Effects of home-based reablement: A micro-econometric approach

Tore Bersvendsen

UiA University of Agder

Doctoral Dissertations at the University of Agder 262

Tore Bersvendsen

Effects of home-based reablement: A micro-econometric approach

Doctoral Dissertation for the Degree *Philosophiae Doctor (PhD)* at the School of Business and Law

University of Agder
School of Business and Law
Department of Economics and Finance
2020

Doctoral Dissertation at the University of Agder 262
ISSN: 1504-9272
ISBN: 978-82-7117-963-2

Printed by 07 Media
Oslo

Acknowledgements

This day has actually arrived. I am writing acknowledgements for my PhD thesis. To be totally honest, I have never ever reflected around the thought of actually finishing a thesis, and the significance of handing in a thesis.

I was never a star pupil at primary and lower secondary school, also not at high school. In fact, I actually struggled in several courses. However, things changed after the University of Agder (at the time named Agder University College) set me up with a speech therapist. After a series of test, the conclusion was straightforward, dyslexia. A diagnosis I have had my entire life, but without knowing it. I do not believe in hiding behind such a diagnosis, but it helps knowing. One realize that trying to gain academic knowledge like all others do, not necessary is beneficial. Find your own way.

I am still not a star student, far from it. However, instead of being demotivated when studying, such a change turned demotivation into curiosity, which later on turned into a PhD thesis.

A PhD project is not done without a significant amount of help and support. I would start by thanking my main supervisor Jochen Jungeilges. Your feedback, support and effort has been highly valued and appreciated. I am very grateful that you always find time in a busy schedule, for believing and trusting in me, and for treating me as a colleague rather then your student. I am also thankful for the help and support from my secondary supervisor Eirik Abildsnes, especially for your calm in stormy times and for always keeping true to your word.

I also want to thank the University of Agder, and especially the Department of Economics and Finance, for support and help whenever I came knocking on your doors. Especially, I would like to thank Kjetil Andersson for countless and stimulating econometric discussions. A special recognition to my fellow PhD students, thanks for sharing both frustration and laughter.

Further, I want to thank the University of Oslo - Department of Health Management and Health Economics for offering me additional office space, so that I could seek new impulses. Over the years I have met a lot of exciting people, particular during courses work. Thanks to the people who have made this possible.

This PhD project would not have been possible without the support in Kristiansand municipality, and especially the former health and social welfare director Lars Dahlen and later on Wenche P. Dehli. Thanks for realizing this public PhD project, for that I am grateful. Thanks to all my fellow colleges whom always have been supporting me along the way. It have meant more than you might have known. Kristin Skjelbred deserves an extra mentioning, thanks for always staying positive on my behalf when I where not.

All my friends have been very important during this process. I know that I have not always been the most chatty and enthusiastic when asked about my research. However, it have meant a lot and very grateful for your support and I am blessed for having you all.

Finally, I pay my gratitude to my family, in particular my parents Tone and Jim alongside my brother Jørn. Thank you Jørn for continued encouragement, motivation and for being my LaTeX-support. Of course, this thesis would not have been possible without my parents. They have always supported and encouraged me to believe that this hold thing could be possible. I am especially thankful for that you both do not know the word “no” when asked for help and, of course, countless leftover dinners over the last year.

Kristiansand, November 2019
Tore Bersvendsen

Contents

Introduction **xiii**

I Empirical evaluation of home-based reablement: A systematic review **1**

I.1 Background . . . 4
I.2 Methods . . . 5
I.2.1 Systematic search . . . 5
I.2.2 Eligibility criteria . . . 6
I.2.3 Selection and categorization . . . 7
I.2.4 Quality assessment . . . 8
I.3 Results . . . 12
I.3.1 Category 1 - Costs and consequences . . . 13
I.3.2 Category 2 - Health benefits . . . 14
I.3.3 Category 3 - Health services usage . . . 16
I.3.4 Assessment . . . 23
I.4 Discussion . . . 27
I.5 Conclusion . . . 32
I.6 References . . . 33
I.A Appendix A: Assessment points given . . . 36

II Absenteeism and care provision: Evidence from a Norwegian cross-sectional dependent panel **37**

II.1 Introduction . . . 40
II.2 Institutional background and sick-leave dynamics . . . 42
II.3 Data collection and description . . . 46
II.4 A linear model for the provision of home-nurse hours . . . 49
II.5 Cross-sectional dependence . . . 49
II.6 Techniques for the correction of standard errors . . . 53
II.6.1 Heteroscedasticity and autocorrelation . . . 54
II.6.2 Heteroscedasticity, autocorrelation and cross-section dependence 55
II.6.3 Fixed-b asymptotics . . . 56
II.7 Ruling out endogeneity . . . 57
II.8 Results . . . 58

II.8.1 Provision of care hours: Estimation results 58
II.8.2 A comparison of HAC correction techniques 60
II.9 Robustness check 63
II.10 Discussion and conclusion 65
II.11 References 69
II.A Appendix A: Variable list 72
II.B Appendix B: Unit-root tests 73
II.C Appendix C: Results - different standard error correction 74

