Using matching methods to account for selection bias in Norway's Primary Care Teams (PCT) pilot

ØYVIND SNILSBERG^{1*}

¹ Department of Health Management and Health Economics, University of Oslo, Norway

Abstract: Norway is piloting team-based primary care delivery models: Honorarmodellen (HM) and Driftstilskuddsmodellen (DM). In addition to organisational changes, the DM transforms provider payment, which seems to attract specific practices. This, coupled with the small number of DM practices, makes it difficult to produce credible evidence regarding the model and its effects on health system performance. I examine whether matching methods—specifically, coarsened exact matching, propensity score matching, and propensity score weighting—can improve evaluation in this demanding situation. As in previous studies on the small sample performance of matching methods, I find no clear best method. This suggests using propensity score weighting, which does not discard data. In the final section of the article, I offer additional advice to help improve the evaluation in similar situations.

JEL classification: I10, I18

Keywords: matching methods, selection bias, difference-in-differences, primary care

1 Introduction

Efficient healthcare systems depend on good primary care. Good primary care requires the right organisation. The Organisation for Economic Co-operation and Development advocates multiprofessional or team-based care (OECD, 2020). Traditionally, the individual general practitioner (GP) assumes sole responsibility for a patient's care. Team-based care redistributes the responsibility among the members of a practice team. The aim is more comprehensive and coordinated care for patients with complex care needs.

Among the Nordic countries, team-based care can be observed in Sweden and Finland, and Norway and Denmark may follow suit. First, the Nordic countries all face the same overall healthcare challenges and may look to each other for solutions (Krasnik and Paulsen, 2009; Olsen et al., 2016). Second, patient and provider surveys suggest a positive outlook towards team-based care in both Norway and Denmark (EY and Vista Analyse, 2019; OECD, 2017).

To date, studies on the effects of team-based care on health system performance have largely been focused on Canada (Glazier et al., 2015; Strumpf et al., 2017a), France (Mousquès and Bourgueil, 2014), and the United States (Friedberg et al., 2015; Jones et al., 2016; Mahmud et al., 2018; Pines et al., 2015; Rosenthal et al., 2016; van Hasselt et al.,

**Correspondence to:* Øyvind Snilsberg, Department of Health Management and Health Economics, University of Oslo, Box 1089, Blindern, 0317 Oslo, Norway. E-mail: oyvind.snilsberg@medisin.uio.no.

Published: Online November 2021. [dx.doi.org/10.5617/njhe.8562](https://doi.org/10.5617/njhe.8562)

2015). While the French and North American experience demonstrates that team-based care can improve health system performance, it is less informative about how this type of primary care delivery model should be implemented elsewhere. To guide implementation, there may be local trials in which practices volunteer to test team-based care. Since volunteers might differ from those that opt out, trials are susceptible to selection bias. To account for selection bias, the French and North American evaluations use matching methods to construct a control group that resembles the treated group at baseline, and difference-in-differences (DD) methods to compare the groups.

The Nordic countries provides an attractive context for this strategy owing to their national registries, which meet the data requirements both for constructing credible control groups and conducting DD analysis. The challenge in Nordic countries, however, is that funders lack the resources to include an adequate number of practices. This might render matching methods unfit, as they were developed under the assumption of an adequate sample size. Moreover, Cenzer et al. (2020) note that the small sample performance of matching methods has only been scarcely studied, and that current evidence based on simulation studies suggests no clear best method. The present article complements the literature with a novel comparison of three common matching methods using a real-world example: Norway's Primary Care Teams (PCT) pilot. The methods considered herein are coarsened exact matching (CEM), propensity score matching (PSM), and propensity score weighting (PSW). However, the main contribution is new insight into the type of practices that team-based primary care delivery models might attract.

2 Norway's PCT pilot

The last comprehensive reform to Norwegian primary care was the regular GP scheme in 2001, which introduced a list patient system. While this system has enabled GPs to fulfil the primary care function for many years, increased GP workload has raised concerns about the quality of primary care (Svedahl et al., 2019). Therefore, Norway is piloting two team-based primary care delivery models: Honorarmodellen (HM) and Driftstilskuddsmodellen (DM). All practices shared by at least three GPs were eligible to apply, but only 13 applied, all of which were admitted. These employed nurses, and GPs, nurses and medical secretaries formed GP-led care teams.

