	1102	0.44
Share of digitalized	0.17	0.38	0.26	

Table 1 Distribution of firms

Source: Register data from Statistics Denmark

Examining the simulation equation, it is clear that if a random sample is used to ensure the same sample size in the two groups, the firms that digitalize will tend to have higher sales, as more of the firms are large. However, if we match on firm size to ensure that the matched sample has the same size distribution as the group of digitalized firms, then the only difference between the two groups should be whether they are digitalized or not, and hence, we have removed selection bias. This is illustrated in Table 2 below.

T data (digitalized)	103 small firms	
S data (non-digitalized)	103 random firms	103 matched firms
Simulations	500	500
Minimum	0.673	0.701
1st quartile	0.923	1.196
Median	**0.924**	**1.200**
Mean	0.937	1.182
3rd quartile	0.939	1.200
Maximum	1.152	1.200
StdDev for mean	0.01	0.006

Table 2 Simulation study of frontier shift

Notes: In column 1, 103 of the 815 non-digitalized small and large firms are used as a random sample in each of the 500 DEAs. In column 2, we use matching including 2 size dummies: the first takes the value 1 if firms have 21-50 employees, and the second take the value 1 if firms have 51-100 employees. The base group is firms with up to 20 employees. In this simple matching exercise, we are able to find a perfect match for each of the firms in the treatment group.

Source: Register data from Statistics Denmark

Table 2 shows an example where we have compared the outcome of 103 small digitalized firms against a sample of non-digitalized firms. The reason for only including the small digitalized firms is to simplify the example. In the first column, we compare the outcomes to a random sample of equal size to get rid of the sample size bias. We repeated the exercise 500 times, and we see that the average effect of digitalization is to lower the output by 7.6% if we focus on the median effect. If we instead use a sample of firms that are matched on the size variable, we obtain an increase in output as the result of a digitalization of 20%, as shown in the last column. The selection bias, in other words, give a very biased estimate of the true underlying digitalization effect of 20% if we use a random sample and thereby eliminate the sample size effect. If we match on size, however, we obtain the correct estimate of the digitalization effect.

Of course, the example above is extreme, and this is why the mistake resulting from a pure randomization is so extreme. What is an advantage from digitalization of 20% is estimated as a disadvantage of 7.6%. We have chosen this example, however, since it allows us to illustrate the risk of not matching, and it allows us to understand the outcome very easily, thereby also ensuring that the simulations have been properly implemented. In the rest of this section, we give an intuitive explanation of the results.

We are investigating frontier shifts as measured by:

$$FC^i(s) = \frac{E^i(s,s)}{E^i(s,t)},$$

but since we are using a constant returns to scale model, the input efficiencies are the inverse of the output efficiencies. Therefore, we have:

$$FC^i(s) = \frac{F^i(s,t)}{F^i(s,s)}.$$

Also, for simplicity, let us normalise the output level of a small non-digitalized firm to 1. The large non-digitalized firms will then have outputs of approximately 1.3 for the same input combination, while a similar small but digitalized firm will have an output of 1.2. We can therefore calculate the output efficiency of a firm from the S sample of non-digitalized firms as illustrated in the following table.

	Small firms	Large firms	Weighted aveage	Index
Share i group of non-digitalized	*511/815*= 0.627	*304/815*= 0.373		
Average output non-dig firms	1.000	1.300		
Output eff of non-dig compared to T sample of small dig firms	1.200	*1.2/1.3*= 0.923	1.097	F(S,T)
Output eff of firms compared to random non-dig firm	1.300	1.000	1.188	F(S,S)
Frontier shift			*1.097/1.18*= 0.923	*F(S,T)/F(S,S)*= FC

Table 3 intuitive explanation of the results in Table 2

Weighting together the share of large and small non-digitalized firms, we obtain an average efficiency of S firms compared to T firms of 1.097. Now, inside the group of non-digitalized firms, the large firms will set the output norm, and therefore, the output efficiency of a small non-digitalized firm will be 1.3, since the size effect is 30%. When we only compare inside the non-digitalized group, we therefore obtain an average output efficiency of 1.188. We see, therefore, that the frontier shift we obtain in the randomized approach is 1.097/1.188=0.923.

4. The effects of electronic information sharing with business partners

Electronic information sharing can be broadly defined as linking information technology (IT), such as computer-to-computer exchange of routine information, across firms. To improve performance, firms need to both emphasize technology investment and choose the appropriate information to share (Zhou et al (2007)). Research on the effects of IT investments and performance using a Cobb–Douglas production function documents

substantial gains for the average firm (Bloom, Sadun, & Van Reenen, 2012; Brynjolfsson & Hitt, 2000; Cardona, Kretschmer, & Strobel, 2013). Still, studies of the linkage between the information system (IS) and firm performance have led to mixed results (Chen et al, (2010)). Two recent studies documenting substantial gains are Prajogo and Olhager (2012) and Leidner et al (2012). Prajogo and Olhager (2012) examined 232 Australian manufacturing firms, and Leidner et al (2012) examined 263 credit unions in the US. However, as Ward (2012) emphasizes, there is a need for further study of whether an IS system makes a difference to firm performance. This study revisits the question of whether electronic information sharing makes a difference to firm performance using a DEA model. Recently, IS systems have been divided into innovative and conservative IS systems, arguing that innovative IS systems have a stronger impact on firm performance. In this study, we are unable to divide electronic information sharing into these two types of systems, which is why the effect might be even larger for the firms using the innovative system.