III Creating a unique panel from Norwegian health register data: Technicalities and difficulties 79
III.1 Introduction 82
III.2 Data collection and anonymization 84
III.2.1 Bird's eye view of the data collection process 84
III.2.2 Data storage 91
III.2.3 Anonymization 91
III.3 Imputation strategy 95
III.3.1 Delete extreme suppressed 95
III.3.2 Creating test data 96
III.3.3 IPLOS data 97
III.3.4 NPR and KUHR data 101
III.4 Data summary 112
III.5 Discussion and concluding remarks 114
III.6 References 116

IV The health cost effects of home-based reablement: Empirical evidence from Norway 119
IV.1 Introduction 122
IV.2 The Norwegian health system and home-based reablement 124
IV.3 Data and cost estimation 127
IV.3.1 Data 127
IV.3.2 Cost estimation 129
IV.3.3 Matching strategy 132
IV.3.4 Descriptive statistics 133
IV.3.5 Treatment variables 136
IV.4 Empirical strategy 137
IV.4.1 Regression setup 137
IV.4.2 Cross-sectional dependence 138
IV.4.3 Heterogenous coefficients 141
IV.4.4 Estimation 143
IV.4.5 Potential endogenous regressors and modeling technicalities . . 146
IV.5 Results 147
IV.5.1 Total costs 148
IV.5.2 Primary care costs 150
IV.5.3 Specialist care costs 153

IV.5.4 Gender-specific regressions 154
IV.6 Robustness tests 156
IV.6.1 Different lags, τ, selections 156
IV.6.2 Jackknife bias-correction 157
IV.6.3 Wild bootstrap standard error correction 159
IV.6.4 Sensitivity analysis 159
IV.7 Discussion and conclusion 162
IV.8 References 167
IV.A Appendix A: Variable list 171
IV.B Appendix B: Unit root and co-integration tests 173
IV.C Appendix C: Gender-specific regression results 175
IV.D Appendix D: Result tables for different τ values 180
IV.E Appendix E: Jackknife bias adjustment 189
IV.F Appendix F: Increased price per hour 194

List of Figures

I.1 Flow chart that captures the main steps of our sequential search and selection process 8

I.2 *Total scores for studies by category.* Y-axis represents our defined study categories, and the X-axis represents total scores from our assessment scheme. Maximum possible score = 15. The triangular symbols represent the total score obtained for different studies, whereas the red multiplication sign indicates category means. Total scores is presented in Appendix I.A. 23

I.3 *Total scores for studies organized by country.* Y-axis are countries of the included studies, and the X-axis represents total scores from our assessment scheme. Maximum possible score = 15. The triangular symbols represent the total score obtained for different studies, whereas the red multiplication sign indicates country means. Total scores is presented in Appendix I.A. 24

II.1 Information flow related to NC notes 44

II.2 Monthly sick-leave rates for GR1, CPs and Home-nurse areas (2014m1-2016m12) 45

II.3 $\text{HOURS}_{i\bullet}$ versus area i augmented by grand mean and time averages . 48

II.4 Each dot represents a new case applying for home-care services. When entering GR1, a case is allocated to a caseworker (smileys) by patients's birth day. Thus the allocation appears to be random for the caseworker, and the decisions of each caseworker will be directed to several different home-nurse areas in an unsystematic way 52

III.1 Data flow and connection process 90

III.2 Illustration of the univariate missing pattern. Variables X_1 - X_5 are all observed and used in the anonymization algorithm. All observations marked as suppressed, due to variables X_1 - X_5, can be structured in the bottom (gray area). Suppression automatically deletes service data Y_1 - Y_k. 102

IV.1 Monthly mean total cost development for HBR and non-HBR patients, 2011 - 2015, measured in NOK. The data are based on the complete sample of 155,080 individuals presented in Section IV.3.1. In addition, the figure illustrates roughly when HBR was implemented in the various municipalities considered here. 126
IV.2 Illustrates the four different treatment indicators of interest. All start at the same point, but their time horizons differ. T_{it} does not have a limited time horizon, while T_{it}^{12}, T_{it}^{8} and T_{it}^{6} and are limited at 12, 8 and 6 months, respectively. 136
IV.3 Difference between FE coefficients and individual-specific MG coefficients, δ_i^{Diff} and ρ_i^{Diff}, for model (IV.2). The covariates included in the estimation are Education, five hospital areas, marital status 2 - 5 and quarterly time dummies 143
IV.4 Box-plot of individual $\hat{\delta}_i$ coefficients from model (1) in Table IV.7 for the HBR group. Upper and lower 5% percentiles excluded 163

Introduction

The common theme of this thesis is estimating effects of home-based reablement (HBR) by applying micro-econometrics. Estimating cost related effects of HBR proved challenging, mainly due to specifics of the Norwegian public health setting and the individual tailoring of the HBR treatment. However, important results for the Norwegian public sector have been generated through an immense data job and by applying state-of-the-art econometric techniques. This thesis adds to the growing literature assessing empirical effects of new health interventions. The thesis is funded by Kristiansand municipality and the Research Council of Norway (Grant number: 247076).

HBR is a fairly new intervention, both in Norway and internationally, which aims to restore or increase patients' level of functioning, thereby increasing patients' self-reliance and consequently decrease their dependence on healthcare services (Tuntland et al., 2014; Whitehead et al., 2015). The treatment is not standardized and content may vary. However, all variations have a common goal and share key features as being time-limited, multidisciplinary, home-based, goal-oriented and person-centered. In HBR, significant resources are provided by a multidisciplinary team, focusing on intensive rehabilitation measures in the patients' home for a limited period. The HBR team typically works with the patient towards a goal defined by the patient (Lewin and Vandermeulen, 2010; Tinetti et al., 2012), usually focusing on skills needed for daily living (Francis et al., 2011). The intervention represents an ecological approach, taking into account patients' preferences and resources and it is in line with the chronic care model suggested by Meld. St. 26 (2014–2015) and W.H.O, 2002. The typical somatic HBR patients are elderly people with a functional decline who are not cognitive impaired.