The regular GP scheme also introduced a default remuneration system that blends capitation (30% of payment) and fee-for-service (ffs) (70% of payment). The HM maintains the traditional remuneration system, but to cover the cost of employing nurses, GPs receive government funding as well as ffs for the services that the nurses perform. In the DM, only 20% of the payment is ffs, while 80% is a practice allowance. The allowance is based on average GP remuneration, adjusted in particular for the cost of employing nurses and the age-sex distribution of list patients.¹ Lastly, some GPs are employed by their local municipality for a fixed salary.

3 Data and descriptive statistics

3.1 Data source and sample selection

To describe practices at baseline, I combine data from several Norwegian registries for three years prior to the announcement of the PCT pilot. Data on practices and GPs comes from the Norwegian Health Economics Administration.² I retain practices shared by at least three

¹ A list of all nine criteria for determining practice allowance is shown in Appendix (Table A1).

² <https://www.helfo.no/fastlegeordninga/basistilskot-for-fastlegeordninga-i-kommunane> (last accessed 28 September 2021).

GPs that were active throughout the baseline period. These data include GP identifiers, which allow me to link practice and GP information to the GP database. The GP database provides the link between GPs and patients, and patients are uniquely identified by a personal identification number. I use patients' personal identification numbers to retain information on year of birth, sex, and socioeconomic background from the Statistics Norway's database, and healthcare utilisation and chronic conditions from the Primary Care Registry of Norway and the Norwegian Patient Registry. The sample consist of five DM practices, eight HM practices and 817 traditional practices. Together, they serve about 75% of the Norwegian population.

3.2 Descriptive statistics

A list of observed characteristics, along with features of their distribution stratified by model, is shown in Table 1.

Table 1: Descriptive statistics of observed baseline characteristics.

Characteristics	DM Mean (Std dev)	HM Mean (Std dev)	Traditional model Mean (Std dev)
Practice characteristics			
Number of GPs	4.8 (1.9)	7.0 (3.5)	4.5 (1.6)
Number of patients per GP	897 (356)	1142 (178)	1169 (276)
Publicly owned (%)	80.0 (44.7)	25.0 (46.3)	27.2 (44.5)
Employs nurses (%)	60.0 (54.8)	37.5 (51.8)	31.1 (46.3)
Expected income gain (%) (a)	-7.0 (14.1)	-14.2 (6.3)	-11.4 (9.4)
Centrality (%) (b)			
1	40.0 (54.8)	25.0 (46.3)	21.3 (41.0)
2	0.0 (0.0)	0.0 (0.0)	29.3 (45.5)
3	0.0 (0.0)	50.0 (53.5)	25.8 (43.8)
4	60.0 (54.8)	25.0 (46.3)	23.6 (42.5)
GP characteristics			
Female (%)	54.4 (24.9)	37.0 (14.3)	40.7 (22.5)
Specialist (%)	47.4 (25.7)	63.3 (24.5)	60.1 (28.4)
Fixed salary (%)	52.3 (31.8)	1.4 (4.0)	3.8 (13.5)
Experience (years) (c)	6.4 (2.4)	9.7 (2.8)	9.7 (3.3)
Patient characteristics			
Female (%)	49.9 (4.4)	49.2 (2.6)	49.9 (3.8)
Age group (years) (%)			
0-19	23.4 (3.5)	25.4 (3.6)	23.6 (4.0)
20-64	63.3 (7.4)	58.2 (2.0)	59.1 (5.0)
65-74	7.8 (3.8)	9.3 (2.3)	9.8 (2.8)
≥75	5.5 (3.0)	7.1 (2.0)	7.6 (2.8)
Education level (%)			
No high school	43.3 (2.2)	44.2 (5.3)	42.1 (6.2)
High school	30.8 (9.9)	34.3 (3.4)	32.8 (6.2)
Higher education	25.9 (10.7)	21.5 (7.5)	25.1 (9.7)
Income quartile (%)			
1	26.6 (6.8)	22.7 (4.2)	24.7 (4.9)
2	25.3 (4.1)	24.8 (4.4)	25.0 (3.5)
3	24.2 (3.7)	26.6 (3.9)	25.4 (3.5)
4	23.9 (10.2)	26.0 (9.9)	24.9 (7.3)
Disability benefits (%) (d)	6.1 (2.2)	6.1 (2.0)	6.5 (2.2)
Sick leave (%) (e)	16.7 (1.8)	17.5 (2.2)	16.8 (2.2)
Healthcare utilisation (f)			
GP services	5.2 (1.6)	5.7 (0.5)	5.9 (1.1)
Emergency department services	0.4 (0.2)	0.3 (0.1)	0.4 (0.2)
Outpatient clinic services	3.8 (1.1)	3.3 (0.7)	3.7 (0.8)
Hospital admissions	0.4 (0.1)	0.3 (0.1)	0.3 (0.1)