4.1 Data

The data for this study are collected by Statistics Denmark. The data on digitization come from a survey on the use of information and communication technology (ICT) in Danish firms. When selecting firms for inclusion in the survey, Statistics Denmark stratified all the Danish firms outside the agricultural sector according to size and industry. Strata with larger firms were overrepresented, but firms were sampled at random within each selected stratum. In 2008, 4,257 firms answered a questionnaire about their ICT use in 2007. These survey data are merged with register data on, e.g., the number of employees, the education levels of employees, and the size of the capital stock. Because the impact of digitization most likely varies considerably across industries, we chose to focus on the manufacturing sector, leaving us with 1,280 firms. When we further exclude observations that are extreme in terms of the input and output variables and observations that are missing treatment variables, we ultimately have 1,102 observations.

In the analysis, we will examine the external digitization of the information flow. External digitization is defined here as the implementation of automated information sharing with suppliers or customers regarding inventory stock, production planning, demand forecasts, or delivery status. To measure external digitization, we create a dummy that takes a value 0 if the firm is not sharing any information electronically with their business partners and a value of 1 if it is sharing information in at least one of the areas with either a customer or a supplier. Of course, we could use other definitions for externally digitized firms. For example, we could measure the degree of digitization as the number of digitized processes. However, because the digitization of the four processes seems to be strongly correlated, we use a binary definition of digitization. In our data, 287 of the 1102 firms (26

percent) are externally digitized. The remaining firms thus constitute the control group that will be selected using the randomized sampling and matching procedures.

4.2 Production model

The production model used in this study is very simple. We consider three inputs, materials, labor and capital, and one output, sales. Summary statistics for the inputs and outputs in the dataset are provided in Table 4 below.

	Full Sample		Externally digitized		Externally non-digitized	
	Mean	StDdev	Mean	StDdev	Mean	StDdev
No. of firms	1,102		287		815	
Output						
Sales	369,748	1,467,264	734,624	2,268,169	241,257	1,020,240
Input						
Capital stock	48,253	298,147	89,326	248,644	33,789	312,581
Materials	251,533	931,599	517,537	1,664,214	157,860	408,699
Employees	212	597	403	905	145	422

	Full Sample		Externally digitized		Externally non-digitized	
	Mean	StDdev	Mean	StDdev	Mean	StDdev
No. of firms	1,102		287		815	
Output						
Sales	369,748	1,467,264	734,624	2,268,169	241,257	1,020,240
Input						
Capital stock	48,253	298,147	89,326	248,644	33,789	312,581
Materials	251,533	931,599	517,537	1,664,214	157,860	408,699
Employees	212	597	403	905	145	422

Table 4 Summary statistics

Notes: Sales, material and capital stock are measured in 1000 DKK.
Source: Register data from Statistics Denmark

We see that the externally digitized firms are generally three times larger than the non-digitized firms.

4.3 Matching

Because the firms in the two groups are not randomly assigned, the groups may differ considerably in terms of their "pre-treatment" characteristics. This may seriously hamper the validity of conclusions about the treatment effects found by comparing the digitized and non-digitized subsamples directly. To create a useful matched control group, the variables included in the logistic regression should ideally be all the variables that potentially confound the treatment effect. According to Heinze and Jüni (2011), a confounder is defined by three conditions: (1) it is a covariate that is available prior to the treatment assignment; (2) it may influence the treatment decision;

and (3) it may influence the firm's outcome, in this case, the transformation of capital, materials, and employees into sales. Thus, by definition, any post-treatment measurements are excluded. Additionally, Rubin and Thomas (1996) theoretically showed and Brookhart et al (2006) demonstrated through a Monte Carlo study that all the variables that are thought to be related to the outcome, whether or not they are related to the treatment exposure, should be included in the matching analysis, as they reduce the bias and variance that results from the analysis not being a random experiment. In addition, Brookhart et al (2006) showed that the inclusion of covariates that correlate with the treatment decision, though not with the outcome, will not improve the results of a propensity analysis. Such covariates instead increase the imprecision of the treatment effect estimate.

Here, size and industry are used as the primary matching variables. They are known before digitization; they are crucial when deciding whether to invest in computer systems to facilitate electronic information sharing; and they likely influence the transformation of production factors into sales. Several other studies have used these two variables; see, for instance, Bharadwaj (2000). We expect that large firms are more likely to use electronic information sharing, as they have more orders, larger inventories and more demands to keep track of. Hence, investments in electronic information sharing are most likely to be more profitable for large firms. In the same way, the amount of information sharing with suppliers and customer differs across industries. For example, in the food industry, some firms share information with their customers (grocery stores) regarding upcoming product campaigns. Both the size and the industry are included as dummy variables. The industry variable is naturally a discrete variable. The size variable will also be more difficult to match as a continuous variable, and we do not believe that the firms must be the exact same size. Of course, we can include much more information in the matching analysis, e.g., the firm's age, the owner type, the export share, sales before introducing digitization, and educational characteristics of the workers. We will illustrate the inclusion of such information in the example below.

We used nearest-neighbor matching without replacement as the matching method. Before we performed the matching, we determined whether any of the propensity score values in the treatment group fell outside the range of the control group's propensity scores. Because none did, no observations from the treatment groups were excluded. We tested other matching methods, but because one of our goals is to have an equal sample size for the two groups, one-to-one matching is preferred. Furthermore, because having as many observations as possible is important, we did not want to exclude any of the observations in the treatment group, making nearest-neighbor matching the only feasible method. We are aware of the limitations of the method: if, for example, we had used so-called caliper matching without replacement, we could have avoided some inferior matches, but the cost of doing so would have been excluding some of the treated observations. Because the balancing tests (Rosenbaum

& Rubin, 1985) are at an acceptable level, there is no reason to exclude any of the observations from the treatment group in our analysis.

To further investigate the matching, summary statistics are provided in Table 5 for four different samples. Column 2 shows the full sample of firms that use electronic information sharing. In the remaining 3 columns, we present three different samples of firms that do not use electronic information sharing. More specifically, column 3 is the full sample of firms that do not use electronic information sharing. Column 4 is a random sample of the same size as the treatment group. Column 5 is the matched sample of firms that do not use electronic information sharing. We observe that using a random sample does not guarantee a sample with the same characteristics as those of the treatment group. In the appendix, we have included the logistic regression, the standardized differences, the associated test, and the distribution of the bias before and after matching.