The history of Norwegian HBR starts back in 2009, when the Norwegian government approved *The Coordination Reform — Proper treatment – at the right place and right time*, St. meld. nr. 47 (2008-2009). The reform addressed three major challenges; need for better coordinated services, an aging population with more complex health status, and that existing services had insufficient initiatives aimed at limiting and preventing diseases. One key step for the latter was to change the municipalities' role, trying to fulfill the goals of prevention and early intervention while addressing the needs of patients with chronic diseases. The reform emphasized that early recognition of functional decline followed by immediate rehabilitation measure, could reduce or postpone the need for healthcare services. Building on the Coordination

Reform the Norwegian government later published the report *Future Care*, Meld. St. 29 (2012-2013), where further development of HBR was recommended. These two national reforms initiated a swift shift towards HBR in Norwegian municipalities.

Most Norwegian HBR initiatives were inspired by early results from Denmark (Kjellberg and Ibsen, 2010). A pivotal motivation for Norwegian municipalities was economic consideration due to the forthcoming demographical challenge, in line with a Danish study (Førland and Skumsnes, 2016; Fersch, 2015). Interestingly, a systematic review of peer-reviewed studies testing empirical effects HBR presented in Chapter I, showed that only three studies discussed cost effects. This chapter also argues that existing evidence regarding the effects of HBR is still inconclusive and to some degree scarce, but by no means insignificant.

Studying HBR effects, current efforts tend to rely on randomized controlled trials (RCTs). Many view the RCT approach as the *gold standard* while others argue that such a standard does not exist (Cartwright, 2007). RCTs are desirable because effects are fairly easy to compute, just difference in means between two groups, and therefore results are often easily explained to policy-makers. However, the most important argument in favor of RCTs is that it solves the problem of selection bias. In other terms, the mean effect of all other causes except treatment is identical between groups, so the net difference is zero (Deaton and Cartwright, 2018). Such perfect balance between groups is more likely when the sample size is large, an assumption that seems to be forgotten in many practical cases.

RCT is only informative about the mean treatment effect, but does not identify other features of the distribution, such as the median, a feature often ignored by practitioners (Deaton, 2010). However, conducted properly, with a large sample size, RCTs is undoubtedly able to produce good estimates of the population mean treatment effect. Conducting such a large scale RCT is often not feasible in social science, usually due to cost and ethical considerations. As an alternative to RCT data collection, Chapter III provides a detailed description on how a large unique dataset is created by merging data from several Norwegian health registers. The alternative is not without ethical dilemmas, and alternative costs are described. On the upside, it is beneficial that the dataset would be large, since statistical properties of regression estimators are typically only known for the asymptotic case. The dataset created is large enough to fulfil the assumptions needed to estimate heterogeneous treatment effect (Pesaran and Smith, 1995; Pesaran, 2006), which is ignored in RCTs.

As stated by Angrist and Pischke, 2009, p. 86, *...we believe regression should be the starting point for most empirical projects*, regression is the workhorse in empirical work. The HBR treatment is implemented in a complex setting, it is executed in a person's home and acts as a minor puzzle piece in a comprehensive health system. Such settings can cause regression variables and error term to be cross-sectionally (CS) dependent, reflecting dependence between patients. This would naturally invalidate the Gauss-Markov assumptions, and could potentially bias estimates in traditional regression models. For the sake of reasonable inference, applied statisticians typically try to get around the assumption that error terms have equal variances for all cross-sections by applying `robust` error correction. This type of correction would however not control for CS dependency.

There is a growing literature on regression models designed to tackle the issue of CS dependency. These models mostly vary with respect to the assumption concerning the source of such a dependence (Driscoll and Kraay, 1998; Bai, 2009). The variables in the dataset explained in Chapter III, turned out to be CS dependent, even though they contain data from several stakeholders. Estimating HBR cost effects with CS dependence is the theme of Chapter IV. HBR proved to reduce the monthly health service related costs, and interestingly the largest cost reductions were found for a stakeholder other than the municipalities, showing the importance of including all major health providers when evaluating potential cost gains.

At first glance, the theme in Chapter II differs slightly from the others, but it is still closely related to the topic presented above. Most of the largest populated municipalities in Norway have authorized an administration office to allocate healthcare services. All new applications for most healthcare services are sent to this office that decides about an acceptance or a rejection. This type of offices could be the source for CS dependence explained above. The employees in these offices usually have the same background, collaborate and follow some norm, which would contribute to some kind of connection between patients, *ergo* to CS dependence. Chapter II contains a study from one such office, and investigates the impact of sick-leave on health service provided, also in a CS dependent setting, but with a different twist compared with Chapter IV. In addition, preliminary effects of HBR are also provided.

Summary of chapters

The review presented in Chapter I is joint work with my two supervisors, *Jochen Jungeilges* and *Eirik Abildsnes*. The study presented in Chapter II study is joint work with *Kjetil Andersson* and *Jochen Jungeilges*. Chapter III is also joint work with my main supervisor *Jochen Jungeilges*, and the study presented in Chapter IV is single authored.