Characteristics	DM	HM	Traditional model
	Mean (Std dev)	Mean (Std dev)	Mean (Std dev)
Home health nursing (%)	4.3 (1.5)	4.2 (1.7)	4.4 (1.7)
Chronic conditions (%) (g)			
Alcohol abuse	0.7 (0.3)	0.6 (0.1)	0.6 (0.2)
Medication abuse	0.3 (0.3)	0.1 (0.1)	0.2 (0.2)
Substance abuse	0.7 (0.6)	0.4 (0.2)	0.5 (0.3)
Depressive disorder	5.1 (2.3)	5.3 (1.6)	6.0 (1.7)
Mental retardation	0.3 (0.2)	0.3 (0.2)	0.3 (0.2)
COPD	1.6 (0.5)	2.0 (1.0)	2.0 (0.8)
Type 2 diabetes	3.4 (1.0)	4.3 (1.7)	3.9 (1.3)
N	5	8	817

Note: Table reports the sample mean of observed baseline characteristics for all Norwegian general practices shared by at least three GPs that were active throughout the baseline period. The samples in Columns 2 and 3 include practices that participate in the PCT pilot with the DM and HM, respectively. Column 4 reports the same characteristics for practices outside the PCT pilot. Standard deviation in parenthesis. (a) Expected income gain is the relative difference between observed and counterfactual income conditional on joining the DM. (b) Centrality ranks municipalities from most to least central on a scale of 1–4, that is, practices with a centrality of 1 are located in urban regions. (c) Experience is the average number of years the GPs have been working in the local municipality. (d) Disability benefits is the annual percentage of list patients who receive disability benefits. (e) Sick leave is the annual percentage of list patients on sick leave. (f) Healthcare utilisation is the annual number of services per list patient. (g) A specific chronic condition is the percentage of list patients with at least one GP service related to that condition. Conditions reflect PCT pilot target patients.

According to these summary statistics, DM practices are distinct in terms of practice and GP characteristics, whereas HM practices resemble traditional practices. This is possibly because the DM attracts salaried GPs who avoid the financial risk associated with changes to provider payment. Salaried GPs work in public practices, which typically possess distinctive characteristics.

Expected income gain, the relative difference between observed and counterfactual income conditional on joining the DM, is of particular interest. According to this measure, the DM is more profitable for DM practices than traditional practices. This is despite the low share of elderly patients in DM practices, which negatively affects the practice allowance. One plausible explanation is that the remuneration of salaried GPs, who submit claims on behalf of their employer, is well below average GP remuneration. If the practice allowance is closer to the average GP remuneration, salaried GPs gain from joining the DM. Table 1 shows that the annual number of GP services is 5.2 in DM practices and 5.9 in traditional practices, lending support to this explanation.

Given that matching methods are most useful in imbalanced data, and may be harmful in balanced data (King and Nielsen, 2019), I will focus on constructing a control group for the DM group. I will also refer to traditional practices as non-DM practices.