Share of firms:	Externally digitizing	Externally non-digitizing		
	Full sample	Full sample	Random samples	Matched sample
Number of observations	287	815	287	287
Up to 20 employees	0.0523	0.1252	0.1394	0.0523
21-50 employees	0.1115	0.2221	0.2195	0.1115
51-100 employees	0.1951	0.2798	0.2753	0.1951
101-250 employees	0.2927	0.2454	0.2404	0.3066
251 or more employees	0.3484	0.1276	0.1254	0.3345
Food, Tobacco	0.1289	0.1227	0.1254	0.1429
Textiles, Leather	0.0279	0.0307	0.0244	0.0348
Wood	0.0209	0.0552	0.0383	0.0209
Paper, Publishing	0.1185	0.0822	0.0836	0.1080
Chemical products	0.0592	0.0417	0.0557	0.0383
Plastic	0.1080	0.1276	0.1289	0.1115
Metal	0.1498	0.1620	0.1602	0.1638
Machinery	0.1429	0.1742	0.1707	0.1882
Electric equipment	0.1359	0.1055	0.1184	0.1080
Transport equipment	0.0488	0.0331	0.0418	0.0244
Furniture and other	0.0592	0.0650	0.0522	0.0592

Table 5 Summary statistics for the matching variables in the analysis of externally digitizing firms

Notes: Size dummies and industry dummies.
Source: Register and survey data from Statistics Denmark

In general, the matched sample is much more similar to the treatment group than any of the other samples in the control group. For example, 35 percent of the firms in the treatment group had more than 250 employees, whereas only 13 percent of the control group had that many employees. After matching, 34 percent of firms in the control group have more than 250 employees—an increase of 21 percentage points. In the same way, the four other size-based groups match much better than if the full control group or a random sample of it is used. The distribution of firms across industries does not vary as much across samples as firm size did. However, using matching rather than a random sample still results in important improvements. For instance, the share of firms in the wood industry falls by 3.5 percentage points toward a better match with the treatment group. Of course, with a relatively small control group, when more variables are included in the matching procedure, finding matches is more difficult. Table 5 also shows that the proportion of firms in some industries is further away from that of the treatment group when matching is used than when the full control group is used. For example, the share of firms in the food industry is 13 percent in the treatment group, 12 percent in the control group and 14 percent in the matched sample.

Earlier studies make two recommendations regarding proper balancing of the matched sample. First, the bias should be less than 5 percent after matching (Rosenbaum & Rubin, 1985); second, a t-test for equality of means should be non-significant after matching for all the matching variables. In the appendix, we have included the matching analysis performed in STATA, where both the bias and the t-test are reported. In our case, all except one of the matching variables satisfy the first recommendation. The second recommendation concerning the non-significant difference in the matching variables across the two groups is satisfied for all of the matching variables.[2]

Figure 1 shows sales per employee and material use per employee in the different firms. The Figure includes four subplots. The two top subgraphs, and the bottom left subgraph show the treatment group, the control group, and the matched firms from the control group, respectively. The bottom right subgraph shows all the firms and the ways in which they have been divided into two groups through matching. The graph shows that the matched sample does not contain firms to the left, close to the y-axis, with relatively low sales, or firms high up with high material use per employees, or the firms with high sales, which are preferred, as no treatment firm is close by. Examining the information to the right, we see that through matching, we have ensured that the four firms in the treatment group with the greatest sales and low material cost per employee are accompanied by similar firms in

[2] For all the matching variables, the test is also performed before matching is conducted, and 6 of 16 variables are significantly different from each other at the 5% level. Not only have these significant differences been excluded through matching, but also the similarity of most of the remaining 10 variables has increased.

the control groups—even though we have not matched on sales and material. Hence, by using matching, we ensure that the matched companies resemble the digitized firms as much as possible, and we can therefore, to a large extent, delete the self-selection effect.