Chapter I: Empirical evaluation of home-based reablement: A systematic review

The increasing costs, economic and otherwise, resulting from the increased need for care will have to be shared by a decreasing proportion of employed individuals. Since home-based care is more cost-effective, many high-income countries are actively bolstering a shift from residential care to home-based care as a potentially more effective and financially sustainable approach (Cochrane et al., 2013). The forthcoming challenges will force the healthcare industry to focus more on preventive measures, early intervention, increased use of technology, rehabilitation and healthcare services that are less manpower-intensive, and services that empower senior citizens to self-manage chronic diseases (W.H.O, 2012). HBR, known as restorative care in Australia, the USA and New Zealand, has its origin from the above background. High-quality care is clearly an essential goal in health care services, but future resources are limited, inevitably leading to priority setting and trade-offs (Emmert et al., 2012). Assessing the efficiency and effects of new interventions, including HBR, is therefore crucial.

To our knowledge, there are few comprehensive and systematic overviews of research relating to the effects of HBR, and its therefore the objective of this study to provide such an overview for studies assessing HBR through empirical evaluation. First, we aim to provide a concise summary of relevant existing findings generated in the course of the research process. In addition, we provide a critical constructive assessment of the publications reflecting the extant research. The application of statistical concepts and models plays a central role in the research efforts we analysed. Consequently, our review adopts a dual perspective: the health-economic angle is augmented by a pronounced statistical/econometric viewpoint.

We designed and implemented a sufficiently sensitive search and selection strategy, and to optimize our search design, we relied on the expertise of an experienced librarian. Given the intrinsically multidisciplinary nature of HBR, we needed to extend our search to multiple databases covering the fields of medicine, health care, social work and economics. Eligibility was guided by a predefined list of inclusion and exclusion criteria. Regarding the quality assessment, we saw the need to devise an instrument suitable for the assessment of research papers that are related to the complex topic of HBR. The resulting tool is rooted in the research and publication culture of economics, econometrics, and statistics. The questions in this instrument target four different aspects of a research effort: i) general introduction, ii) data sampling and description, iii) statistics and iv) external validity. Our checklist contains 15 different questions, each checking for a specific attribute of the study scrutinized. Given the binary scaling, the total score of a paper will range from 0 to 15 points.

The largest score observed (7) also constitutes the modal score, and the mean and median total score were estimated as 5.4 and 6.0, respectively. The assessment scheme revealed one specific common pattern among the studies. In our view, none of the papers scrutinized provided sufficient information about the data or the statistics employed. Almost all studies lacked external validity. For data description, statistics, external validity, only three points were given. We do not believe that this evidence is indicative of the quality of the underlying research process. More likely, our findings reflect an established publication standard idiosyncratic to the health and medical journals where these studies were published. Failure to provide full information on data and statistics can create uncertainty in an informed reader.

Twelve articles met our eligibility criteria and where included in the study, five were conducted in Australia, three in New Zealand, three in Norway and two in the US. Physical functioning or independence were the potential benefit categories where we found the most studies, and often these focused on ADL. There is no clear evidence supporting the notion that HBR significantly increases physical functioning. Encouragingly, in the studies that produced no significant difference, HBR tended to lead to superior results on the selected instruments. A general pattern for all different measures and existing evidence regarding the effects of HBR is still inconclusive. However, so far it has not been established that HBR renders negative effects.

Chapter II: Absenteeism and care provision: Evidence from a Norwegian cross-sectional dependent panel

Reflection on the phenomenon of absenteeism from the workplace readily produces a list of non-beneficial correlates that could be interpreted as costs. Such cost might be severe for the employee, family members, co-workers and the health system as such. According to Barmby et al., 2002, the absence rates due to sickness are quite substantial in several industrial countries, and they vary significantly across gender. Moreover, there is empirical evidence that workers in the public sector tend to have more days absent than employees in the private sector (Winkelmann, 1999). The direct cost of absenteeism is found to be substantial, and a Norwegian study (Markussen, 2012) shows that on average, a 1% increase in an individual's sick leave rate reduces individual earnings 2 years later by 1.2 %.

In general, absence can either be valid, if due to sickness, or invalid, if due to shirking. So absenteeism is important since it provides information concerning the determinants of worker behavior (Brown and Sessions, 1996). There exist theoretical positions linking productivity to social activities in the workplace. For instance, Rotemberg, 1994 argues that more (less) opportunities to socialize promote (impede) altruism among co-workers which, in turn, tends to increase (decrease) productivity. Another strand of the theoretical literature emphasizes links between productivity and peer pressure.

This study, first of all, examines the effects of increased sick leave on indirect expenses experienced by a public-health entity. That is, we scrutinize cost associated with absenteeism occurring over and above the direct costs as, for instance, the wage rate. A *unique 36-month panel from a micro environment in a large Norwegian municipality* is utilized to estimate the variation in the amount of additional home-nurse hours provided due to changes in administrative workers sick leave occurring in different organizational units of the care providing entity. Secondly, since we deal with idiosyncrasies of the panel data set at hand, using unique insights into the administrative background of the municipality studied, the current effort constitutes an informative case study in applied econometrics. In particular by concentrating on the effects of techniques designed to correct standard errors of estimates for heteroscedasticity, autocorrelation and *cross sectional dependence* on inference, we touch upon a technical matter that has recently received considerable attention in applied statistics.