4 Matching methods

The term *matching* refers to equating groups in terms of observed baseline characteristics to reduce bias due to confounding. Ideally, for each DM practice, one would find one or more non-DM practices with identical values on all covariates $\mathbf{x}$ in Table 1, against which treatment effects could be assessed. However, exact matching is not possible because $\mathbf{x}$ is high dimensional, and most of the covariates are continuous. This problem motivates inexact matching methods, which find or upweight non-DM practices whose $\mathbf{x}$ is similar to that of DM practices. A common balance check is the absolute standardised mean difference (ASMD), which refers to the absolute mean difference between the DM and non-DM groups

divided by the standard deviation in the DM group, $|\bar{x}_{DM} - \bar{x}_{non-DM}|/\sigma_{DM}$. An ASMD less than 0.2 is usually considered negligible (Austin, 2009). However, because matching is implemented iteratively, I need a single summary statistic, rather than the ASMD for each covariate, to determine easily whether new iterations improve balance. Therefore, I use the sum of the ASMDs as a balance check, $\sum_{k=1}^K (|\bar{x}_{DMk} - \bar{x}_{non-DMk}|/\sigma_{DMk})$.

4.1 Coarsened exact matching

Exact matching on continuous covariates is not possible because there are few observations at each value $\mathbf{x}_i$. CEM temporarily coarsens continuous covariates, making them discrete, and thus ensures more observations at each value $\mathbf{x}_i$. DM and non-DM practices with identical values for categorical or coarsened continuous covariates are matched (Iacus et al., 2012). I implement CEM based on a forward selection procedure. That is, I begin with no covariates in the CEM and test the addition of each covariate for different levels of coarsening using the balance check. Specifically, I coarsen continuous covariates using between 1 and 10 evenly spaced cut points, and categorical covariates are left as is. I add the covariate whose inclusion improves balance the most. This process is repeated until no further improvements can be made.

4.2 Propensity score methods

The propensity score, $p(\mathbf{x}_i)$, is the predicted probability that a practice joins the DM given $\mathbf{x}_i$. Rosenbaum and Rubin (1983) show that if selection bias can be accounted for by controlling for $\mathbf{x}_i$, it can also be accounted for by controlling for $p(\mathbf{x}_i)$, thus solving the problem of $\mathbf{x}_i$ being high dimensional. The problem then becomes estimating $p(\mathbf{x}_i)$.

4.2.1 Propensity score matching

Even when matching on $p(\mathbf{x}_i)$ alone, exact matching is not possible, so DM and non-DM practices with similar propensity scores are matched. To predict $p(\mathbf{x}_i)$, I use logistic regression. I develop the model by progressively adding the covariates that result in the largest improvement in balance in the propensity score-matched sample. I also consider adding derived functions of covariates already in the model, such as interactions and quadratic terms. When matching on propensity scores, two issues are relevant: (1) the number of matched non-DM practices, and (2) whether multiple DM practices can be matched to the same non-DM practice. I let the number of matched non-DM practices vary between 1 and 10 and allow multiple DM practices to be matched to the same non-DM practice.

4.2.2 Propensity score weighting

In this technique, non-DM practices are weighted by the inverse probability of being in the non-DM group given $\mathbf{x}_i$, $w_i = p(\mathbf{x}_i)/(1 - p(\mathbf{x}_i))$. That is, non-DM practices that resemble the DM group and differ from their own group receive larger weights. The prediction of $p(\mathbf{x}_i)$ for weighting is analogous to the prediction of $p(\mathbf{x}_i)$ for matching.

5 Results

Using the procedures outlined above, I implement each method to maximise the balance between the DM and non-DM groups.³ Specifically, to implement CEM, I match DM and non-DM practices with identical values for coarsened centrality, experience, number of hospital admissions, and prevalence of alcohol abuse, COPD, and type 2 diabetes. To implement PSM, I use logistic regression to estimate $p(\mathbf{x}_i)$ by modelling the main effects

³ I implement the matching methods using Stata version 16.1 and the commands `cem` (Iacus et al., 2009) and `psmatch2` (Leuven and Sianesi, 2003).

of centrality, experience, share of female GPs, and number of GP services. I match each DM practice to the three nearest non-DM practices in terms of estimated propensity scores. To implement PSW, I use logistic regression to estimate $p(x_i)$ by modelling the main effects of centrality, experience, share of female GPs, number of GP services and hospital admissions, patient age groups and income quartiles, and prevalence of substance abuse. I weigh non-DM practices using estimated propensity scores. It should be stated that all choices regarding implementation add uncertainty to the treatment effect estimates. However, this is usually not seen as a major issue because the correct choices are those that maximise balance (Ho et al., 2007).