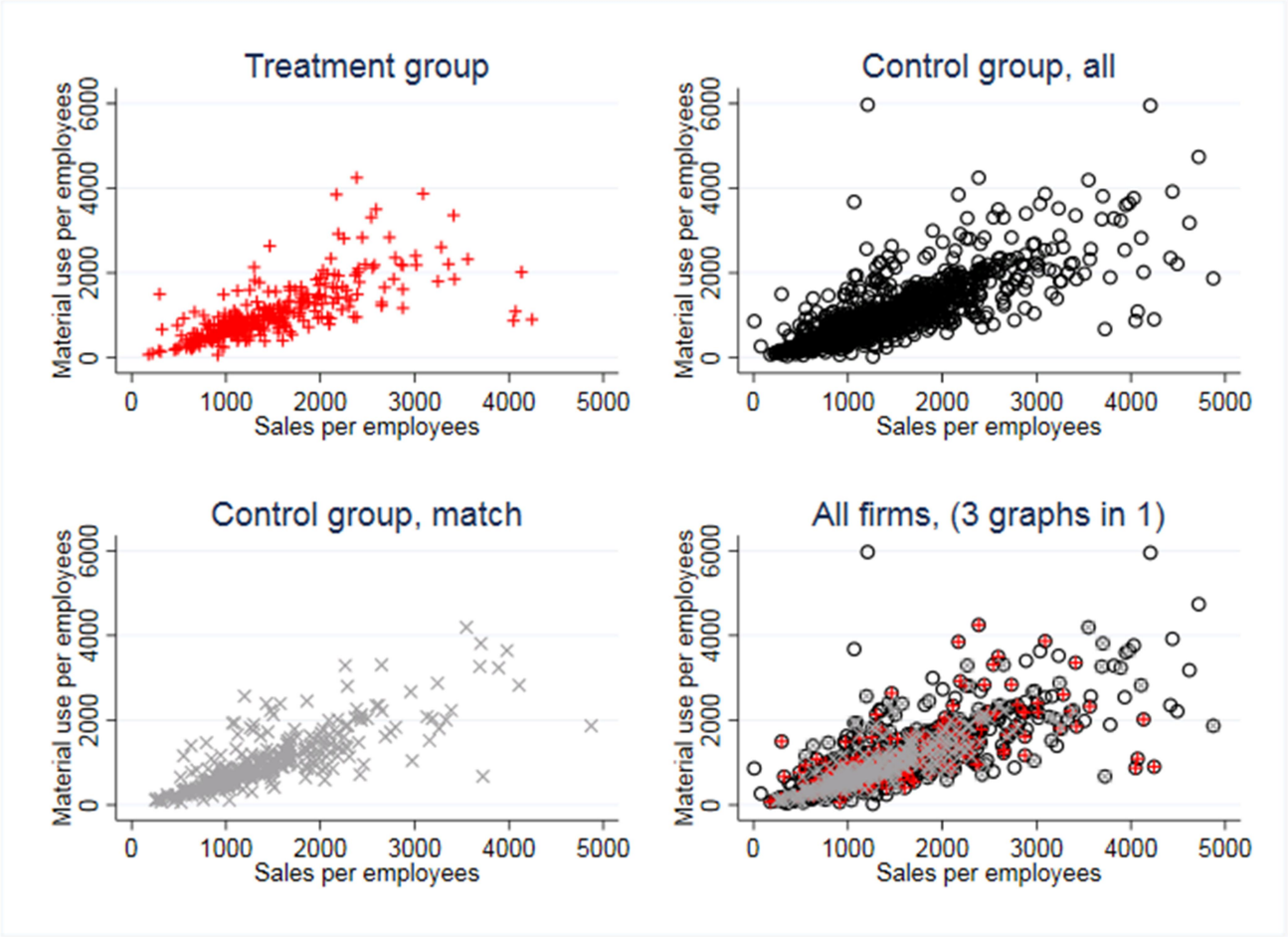

Figure 1 The distribution of the firms across treatment and control groups

4.4 Measuring frontier shifts due to the treatment

In Table 6, the frontier shifts from group S, the control group, to group T, the treatment group, are shown using equations (1). Both $FC^i(s)$ and $FC^i(t)$ are shown for four different cases. Thus, the first two columns show them for the full sample for both groups of firms. Columns 3 and 4 show the average $FC^i(s)$ and $FC^i(t))$ across

500 random samples of firms from the control group using the same sample size as the treatment group. In columns 5 and 6, the results are shown using matching to obtain a group of similar firms for the control group. Finally, in columns 7 and 8, the results are shown for another matched sample, where we have matched additional variables. To ensure that outliers do not drive the results, six firms with a negative capital stock or a capital stock per employee of more than DKK 1,500,000 have been excluded. In the same way, two firms with sales per employee of more than DKK 10,000,000 have been excluded.

	Full sample		Random sample		Matching 1		Matching 2	
	$FC^i(s)$	$FC^i(t)$	$FC^i(s)$	$FC^i(t)$	$FC^i(s)$	$FC^i(t)$	$FC^i(s)$	$FC^i(t)$
Observations	815	287	287	287	287	287	286	286
Minimum	0.252	0.718	0.514	0.821	0.701	0.871	0.252	0.907
1st quartile	0.985	0.997	1.207	1.213	1.031	1.017	1.14	1.156
Median	1.091	1.073	1.374	1.366	1.178	1.157	1.256	1.275
Mean	1.101	1.101	1.35	1.361	1.236	1.242	1.236	1.26
3rd quartile	1.224	1.216	1.506	1.501	1.418	1.428	1.339	1.37
Maximum	1.486	1.508	1.873	2.891	1.89	2.563	1.504	2.207

Table 6 Frontier shift

Notes: Frontier information from a DEA model assuming constant return to scale technology. Output is sales measured in DKK 1,000. Input is material and the capital stock measured in DKK 1,000 and number of employees in the firm. The first two columns show results for full samples of both groups. In columns 3 and 4, a random sample of 287 of the firms from the control group is used, and the result is an average across 500 random samples of the control group. Columns 5 and 6 use matching to pick 287 firms from the control group using industry and firm size, and columns 7 and 8 uses matching to pick 287 firms from the control group using industry, firm size, education dummies, export, and sales.
Source: Register and survey data from Statistics Denmark

The results of the full sample show that firms that use external electronic information sharing cause an average frontier shift of approximately 10 percent (1.101) compared with firms that do not use external electronic information sharing. However, this result is most likely driven largely by the bias problem of DEA. Because we have a much larger sample of non-digitized firms, the best of these firms are likely to outperform the best from the smaller sample of digitized firms simply because we have a larger bias in the small sample of digitalized firms. For this reason, we expect that the shift in the full sample is biased downward.

Since Zhang and Bartels (1998), a common method is to take the average across 500 random samples of the same size as the smaller of the two groups. Using this method, firms that use external electronic information

sharing improve their performances by approximately 35 percent (1.350 and 1.361)—a much larger shift than in the previous case—confirming that the shift in the full sample is biased downward.