The findings concerning the effects of sick leave are in line with theoretical (Pauly et al., 2002) and empirical (Nicholson et al., 2006) findings, implying the cost of work loss will exceed wages if the firm has difficulties substituting the absentee with an equally productive worker. In the case at hand, the administration office has little personnel slack and the typical job responsibilities are fairly specialized. Therefore replacing an absent worker by one of equal competence and productivity is close to impossible. Interestingly, we find the opposite effects of increased sick leave in the home-nurse areas. In contrast to the administrative unit, the management of these areas can rely on a pool of potential substitute nurses temporarily replacing the nurse(s) on sick leave.

Apart from health-care related results, we put forward a case study in applied econometrics. We demonstrate how knowledge about the basic "architecture" of the system designed to order and provide home-nurse hours helps to choose an adequate estimator in the context of our linear model of home-nurse service provision. We argue below that due to key features of the system, we can assume that unknown common factor(s) generating cross-sectional correlation in the variables are uncorrelated with those factors causing correlated errors. As a consequence, fixed effects estimation will guarantee unbiased and consistent parameter estimates. Moreover, we rule out endogeneity by considering the design/organization of the administrative system in place to facilitate home-nurse hours in combination with empirical evidence. And finally, in the context of correcting standard errors in the presence of the triad "heteroscedasticity, autocorrelation and cross-sectional dependence", our study reveals that "no-correction" is no reasonable alternative. We substantiate that simply relying on default correction options of standard software can lead to suboptimal inference with respect to the effects of our variables of interest on home-nurse hours.

Chapter III: Creating a unique panel from Norwegian health register data: Technicalities and difficulties

Over the past decades we have witnessed an explosion in the quantity and quality of data bases holding information about individuals. The public discourse resonating this development has generated a wide spectrum of views on the phenomenon. Portrayed by some as a curse, others see a blessing in our ability to amass individual-level data. For science in general, and for social scientists/economists in particular, such rich datasets provide additional opportunities to answer new relevant economic research questions. In response to the growing availability of individual data, distinct subfields of statistics/econometrics have developed.

The panel-data methodology developed in microeconometrics, for example, provides the methodological backbone of a project which tries to asses an innovative care intervention (home based reablement HBR) offered in Norwegian municipalities. Our effort to address a new research problem in the context of health economics involves a large dataset. The insight into the economic aspects of the care strategy that we provide today might be considered by policy makers in designing an economically viable and sustainable future care sector. We believe that society might benefit from the use of a large set of individual patient data in the case at hand. This paper documents the construction of the dataset we used in the HBR study. Focussing on the various dimensions of the construction process - institutional matters, legal aspects, organizational issues, IT-problems and statistical issues - we hope to provide a guideline/orientation for researchers planning this type of work in the future.

Finally, we intent to provide a unique view of the statistical technique of data imputation as a solution to an apparent dilemma experienced by many researchers. Owners of sensible private information (individual data) have to adhere to legal rules protecting the private sphere of individuals when data are made available to researchers. Therefore data are anonymized. The datasets received by the researcher often allow, at best, for suboptimal statistical inference. Thus the information available for the policy maker will not be optimal. The subsequent decision making may not be in the best interest of society. We demonstrate how statistical imputation techniques can be used to reconstruct a research dataset with the goal to improve the precision of statistical inference.

The dataset we constructed is large in terms of both N and T. It was created based on information held in several separate Norwegian health registers. Norway is in a rather rare situation that key services as schooling, social welfare, and health, are exclusively provided by the public sphere. All national register data is therefor often mandatory, of high quality, and covers almost the entire Norwegian population. These features definitely characterize the four health registers that delivered the data which, after being merged, constituted an unique individual monthly panel covering the time-period from 2011 to 2015.

Legal requirements needed the dataset to be anonymized. In the course of this process some observations had to be deleted or suppressed. An algorithm based on four variables was used to mark observations for suppression, and thereafter the associated service data where deleted for the suppressed observations. We tried to counter the negative statistical effects of data suppression by applying imputation techniques.

The operational characteristics of the imputation techniques to be employed were carefully assessed. All test imputation proved accurate for the mean and standard deviation. The accuracy is so high that for instance on binary variables, the mean and standard deviation is identical up to the fourth decimal. However, all techniques underestimate the frequency of service usages, resulting in an overestimated "within" standard deviation. With a large dataset, it is argued, that it is beneficial to have only one complete dataset. Interestingly, with great variable insight, simple interpolation proved to work remarkably well for ceratin variables. Test results for the latter, showed that one is able to perfectly estimate the true value in 97% of the case with a 50% missing rate. When applied, the imputation strategy is used on 1.9% of the total 7,951,682 observations on 154,839 patients.

Documenting the process of constructing the dataset, we naturally focus on the HBR context. However, the process could be replicated for other new interventions of special interest, like new home-based e-health services. Apparently the value-added of our contribution can be derived from the fact that the process described below can be implemented in a wide range of health care related research problems. Future researchers can benefit from the experiences we outline.