Columns 4–6 in Table 2 show the ASMDs between the DM group and the CEM, PSM, and PSW control groups, respectively. For reference, the DM group is compared with the non-DM group in the original data in Column 3. The sum of the ASMDs is shown at the end of Table 2. According to this summary statistic, all methods significantly improve the overall mean balance. However, CEM and PSW slightly outperform PSM

Table 2: Balance of observed baseline characteristics.

Characteristics	DM	Non-DM			
	Mean (Std dev)	Original Data ASMD	CEM ASMD	PSM ASMD	PSW ASMD
Practice characteristics					
Number of GPs	4.8 (1.9)	0.2	0.2	0.5	0.1
Number of patients per GP	897 (356)	0.9	0.3	0.1	0.2
Publicly owned (%)	80.0 (44.7)	1.2	0.4	0.4	0.4
Employs nurses (%)	60.0 (54.8)	0.6	0.0	0.0	0.1
Expected income gain (%) (a)	-7.0 (14.1)	0.4	0.5	0.2	0.3
Centrality (%) (b)					
1	40.0 (54.8)	0.4	0.0	0.0	0.1
2	0.0 (0.0)	0.9	0.0	0.0	0.0
3	0.0 (0.0)	0.8	0.0	0.0	0.0
4	60.0 (54.8)	0.7	0.0	0.0	0.1
GP characteristics					
Female (%)	54.4 (24.9)	0.6	0.1	0.1	0.0
Specialist (%)	47.4 (25.7)	0.5	0.4	0.3	0.3
Fixed salary (%)	52.3 (31.8)	2.0	1.2	1.5	1.6
Experience (years) (c)	6.4 (2.4)	1.1	0.1	0.1	0.2
Patient characteristics					
Female (%)	49.9 (4.4)	0.0	0.1	0.1	0.2
Age group (years) (%)					
0-19	23.4 (3.5)	0.0	0.5	0.2	0.0
20-64	63.3 (7.4)	0.7	0.4	0.4	0.3
65-74	7.8 (3.8)	0.6	0.0	0.2	0.2
≥75	5.5 (3.0)	0.7	0.2	0.3	0.1
Education level (%)					
No high school	44.2 (2.2)	0.3	0.2	0.0	0.1
High school	30.8 (9.9)	0.2	0.2	0.1	0.1
Higher education	25.9 (10.7)	0.1	0.1	0.1	0.1
Income quartile (%)					
1	26.6 (6.8)	0.3	0.1	0.0	0.2
2	25.3 (4.1)	0.1	0.1	0.0	0.1
3	24.2 (3.7)	0.3	0.0	0.1	0.1
4	23.9 (10.2)	0.1	0.0	0.0	0.1
Disability benefits (%) (d)	6.1 (2.2)	0.2	0.2	0.6	0.1
Sick leave (%) (e)	16.7 (1.8)	0.1	0.0	0.1	0.1
Healthcare utilisation (f)					
GP services	5.2 (1.6)	0.5	0.0	0.0	0.0

	DM	Non-DM			
Emergency department services	0.4 (0.2)	0.3	0.2	0.1	0.0
Outpatient clinic services	3.8 (1.1)	0.1	0.1	0.4	0.1
Hospital admissions	0.4 (0.1)	0.5	0.1	0.7	0.1
Home health nursing (%)	4.3 (1.5)	0.1	0.0	0.1	0.1
Chronic conditions (%) (g)					
Alcohol abuse	0.7 (0.3)	0.4	0.3	0.5	0.2
Medication abuse	0.3 (0.3)	0.5	0.4	0.6	0.5
Substance abuse	0.7 (0.6)	0.3	0.4	0.7	0.3
Depressive disorder	5.1 (2.3)	0.5	0.2	0.1	0.1
Mental retardation	0.3 (0.2)	0.2	0.1	0.1	0.2
COPD	1.6 (0.5)	0.5	0.1	0.0	0.2
Type 2 diabetes	3.4 (1.0)	0.4	0.0	0.0	0.3
Sum		18.3	7.3	8.7	7.4
N	5	817	14	15	367