This solution is perfectly adequate if all the firms in the population are equally likely to use external electronic information sharing. However, if large firms with more orders, larger inventories and more demands are expected to be more likely to invest in electronic information sharing, the likelihood of large firms being in the control group should increase. In the same way, across industries, the amount of information that is shared with suppliers and customers differs. Hence, the distribution of firms across industries should be the same in the two groups to avoid overestimating or underestimating the effect of external electronic information sharing. Columns 5 and 6 of Table 6 show the results for the DEA model, where the control group has been chosen using matching for size and industry. The average frontier shift is approximately 24 percent, which falls between the two earlier cases. This minor example shows that firms do not decide to use external electronic information sharing at random. In addition, it shows that if only a random sample is used, the effect will be overestimated by nearly 10 percentage points. The importance of comparisons with similar firms is observed by examining not only the means but also the quartiles. For example, in the case where similar firms are compared, approximately 75 percent of them outperformed those that did not invest because the 1st quartile is just above 1. This is not the case if a random sample of firms is used; here, the 1st quartile is well above 1—again, a sign of overestimation.

Of course, one of the limitations of matching is that the number of firms in the control group has to be somewhat larger to be able to ensure good matches. This problem increases if we seek to use more variables for matching. In our case, we only have 815 firms in the control groups, and 1/3 of them have to be assigned to the control group each time. In the appendix, Table A1.1 shows the matching results for a case in which only firm size is used as a matching variable. In this case, we are able to find a perfect match for each of the firms. Table A1.2 shows the results for a case in which both size and industry are matched—i.e., the control group used in columns 5 and 6 in Table 6. In Table A1.2, the percentage bias is less than 5 percent after matching for all but one variable—the electronic equipment industry—and the test is non-significant for all the variables. Therefore, we can conclude that overall, we have a good, though imperfect, match. Should we add more matching variables? In general, the matching will be less and less precise as we add more and more variables to the matching procedure.

As argued earlier, matching is used to minimize the variability in firm characteristics across the groups to ensure that the effect we find is actually attributable to the treatment analyzed. The challenge involves using the firm characteristics that confound the treatment effect—and only those. Barber and Lyon (1996) argue that by matching performance, one can control for various factors that affect performance but that are unrelated to the

treatment. For instance, if a firm has enjoyed an unusually good performance and continues to have an above-average performance after an event, this firm seemingly enjoys a performance that exceeds the performance expected in the absence of the event. Therefore, matching with 3-year-old performance measures may provide more-accurate estimates. As such, we have also tried an extended matching procedure in which we match not only size and industry but also two education variables (the firm's shares of highly educated workers and of workers with mid-level education), an export dummy, and a sales variable (value added per employee)—all from three years earlier. We experimented with many other firm characteristics in the educational, occupational, age, export and 3-year-old sales categories, but most of them did not ultimately differ across the two groups. Table A1.3 shows the results for the extended matching. Unfortunately, the matching is not convincing, as the percent bias is not less than 5 percent after matching for 5 of the 14 variables used for matching. The obvious explanation for these unconvincing results is that we only have a small group of firms from which to choose, making good matches difficult to find. Now, applying this less-than-perfect matched sample, we obtain the frontier shifts in columns 7 and 8 of Table 6. We observe that the frontier shifts are somewhat higher than they are for the first matching, but they are still much lower than the random sample results. We have mainly included these columns to show that matching is not something that can always be performed satisfactorily and to emphasize the importance of carefully choosing the matching variables, especially if the control group is relatively small.

4.5 Second-stage analysis

As mentioned earlier, one common solution in the DEA literature involves making second-stage adjustments using a Tobit model. Thus, we can include all the variables that might have an impact on efficiency and make adjustments, as described above. Alternatively, because we are actually interested in the treatment effect, we can use a Tobit regression of the efficiency scores from the full sample on the matching variables and a dummy that divides the firms into a treatment group and a control group to examine whether the treated firms, on average, have higher efficiency scores (see Table 7). Compared with the matching procedure above, the analysis based on a Tobit regression only provides us with one number—the average treatment effect—whereas the DEA model that uses matching provides us with an entire distribution of individual firm effects. Because Table 6 shows that, independent of the control group used, we found a positive frontier shift, we expect that the treatment dummy is positive and significant, with the treatment group having higher efficiency scores than the full sample. Somewhat surprisingly, the treatment dummy is insignificant in all the Tobit models in Table 7, irrespective of which variables are used as controls. Thus, in our case, the second-stage adjustment did not catch the selection bias.

	R1	R2	R3	R4
Treatment dummy	0.016	0.009	0.005	0.003
	(0.011)	(0.011)	(0.011)	(0.011)
21–50 employees		0.001	-0.002	0.001
		(0.018)	(0.018)	(0.018)
51–100 employees		0.023	0.019	0.003
		(0.017)	(0.017)	(0.017)
101–250 employees		0.030*	0.024	0.01
		(0.018)	(0.017)	(0.017)
251 or more employees		0.035*	0.03	0.008
		(0.019)	(0.019)	(0.019)
Textiles, leather			-0.024	-0.02
			(0.028)	(0.027)
Wood			-0.102***	-0.030***
			(0.023)	(0.023)
Paper, Publishing			-0.031*	-0.043*
			(0.017)	(0.017)
Plastic			-0.066***	-0.055***
			(0.015)	(0.015)
Metal			-0.055***	-0.041***
			(0.014)	(0.014)
Electric Equipment			0.002	-0.024
			(0.016)	(0.016)
Share of workers with mid-level education, 3 years ago				0.416***
				(0.072)
Share of highly educated workers, 3 years ago				0.257***
				(0.088)
Sales, 3 years ago				0.000***
				(0.000)
Exports, 3 years ago				0.006
				(0.011)
No. of firms	1,102	1,102	1,102	1102
No. of right-censored firms	10	10	10	10

Table 7 Tobit coefficients

*Notes: Tobit regression of the efficiency scores from the full sample on the matching variables and a dummy that divides the firms into a treatment group and a control group. We use a modified Tobit model that is right-censored at 1. * 10%, ** 5%, and *** 1% significance level.*

Source: Register data from Statistics Denmark

4.6 Who is affected by electronic information sharing and how?

One explanation of the difference between the matching and the Tobit approach may be that the latter captures the average effect of the treatment, while our combination of DEA and matching is designed to capture the treatment effects on the frontier, i.e., how the most-efficient firms are benefitting from increased external digitization.