Chapter IV: The health cost effects of home-based reablement: Empirical evidence from Norway

The Western world is facing significant demographic change, and Norway is no exception. For the last decade, the Norwegian population growth rate has exceeded that of other European countries, a trend that will continue in the near future (Leknes et al., 2018). Norwegian population projections indicate that by 2060, every fifth citizen will be 70 years of age or older. As the population of older adults grows, the number of individuals facing age-related diseases, along with the number of diseases occurring simultaneously within the same person, i.e., multimorbidity, will increase (Barnett et al., 2012). In a recent paper by Eckardt et al., 2017, it is shown that health care costs tend to increase with the number of comorbidities. The estimated cost of long-term care to people over 65 years of age will double or triple by 2050 in countries in the Organisation for Economic Co-operation and Development (Oliveira Martins and Maisonneuve, 2007).

In this paper, I examine the individual cost effect of HBR, compared to usual care. The present study is the first large-scale attempt to estimate the cost effects of HBR, including services from all key Norwegian healthcare providers. This is made possible by merging three national health registers in Norway at the individual level, creating a unique monthly panel explained in Chapter III. Estimating the cost effect of HBR based on individual data poses some interesting econometric challenges. Due to the individuality of the treatment, one should allow for heterogeneous effects in a dynamic model because current overall health costs naturally are affected by previous costs. This needs to be addressed in a setting where variables and the error term are cross-sectionally dependent. Since it is difficult to argue that the factors driving the cross-section dependence in the variables and error term are not correlated, standard estimators might contain bias. To avoid this problem, a recent large N and T estimator developed by Chudik and Pesaran, 2015 is used, and an introduction to the estimator is provided.

The results indicate that HBR patients on average have lower costs after treatment than non-HBR patients. For the main treatment variable, the estimated monthly short-run total cost savings varies between -4,875 and -6,007 NOK. This includes the first period of HBR and the entire subsequent period. On average, the mean length of the main binary treatment variable is over 15 months. The short-run effect translates into a predicted long-run cost reduction effect between -4,637 and -6,373 NOK. Moreover, the cost reduction increases as the time horizon between the first HBR period and the observation period increases. The effect is mainly found in specialist health services, which is a different area from where the intervention is implemented. Interestingly, especially for policy-makers, no effect is shown for men when regressions are conditioned on gender. Explaining the gender difference could be of interest for future research. All results pass a series of robustness tests.

The present study is not without limitations. One limitation, is that one are not able to use the function group variables in the regression because of endogeneity issues. Both groups have equal scores the period before the first entry into the service. However, the function scores prior to first entry are based on only the few patients who received some other service, such as a *support person*. Most persons in the sample actually have a zero function score prior to the first entry. The dataset therefore does not contain any good measure of the function level prior to first entry. Lagged costs are potentially the best measure available in the dataset. Costs should reflect service usage, and service usage should to some degree reflect function or sickness level. The model used controls for all unobserved time-invariant factors and potential trends, so a large degree of uncertainty could be captured. However, one cannot completely rule out the possibility that the effects observed above are solely due to HBR itself; they might also capture some unobserved pre-treatment functional differences or other unobserved time-varying changes. Such a change could be motivated by HBR patients changing their motivation for living more healthfully with increased knowledge after treatment. The effect could therefore be caused by a change in the mental state and not due to HBR training per se. Either way, HBR had a positive effect, what causes the effect would just differ.

References

Angrist, Joshua and Jorn-Steffen Pischke (2009). *Mostly Harmless Econometrics: An Empiricist's Companion.* 1st ed. Princeton University Press.

Bai, Jushan (2009). "Panel Data Models with Interactive Fixed Effects". In: *Econometrica* 77.4, pp. 1229–1279.

Barmby, Tim A., Marco G. Ercolani, and John G. Treble (2002). "Sickness Absence. An international comparison". In: *The Economic Journal* 112.480, pp. 315–331.

Barnett, Karen et al. (2012). "Epidemiology of multimorbidity and implications for health care, research, and medical education: a cross-sectional study". In: *The Lancet* 380.9836, pp. 37 –43.

Brown, Sarah and John G. Sessions (1996). "The economics of absence: Theory and Evidence". In: *Journal of Economic Surveys* 10.1, p. 23.

Cartwright, Nancy (2007). "Are RCTs the Gold Standard?" In: *BioSocieties* 2.1, pp. 11–20.

Chudik, Alexander and M. Hashem Pesaran (2015). "Common correlated effects estimation of heterogeneous dynamic panel data models with weakly exogenous regressors". In: *Journal of Econometrics* 188.2. Heterogeneity in Panel Data and in Nonparametric Analysis in honor of Professor Cheng Hsiao, pp. 393 –420.

Cochrane, A. et al. (2013). "Home-care 're-ablement' services for maintaining and improving older adults' functional independence". In: *Cochrane Database of Systematic Reviews* 11.

Deaton, Angus (2010). "Instruments, Randomization, and Learning about Development". In: *Journal of Economic Literature* 48.2, pp. 424–455.

Deaton, Angus and Nancy Cartwright (2018). "Understanding and misunderstanding randomized controlled trials". In: *Social Science & Medicine* 210. Randomized Controlled Trials and Evidence-based Policy: A Multidisciplinary Dialogue, pp. 2 –21.

Driscoll, John C. and Aart C. Kraay (1998). "Consistent Covariance Matrix Estimation with Spatially Dependent Panel Data". In: *The Review of Economics and Statistics* 80.4, pp. 549–560.