Note: Column 2 reports the sample mean of observed baseline characteristics for practices that participate in the PCT pilot with the DM. Standard deviation in parenthesis. Columns 4–6 report the absolute standardised mean difference (ASMD) between the DM group and the coarsened exact matching (CEM), propensity score matching (PSM), and propensity score weighting (PSW) control groups, respectively. For reference, the DM group is compared to the non-DM group in the original data in Column 3. The PSW control group only has 367 observations. This is because no DM practice has a centrality of 2 or 3, causing non-DM practices with a centrality of 2 or 3 to be dropped when I include this covariate as a predictor in the logistic regression model.

The results are hopeful, but to adjust completely for confounding, the entire distribution of covariates must be balanced, not just means. Although techniques for comparing covariate distributions are available, they are not applicable because of the size of the DM group. Instead, it is useful to investigate the crucial identifying assumption of DD; that is, that the DM and non-DM groups shared a common trend in outcome before the PCT pilot was announced. The degree to which this assumption holds can be assessed graphically. To this end, Figures 1–3 plot the average rates of three utilisation outcomes for the DM and non-DM groups over five years, three of which were pre-announcement. These figures show that pre-trends in outcomes were similar, suggesting that the common trend assumption is reasonable.⁴

⁴ Interest lies in the pre-period, but it should be noted that the sharp decline in the number of GP services in the post-period (Figure 1) is due to many services not being billable in the DM.

Figure 1: Annual number of GP services for DM and non-DM patients in the original, CEM, PSM and PSW samples.

Figure 2: Annual number of emergency department services for DM and non-DM patients in the original, CEM, PSM and PSW samples.

Figure 3: Annual number of hospital admissions for DM and non-DM patients in the original, CEM, PSM and PSW samples.

6 Discussion

Although team-based care has shown promise in delivering the comprehensive and coordinated care that is expected from modern primary care, the promotion of this type of primary care delivery model should be guided by carefully evaluated trials. A key challenge facing evaluation is selection bias. To account for selection bias in case of the DM, I compare three common matching methods. I find that CEM, PSM, and PSW all produce a control group that resembles the treated group in terms of both observed baseline characteristics and pre-trends in utilisation outcomes. This makes a convincing qualitative argument that either of the control groups are suitable for DD analysis, which suggests using PSW because it uses all available data. However, given the size of the DM group, the power to detect real effects remains a serious concern. Without adequate statistical power, a lack of significant findings would not necessarily indicate that the DM does not work. Instead, the findings may be false negatives. Moreover, DD analysis relies on the interaction between time and treatment status, which is notoriously lacking in power (Strumpf et al., 2017b).

That said, identification in DD analysis often arises from changes in treatment status by a small number of units, and remedies are available. Some remedies aim to reduce the minimum detectable effect (MDE). For example, Peikes et al. (2011) show how measuring healthcare utilisation among high utilisers, such as the chronically ill, greatly reduces MDEs. Redefining healthcare utilisation to be, for example, binary, also reduces MDEs, although possibly at the expense of lower apparent treatment effects. Notably, the concern that linear probability models might predict outcomes outside the $[0,1]$ range does not carry much weight in DD analysis. In contrast, patient-reported measures have low MDEs and should be easily estimated. Other remedies involve the use of alternative inference

approaches. Two prominent examples are Brewer et al. (2018) and Conley and Taber (2011). Both feature simulations that show how their proposed estimators improve power considerably compared to the standard ordinary least squares estimator in DD analysis. A drawback of these remedies is that they require patient-level data. If neither patient-level data nor adequate statistical power is attainable, confidence intervals can still be informative about the presence of clinically or economically important differences between the treated and control groups.