Using the matching procedure, we can calculate the effect of external digitization on the average firm as the average Malmquist index, i.e., as the combined effect of the frontier shifts and catch-up. These calculations are shown in Table 8 below.

Matched sample for the control group: industry and firm size

	Frontier shift	Catch-up	Malmquist index (product)
Observations	287	287	287
Minimum	0.94	0.244	0.289
1st quartile	1.066	0.56	0.719
Median	**1.2**	**0.824**	**0.986**
Mean	1.227	0.928	1.126
3rd quartile	1.334	1.117	1.346
Maximum	1.91	3.666	4.476

Table 8 Frontier shift, catch-up and Malmquist index using matching

Source: Register data from Statistics Denmark

Table 8 shows that the median frontier shift is 20 percent. Catch-up however, is negative 18 percent, making the combined improvement from the treatment approximately zero. The latter is consistent with the findings from the Tobit approach. The Tobit approach thus focuses on the combined effect of the frontier shift and catch-up, i.e., the effect on the average firm. However, the matching approach allows us to divide this combined effect into non-biased estimates of the frontier effect and the catch-up effect. This finding is interesting from an industrial economics perspective because it highlights the industry dynamics under which the best firms benefit more from new technology and less-efficient firms fall further behind. Likewise, such information is interesting from a policy perspective because it shows which firms to target in an attempt to improve welfare through the increased use of electronic information exchange.

5. Conclusions

DEA allows us to evaluate not only changes in average performance but also changes in frontier performance when some firms are subject to a new treatment, for example, a new regulation or a new business model. Unfortunately, DEA estimates are biased if the treated firms do not select into the treatment and control group at random. In this case, a treatment effect may be confounded with a bias effect. In this paper, we argued that matching can be useful in such cases. It not only allows us to construct a relevant counterfactual group of non-treated firms but also allows us to balance the comparisons so that bias effects are not confounded with treatment effects.

We illustrated the approach by evaluating how information sharing has an impact on the performance of Danish manufacturing firms. The frontier shift from information sharing is shown to be approximately 20 percent when the control group consists of a matched sample of comparable firms. Ignoring the size difference between the treatment and non-treatment groups, we estimate a frontier shift of only 10 percent. This result underestimates the gains from information sharing, because the sample size bias in the large non-treatment group is smaller than that in the small treatment group. The traditional approach of correcting for the sample size effect on the bias is to use a randomized subsampling in the large group. Using this approach leads to a 35-percent frontier shift. However, this approach may exaggerate the benefits if the random samples differ from the treatment sample in terms of a series of covariates. When firms self-select into the treatment group, we expect the treatment group to have somewhat different characteristics. To see the real effect of the treatment, we need the control group to have similar characteristics. The proposed approach of using a matched sample is thus expected to provide the most-reliable estimates of both the average and the frontier effects of treatment.

We have also illustrated the gains from using a matched sample compared with second-stage adjustments using a Tobit model. A Tobit approach focuses on the combined effect of the frontier shift and catch-up, and the matching approach allows us to separate these two effects. We observed that the combined effect is non-significant in the Tobit approach. The matching approach suggests a positive frontier impact of approximately 20 percent and a negative catch-up effect of approximately 18 percent. This finding suggests industry dynamics under which the best firms benefit more from new technology and the less efficient firms suffer and fall further behind. This information can also guide industry policy because it provides information regarding which firms to target in an attempt to improve welfare by changing a specific business practice.

The ability to provide unbiased estimates of the frontier and average effects of a treatment is interesting in many

applications. Our example illustrates an application that is relevant to industrial economics. To provide another example, one may examine the effects of a new medical procedure or a new drug. From a health economics and health policy perspective, whether this procedure or drug is likely to affect the average patient or only the strong (or weak) patients is of interest.

References

Abri, A. G., & Mahmoudzadeh, M. (2015). Impact of information technology on productivity and efficiency in Iranian manufacturing industries. *Journal of Industrial Engineering International, 11*, 143-157.

Barber, B. M., & Lyon, J. D. (1996). Detecting abnormal operating performance: The empirical power and specification of test statistics. *Journal of Financial Economics, 41*, 359-399. doi:10.1016/0304-405X(96)84701-5.

Bharadwaj, A. S. (2000). A resource-based perspective on information technology capability and firm performance: An empirical investigation. *MIS Quarterly, 24*, 169-196. doi:10.2307/3250983.

Bogetoft, P. (2012). *Performance benchmarking.* New York: Springer Verlag.

Bogetoft, P., & Otto, L. (2011). *Benchmarking with DEA, SFA, and R.* New York: Springer Verlag.

Brookhart, M. A., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J., & Stürmer, T. (2006). Variable selection for propensity score models. *American Journal of Epidemiology, 163*, 1149-1156. doi:10.1093/aje/kwj149.

Caves, D. W., Christensen, L. R., & Diewert, W. E. (1982). The economic theory of index numbers and the measurement of input, output, and productivity. *Econometrica, 50*, 1393-1414. doi:10.2307/1913388.

Charnes, A., Cooper, W. W., & Rhodes, E. (1978). Measuring the efficiency of decision-making units. *European Journal of Operational Research, 2*, 429-444. doi:10.1016/0377-2217(78)90138-8.

Charnes, A., Cooper, W. W., & Rhodes, E. (1979). Short communication: Measuring the efficiency of decision making units. *European Journal of Operational Research*, *3*, 339. doi:10.1016/0377- 2217(79)90229-7.