Eckardt, Matthias et al. (2017). "Analysis of Health Care Costs in Elderly Patients with Multiple Chronic Conditions Using a Finite Mixture of Generalized Linear Models". In: *Health Economics* 26.5, pp. 582–599.

Emmert, Martin et al. (2012). "Economic evaluation of pay-for-performance in health care: a systematic review". In: *The European Journal of Health Economics* 13.6, pp. 755–767.

Fersch, Barbara (2015). "Expectations towards home care re-ablement in Danish municipalities". eng. In: *International Journal of Sociology and Social Policy* 35.3/4, pp. 126–140.

Førland, Oddvar and Randi Skumsnes (2016). *En oppsummering av kunnskap: Hverdagsrehabilitering.*

Francis, Jennifer, Mike Fisher, and Deborah Rutter (2011). Tech. rep.

Kjellberg, Jakob and Rikke Ibsen (2010). *Økonomisk evaluering af Længst Muligt i Eget Liv i Fredericia Kommune.*

Leknes, Stefan et al. (2018). *Befolkningsframskrivingene 2018. Modeller, forutsetninger og resultater.* Tech. rep. Rapporter 2018/21. Statistics Norway.

Lewin, Gill and Suzanne Vandermeulen (2010). "A non-randomised controlled trial of the Home Independence Program (HIP): an Australian restorative programme for older home-care clients". In: *Health & Social Care in the Community* 18.1, pp. 91–99.

Markussen, Simen (2012). "The individual cost of sick leave". In: *Journal of Population Economics* 25.4, pp. 1287 –1306.

Meld. St. 26 (2014–2015) (2015). *The primary health and care services of tomorrow – localised and integrated.* Norwegian.

Meld. St. 29 (2012-2013) (2013). *Future Care.* Norwegian. Report to the Storting (White Paper) Chapter 1–3.

Nicholson, Sean et al. (2006). "Measuring the effects of work loss on productivity with team production". In: *Health Economics* 15.2, pp. 111–123.

Oliveira Martins, Joaquim and Christine Maisonneuve (2007). "The drivers of public expenditure on health and long-term care". In: *OECD Economic Studies* 2006, pp. 115–154.

Pauly, Mark V. et al. (2002). "A general model of the impact of absenteeism on employers and employees". In: *Health Economics* 11.3, pp. 221–231.

Pesaran, M. Hashem (2006). "Estimation and Inference in Large Heterogeneous Panels with a Multifactor Error Structure". In: *Econometrica* 74.4, pp. 967–1012.

Pesaran, M.Hashem and Ron Smith (1995). "Estimating long-run relationships from dynamic heterogeneous panels". In: *Journal of Econometrics* 68.1, pp. 79–113.

Rotemberg, Julio J. (1994). "Human relations in the workplace." In: *Journal of Political Economy* 102.4, p. 684.

St. meld. nr. 47 (2008-2009) (2009). *The Coordination Reform. Proper treatment – at the right place and right time.* Norwegian.

Tinetti, Mary E. et al. (2012). "Effect of a Restorative Model of Posthospital Home Care on Hospital Readmissions". In: *Journal of the American Geriatrics Society* 60.8, pp. 1521–1526.

Tuntland, Hanne et al. (2014). "Reablement in community-dwelling adults: study protocol for a randomised controlled trial". In: *BMC Geriatrics* 14.1, p. 139.

Whitehead, Phillip J et al. (2015). "Interventions to reduce dependency in personal activities of daily living in community dwelling adults who use homecare services: a systematic review". In: *Clinical Rehabilitation* 29.11. PMID: 25587088, pp. 1064–1076.

W.H.O (2002). *Innovative Care for Chronic Conditions: Building Blocks for Action.* Tech. rep.

W.H.O, Europe (2012). *Strategy and Action Plan for Healthy Ageing in Europe, 2012-2020.* Tech. rep. 122346. World Health Organization.

Winkelmann, Rainer (1999). "Wages, firm size and absenteeism". In: *Applied Economics Letters* 6.6, pp. 337–341.

Paper I

Empirical evaluation of home-based reablement: A systematic review

Empirical evaluation of home-based reablement: A systematic review

Tore Bersvendsen[1*], Jochen Jungeilges[1,2] and Eirik Abildsnes[3]

[1]Department of Economics and Finance, University of Agder, Norway
[2]Institute of Mathematics and Natural Sciences Ural Federal University, Ekaterinburg, Russian Federation
[3]Department of Health Promotion and Innovation, Health and Social Services, Kristiansand Municipality, Norway

Home-based reablement (HBR) aims to restore or increase a patient's level of functioning, thereby increasing the patient's self-reliance and consequently decreasing their dependence on healthcare services. To date, the evidence on whether HBR is an efficient method has not been comprehensively reviewed. The aim of this study was to provide a concise summary of relevant existing findings. In addition, we provide a critical constructive assessment of the publications reflecting the extant research. The relevant literature on this topic was identified through a systematic search of appropriate databases. Thereafter, we screened the studies, first by title, followed by abstract and then by assessing full-text eligibility. A checklist of 15 criteria was developed and used as the basis for the quality assessment. In total, 11 studies from Australia, New Zealand, the USA and Norway were included in the full-text review. The studies reported estimated cost differences between HBR and usual care after the intervention. All the studies indicated lower costs for HBR, but not all of them reported a significant difference. The same pattern was also found for other measures of physical functioning and quality of life. The assessment revealed one specific common pattern: None of the papers scrutinized provided sufficient information about the data or the statistics employed, and all lacked external validity. Some promising results have been reported with respect to HBR reducing the need for specialist or residential care. In short, the existing evidence regarding the effects of HBR is still inconclusive. The findings from the quality assessment should motivate a multidisciplinary approach for future research on HBR.