It should also be stated that neither CEM, PSM, nor PSW is able to balance all the observed characteristics. This can occur when the treated group is small because one cannot afford to discard treated units regardless of whether suitable untreated units are available. While some differences in observed characteristics that can be expected to persist over time may be acceptable, as DD adjusts for all fixed differences between groups, such differences may raise concerns about potential unobserved time-varying differences. In this case, one may prefer the synthetic control (SC) method (Abadie et al., 2010), which is a weighted combination of untreated units where the weights are chosen so that the pre-trend in the outcome of the synthetic control mirrors that of the treated unit, thus allowing for unobserved time-varying differences. A practical drawback of this method is that one must fit separate synthetic controls for each treated unit. Moreover, to ensure its credibility, one should restrict the untreated group to units with characteristics similar to the treated unit, have many pre-period data points, and avoid solely optimising the pre-trend in outcome while ignoring other characteristics that influence future outcome values (Kaul et al., 2017). However, if these requirements are met, then this is an attractive method to account for selection bias. Moreover, recent extensions accommodate not only multiple treated units, but also complex designs, such as staggered adoption and treatment reversal (Abadie, 2021).

Ultimately, my findings suggest that matching methods can be used in comparative case studies to account for selection bias when the treated group is very small and distinct. This result is especially useful in situations where the SC method is unfit. Additionally, I identify important predictors of joining the DM. This may be of particular interest to policymakers seeking to increase participation in new team-based primary care delivery models.

References

- Abadie, A., Diamond, A. and Hainmueller, J. (2010). Synthetic control methods for comparative case studies: estimating the effect of California's tobacco control program, *Journal of the American Statistical Association*, 105(490), pp. 493–505.
- Abadie, A. (2021). Using synthetic controls: feasibility, data requirements, and methodological aspects, *Journal of Economic Literature*, 59(2), pp. 391–425.
- Austin, P.C. (2009). Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples, *Statistics in Medicine*, 28(25), pp. 3083–3107.
- Brewer, M., Crossley, T.F. and Joyce, R. (2017). Inference with difference-in-differences revisited, *Journal of Econometric Methods*, 7(1).
- Cenzer, I., Boscardin, J.W. and Berger, K. (2020). Performance of matching methods in studies of rare diseases: a simulation study, *Intractable & Rare Diseases Research*, 9(2), pp. 79–88.
- Conley, T.G. and Taber, C.R. (2011). Inference with difference in differences with a small number of policy changes, *The Review of Economics and Statistics*, 93(1), pp. 113–125.
- EY and Vista Analyse. (2019). Evaluering av Fastlegeordningen. EY and Vista Analyse.