Chen, Y., & Ail, A. I. (2004). DEA Malmquist productivity measure: New insights with an application to computer industry. *European Journal of Operational Research, 159,* 239-249.

Chumpitaz, R., Kerstens, . K, Paparoidamis, N,& Staat, M. (2010). Comparing efficiency across markets: An extension and critique of the Zhang and Bartels (1998) methodology. *European Journal of Operational Research,* 719-728.

Coelli, T., Rao, D., & Battese, G. (1998). *An introduction to efficiency and productivity analysis*. *Boston Dordrecht red*. London: Kluwer Academic Publishers.

Cooper, W., Seiford, L., & Tone, K. (2007). *Data envelopment analysis: A comprehensive text with models, applications, references and DEA–solver software.* (2nd ed.). Secaucus: Springer.

Cubbin, J. (2004). Some more statistical properties of data envelopment analysis. Working paper City University.

Deprins, D., Simar, L., & Tulkens, H. (1984). Measuring labor efficiency in post offices In M. Marchand, P. Pestieau, & H. Tulkens (Eds.), *The Performance of Public Enterprises: Concepts and measurements*. Amsterdam: North Holland.

Färe, R., Grosskopf, S., Lindgren, B., & Ross, P. (1994). Productivity development in Swedish hospitals: A Malmquist output index approach. In: *Data envelopment analysis: Theory, methodology, and application,* 253-272. Boston: Kluwer.

Farrell, M. J. (1957). The measurement of productive efficiency. *Journal of the Royal Statistical Society, 120*, 253-281. doi:10.2307/2343100.

Gijbels, I., Mammen, E., Park, B. U., & Simar, L. (1999). On estimation of monotone and concave frontier functions. *Journal of the American Statistical Association, 94*, 220-228. doi:10.1080/ 01621459.1999.10473837.

Giokas, D., Eriotis, N., & Dokas, I. (2015). Efficiency and productivity of the food and beverage listed firms in the pre-recession and recessionary periods in Greece, *Applied Economics, 47, 19,* 1927-1941

Heinze, G., & Jüni, P. (2011). An overview of the objectives of and the approaches to propensity score analyses. *European Heart Journal*, 32, 1704-1708. doi:10.1093/eurheartj/ehr031.

Kittelsen, S. (1995). Monte Carlo simulations of DEA efficiency measures and hypothesis tests. University of Oslo doctoral thesis.

Kneip, A., Park, B. U., & Simar, L. (1998). A note on the convergence of nonparametric DEA estimators for production efficiency scores. *Econometric Theory, 14*, 783-793. doi:10.1017/S0266466698146042.

Malmquist, S. (1953). Index numbers and indifference surfaces. Trabajos de Estadistica*, 4*, 209-242. doi:10.1007/BF03006863.

Pedraja-Chaparro, F., Salinas-Jiménez, J., & Smith, P. (1999). On the quality of the data envelopment analysis model. *Journal of the Operational Research Society, 50*, 636-644. doi:10.1057/palgrave.jors.2600741.

Rosenbaum, P. R., & Rubin, D. B. (1983). The Central role of the propensity score in observational studies for causal effects. *Biometrika, 70*, 41-55. doi:10.1093/biomet/70.1.41.

Rubin, D. B., & Thomas, N. (1996). Matching using estimated propensity scores: Relating theory to practice. *Biometrics, 52*, 249-264. doi:10.2307/2533160.

Seiford, L. (1994). A DEA bibliography (1978–1992). In A. Charnes, W. Cooper, & A. Lewin (Guest Eds.), *Data envelopment analysis: theory, methodology, and application*, 437-469. Boston: Kluwer Publishing House.

Simar, L., & Wilson, P. W. (2000). Statistical inference in nonparametric Frontier models; the state of the art. *Journal of Productivity Analysis, 13*, 48-78.

Smith, J., & Todd, P. (2005). Rejoinder. *Journal of Econometrics, 125*, 365-375. doi:10.1016/ j.jeconom.2004.04.013.

Tavaras, G. (2002). A bibliography of data envelopment analysis (1978-2001). *RUTCOR, Rutgers University, 11,* 14.

Zhang, Y., & Bartels, R. (1998). The effect of sample size on the mean efficiency in DEA with an application to electricity distribution in Australia, Sweden and New Zealand. *Journal of Productivity Analysis, 9,* 187-204.

Appendix: Matching

This appendix shows the matching analysis, where the treatment variable is the "external digitization of the information flow". Three tables are shown: in the first one, firm size is the only matching variable; in the second one, both size and industry are used as matching variables; and, in the last table, information about export, the share of workers with a high level of education, and sales figures, all three-year-old, are added as matching variables. The data were sorted randomly before performing the matching analysis because the matching method depends on the order of the observations. The logistic regression was used to predict each firm's propensity score. There were 1102 observations, and the R-squared is between 6.7 percent and 8.3 percent. The first column in Tables A1.1 to A1.3 shows the coefficient from the logistic regression. The tables also show the mean for each binary variable included in the logistic regression before and after matching divided into the two groups—treated and control. In columns 5 and 6, the absolute standardized differences (% Bias) and the reduction in the bias (% Reduction) after matching, respectively, are shown. Rosenbaum and Rubin (1985) suggest that the percent bias should be less than 5 percent after matching for good balance. Finally, the t-test for the equality of means in the treated and control groups and the p-value are shown in columns 7 and 8, respectively. For good balance, the test should be non-significant after matching.

Logistic regression (Treatment group/control group)		**Means before and after matching**						
	Coef. (SE)	Sample	Treated	Control	% Bias	% Reduction	t-test	p>|t|
Up to 20 employees	Base							
21–50 employees	0.1841	Unmatched	0.115	0.222	-30		-4.11	0
		Matched	0.115	0.115	0	100	0	1
51–100 employees	0.5129	Unmatched	0.195	0.28	-20		-2.83	0.01
		Matched	0.195	0.195	0	100	0	1
101–250 employees	1.049***	Unmatched	0.293	0.245	10.7		1.58	0.12
		Matched	0.293	0.293	0	100	0	1
251 or more	1.878***	Unmatched	0.348	0.128	53.6		8.55	0
		Matched	0.348	0.348	0	100	0	1
R-squared	0.067							

Table A1.1: Matching using firm size dummies

*Notes: *10%, ** 5%, and *** 1% significance level.*

Source: Register data from Statistics Denmark

In Table A1.1, the percent bias is less than 5 percent after matching for all variables, and the test is non-significant for all variables, thereby allowing us to conclude that, overall, we have a good match.

Logistic regression (Treatment group/control group)		Means before and after matching						
	Coef.	Sample	Treated	Control	% Bias	% Reduction	t-test	p>|t|
21–50 employees	0.171	Unmatched	0.112	0.222	-30		-4.11	0
		Matched	0.112	0.112	0	100	0	1
51–100 employees	0.543	Unmatched	0.195	0.28	-20		-2.83	0.01
		Matched	0.195	0.195	0	100	0	1
101–250 employees	1.074***	Unmatched	0.293	0.245	10.7		1.58	0.12
		Matched	0.293	0.307	-3.1	71	-0.36	0.72
251 or more	1.903***	Unmatched	0.348	0.128	53.6		8.55	0
		Matched	0.348	0.334	3.4	94	0.35	0.73
Textile, leather	0.204	Unmatched	0.028	0.031	-1.7		-0.24	0.81
		Matched	0.028	0.035	-4.1	-149	-0.48	0.63
Wood	-0.913**	Unmatched	0.118	0.082	12.1		1.83	0.07
		Matched	0.118	0.108	3.5	71	0.39	0.68
Paper, Publishing	0.42*	Unmatched	0.021	0.055	-18		1.36	0.18
		Matched	0.021	0.021	0	100	0.35	0.73
Plastic	-0.205	Unmatched	0.108	0.128	-6.1		-0.87	0.38
		Matched	0.108	0.112	-1.1	82.2	-0.13	0.89
Metal	0.134	Unmatched	0.15	0.162	-3.3		-0.48	0.63
		Matched	0.15	0.164	-3.8	-14.8	-0.46	0.65
Electric Equipment	0.258	Unmatched	0.136	0.106	9.3		1.4	0.16
		Matched	0.136	0.108	8.6	8.2	1.02	0.31
R-squared	0.076							

Table A1.2: Matching using firm size and industry dummies

*Notes: * 10%, ** 5%, and *** 1% significance level.*

Source: Register data from Statistics Denmark

In Table A1.2, the percent bias is less than 5 percent after matching for all but one variable - the electric equipment industry - and the test is non-significant for all variables, thereby allowing us to conclude that, overall, we have a good match.

Logistic regression (Treatment group/control group)		Means before and after matching						
	Coef.	Sample	Treated	Control	% Bias	% Reducti	t-test	p>|t|
21–50 employees	0.199	Unmatched	0.112	0.222	-30		-4.11	0
		Matched	0.112	0.105	1.9	934	0.27	0.79
51–100 employees	0.536	Unmatched	0.195	0.28	-20		-2.83	0.01
		Matched	0.195	0.175	0.8	75	0.64	0.52
101–250 employees	1.031***	Unmatched	0.293	0.245	10.7		1.58	0.12
		Matched	0.293	0.315	4.7	56	-0.54	0.59
251 or more employees	1.743***	Unmatched	0.348	0.128	53.6		8.55	0
		Matched	0.348	0.346	0.8	98	0.09	0.93
Textile, leather	0.202	Unmatched	0.028	0.031	-1.7		-0.24	0.81
		Matched	0.028	0.032	-2.1	-25	-0.25	0.81
Wood	-0.792*	Unmatched	0.021	0.055	-18		-2.38	0.02
		Matched	0.021	0.021	-1.8	90	-0.3	0.76
Paper, Publishing	0.621**	Unmatched	0.118	0.082	12.1		1.83	0.07
		Matched	0.118	0.108	3.5	71	0.27	0.79
Plastic	-0.115	Unmatched	0.108	0.128	-6.1		0.39	0.69
		Matched	0.108	0.133	-7.6	-25	-0.9	0.37
Metal	0.21	Unmatched	0.15	0.162	-3.3		-0.48	0.63
		Matched	0.15	0.171	-5.8	-73	-0.68	0.45
Electric Equipment	0.25	Unmatched	0.136	0.106	9.3		1.4	0.16
		Matched	0.136	0.115	6.4	31	0.76	0.45
Share of workers with mid-level education, 3 years ago	-1.407	Unmatched	0.084	0.072	16.6		2.47	0.01
		Matched	0.084	0.086	-1.6	91	-0.19	0.85
Share of highly educated workers, 3 years ago	2.810**	Unmatched	0.043	0.026	26.1		4.19	0
		Matched	0.043	0.039	6.1	77	0.66	0.51
Sales, 3 years ago	0	Unmatched	0	0	25.6		4.55	0
		Matched	0	0	12.5	51	1.33	0.18
Exports, 3 years ago	0.308**	Unmatched	0.777	0.662	26.4		3.63	0
		Matched	0.777	0.762	3.9	85	0.5	0.62
R-squared	0.084							

Table A1.3: Matching using industry, firm size, education level, export and sales

*Notes: * 10%, ** 5%, and *** 1% significance level.*

Source: Register data from Statistics Denmark

In Table A1.3, the test is non-significant for all variables, but the percent bias is not less than 5 percent after matching for 5 of the 14 variables used for matching. Because our sample only contains 815 firms, of which 287 should be matched, matching becomes increasingly difficult as more matching variables are included.