- Friedberg, M.W., Rosenthal, M.B., Werner, R.M., Volpp, K.G. and Schneider, E.C. (2015). Effects of a medical home and shared savings intervention on quality and utilization of care, *JAMA Internal Medicine*, 175(8), pp. 1362–1368.
- Glazier, R.H., Hutchison, B. and Kopp, A. (2015). Comparison of family health teams to other Ontario primary care models, 2004/05 to 2011/12. Institute for Clinical Evaluative Sciences.
- Ho, D.E., Imai, K., King, G. and Stuart, E.A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference, *Political Analysis*, 15(3), pp. 199–236.
- Iacus, S.M., King, G. and Porro, G. (2009). Cem: software for coarsened exact matching, *Journal of Statistical Software*, 30(9), pp. 1–27.
- Iacus, S.M., King, G. and Porro, G. (2012). Causal inference without balance checking: Coarsened exact matching, *Political Analysis*, 20(1), pp. 1–24.
- Jones, C., Finison, K., McGraves-Lloyd, K., Tremblay, T., Mohlman, M.K., Tanzman, B., Hazard, M., Maier, S. and Samuelson, J. (2016). Vermont's community-oriented all-payer medical home model reduces expenditures and utilization while delivering high-quality care, *Population Health Management*, 19(3), pp. 196–205.
- Kaul, A., Klößner, S., Pfeifer, G. and Schieler, M. (2017). Synthetic control methods: never use all pre-intervention outcomes together with covariates. MPRA, 83790.
- King, G. and Nielsen, R. (2019). Why propensity scores should not be used for matching, *Political Analysis*, 27(4), pp. 1–20.
- Krasnik, A. and Paulsen, B. (2009). Reforming primary health care, in Magnussen, J., Vrangbæk, K. and Saltman, R.B. (eds.), *Nordic health care systems. Recent reforms and current policy challenges*. London: Open University Press, pp. 233–254.
- Leuven, E. and Sianesi, B. (2003) PSMATCH2: Stata module to perform full Mahalanobis and propensity score matching, common support graphing, and covariate imbalance testing.
- Mahmud, A., Timbie, J.W., Malsberger, R., Setodji, C.M., Kress, A., Liisa, H., Mendel, P. and Kahn, K.L. (2018). Examining differential performance of 3 medical home recognition programs, *The American Journal of Managed Care*, 24(7), pp. 334–340.
- Mousquès, J. and Bourgueil, Y. (2014). L'évaluation de la performance des maisons, pôles et centres de santé dans le cadre des Expérimentations des nouveaux modes de rémunération (ENMR) sur la période 2009-2012. Irdes, 559.
- OECD. (2017). Primary care in Denmark, OECD Reviews of Health Systems. OECD Publishing.
- OECD. (2020). Realising the potential of primary health care, OECD Health Policy Studies. OECD Publishing.
- Olsen, K.R., Anell, A., Häkkinen, U., Iversen, T., Ólafsdóttir, T. and Sutton, M. (2016). General practice in the Nordic countries, *Nordic Journal of Health Economics*, 4(1), pp. 56–67.
- Peikes, D., Dale, S., Lundquist, E., Genevro, J. and Meyers, D. (2011). Building the evidence base for the medical home: what sample and sample size do studies need? Agency for Healthcare Research and Quality, 11-0100-EF.
- Pines, J.M., Keyes, V., van Hasselt, M. and McCall, N. (2015). Emergency department and inpatient hospital use by Medicare beneficiaries in patient-centred medical homes, *Annals of Emergency Medicine*, 65(6), pp. 652–660.
- Rosenbaum, P. and Rubin, D. (1983). The central role of the propensity score in observational studies for causal effects, *Biometrika*, 70(1), pp. 41–55.
- Rosenthal, M.B., Alidina, S., Friedberg, M.W., Singer, S.J., Eastman, D., Li, Z. and Schneider, E.C. (2016). A difference-in-difference analysis of changes in quality, utilization and cost following the Colorado multi-payer patient-centered medical home pilot, *Journal of General Internal Medicine*, 31(3), pp. 289–296.

- Strumpf, E.C., Ammi, M., Diop, M., Fiset-Laniel, J. and Tousignant, P. (2017a). The impact of team-based primary care on health services utilization and costs: Quebec's family medicine groups, *Journal of Health Economics*, 55, pp. 76–93.
- Strumpf, E.C., Harper, S. and Kaufman, J.S. (2017b). Fixed effects and difference-in-differences, in Oakes, J.M. and Kaufman, J.S. (eds.), *Methods in Social Epidemiology*. Hoboken, NJ: Jossey-Bass, pp. 341–368.
- Svedahl, E.R., Pape, K., Toch-Marquardt, M., Skarshaug, L.J., Kaspersen, S.L., Bjørngaard, J.H. and Austad, B. (2019). Increasing workload in Norwegian general practice – a qualitative study, *BMC Family Practice*, 20(68).
- van Hasselt, M., McCall, N., Keyes, V., Wensky, S.G. and Smith, K.W. (2015). Total cost of care lower among Medicare fee-for-services beneficiaries receiving care from patient-centered medical homes, *Health Services Research*, 50(1), pp. 253–272.

© 2021 by the author(s). This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution license (<http://creativecommons.org/licenses/by/4.0/>).

Appendix

Table A1: Criteria for determining practice allowance.

Criteria	Weight
Age-sex distribution of list patients	0.8155
Quality of life in local municipality	0.0250
Travel distance in local municipality	0.0200
Share of list patients' services in own practice	0.0500
Number of GPs with specialisation in general medicine	0.0600
Number of nurses with Master of Advanced Clinical Nursing	0.0025
Share of COPD and type 2 diabetes list patients with treatment plan	0.0100
Share of COPD and type 2 diabetes list patients with medication review	0.0150
Online appointment system	0.0020

Table reports criteria for determining practice allowance. The importance of each criterion is given by their respective weights.

Table A2: Identification codes for chronic conditions.

Conditions	ICPC-2 code
Alcohol abuse	P15
Medication abuse	P18
Substance abuse	P19
Depressive disorder	P76
Mental retardation	P85
COPD	R95
Type 2 diabetes	T90