Keywords: Reablement, economic/econometric evaluation, rehabilitation, RCT technology, assessment tool

JEL classification: I19, C18

*Corresponding author: Tore Bersvendsen, tore.bersvendsen@uia.no

I.1 Background

The western world is facing a significant demographic change in the years to come. These forthcoming developments are expected to lead to a persisting change in the age distribution of the population. As the elderly population grows, the number of individuals facing age-related diseases and multimorbidity will increase (Barnett et al., 2012). Costs of healthcare services increase with age and with the degree of multimorbidity (Yoon et al., 2014). According to Oliveira Martins and Maisonneuve, 2007, the costs of long-term care for people over 65 years old are predicted to double or triple by 2050 in countries belonging to the Organization for Economic Co-operation and Development (OECD). Along with these upcoming demographic challenges, the number of participants in the workforce per senior citizen in OECD countries will decrease (OECD).

The increasing costs, economic and otherwise, resulting from the increased need for care will have to be shared by a decreasing proportion of employed individuals. The upcoming challenges will increase the demand for long-term services as home-based care (Ryburn et al., 2009). Since home-based care is more cost-effective, many high-income countries are actively bolstering a shift from residential care to home-based care as a potentially more effective and financially sustainable approach (Cochrane et al., 2013). Another incentive for this shift is that older people prefer to 'age in place' (Wiles et al., 2011). The forthcoming challenges will force the healthcare industry to focus more on preventive measures, early intervention, increased use of technology, rehabilitation and healthcare services that are less manpower-intensive, and services that empower senior citizens to self-manage chronic diseases (W.H.O, 2012).

Home-based reablement (HBR), known as restorative care in Australia, the USA and New Zealand, is one fairly new way of providing homecare services. The main goal of HBR is to restore or increase a patient's level of functioning, thereby increasing the patient's self-reliance and consequently decreasing their dependence on healthcare services. Even though HBR is not a standardized treatment and the content of HBR varies, all such interventions intend to restore or increase the level of functioning (Tuntland et al., 2014; Whitehead et al., 2015). This type of intervention has gained significant prominence internationally in recent years (Cochrane et al., 2013). HBR represents an ecological approach, taking into account patients' preferences and resources. The main features of being time-limited, multidisciplinary, home-based, goal-oriented and person-centred are homogenous across HBR programmes. Patients are mainly senior citizens with or at risk of functional decline (Aspinal et al., 2016). Typically, a multidisciplinary team works towards a patient-defined goal that focuses on everyday activities important to the patient (Tuntland et al., 2014). When developing an HBR model, Baker et al., 2001 learned that several traditional homecare routines created barriers for functional independence in older patients. They also agreed that involving the patient in goal setting was essential. One might expect that HBR may lead to increased risks of falling, injury or other adverse events. There is evidence that reablement in nursing homes does not have such effects (Gruber-Baldini et al., 2011). We are not aware of similar evidence concerning HBR. A Danish study concluded that policy-makers mainly motivated by economic considerations were pivotal for the implementation of HBR (Fersch, 2015). High-quality care is clearly an essential goal in health care services, but future resources are limited, inevitably leading to priority setting and trade-offs (Emmert et al., 2012). Assessing the efficiency and effects of new interventions, including HBR, is therefore crucial.

To our knowledge, there are few comprehensive and systematic overviews of research relating to the effects of HBR. Interestingly, no HBR studies were included in an overview of systematic reviews on economic evaluations of health-related rehabilitation (Howard-Wilsher et al., 2016). Five HBR studies were included in a systematic review identifying interventions that aimed to reduce dependency in activities of daily living (ADL) (Whitehead et al., 2015). The latter review had two objectives, of which the second was to determine the effect an intervention had on improving a person's ADL. HBR could have effects on factors other than ADL. The two studies most similar to our paper are those by Legg et al., 2016 and "Effectiveness of Reablement: A Systematic Review", both systematic reviews from 2016. "Effectiveness of Reablement: A Systematic Review" examined the effectiveness of HBR and factors that might contribute to successful implementation for Canadian policy makers. They focused on three outcomes, function, health-related quality of life (HRQoL) and service utilization, concluding that there is good evidence supporting the effectiveness of HBR, especially regarding HRQoL and service utilization. Interestingly, Legg et al., 2016 studied whether publicly funded HBR affects patient health or use of services. They found no data suitable for evaluating the effects of HBR and concluded that there is no evidence that HBR fulfils its goals. In sum, previous reviews either focus on minor aspects of potential benefits of HBR alone or do not include studies on HBR, as such studies failed to meet the inclusion criteria defined by the respective reviewers. It is therefore the objective of this paper to provide a comprehensive and systematic review of current literature assessing HBR through empirical evaluation. First, we aim to provide a concise summary of relevant existing findings generated in the course of the research process. In addition, we provide a critical constructive assessment of the publications reflecting the extant research. The application of statistical concepts and models plays a central role in the research efforts we analysed. Consequently, our review adopts a dual perspective: the health-economic angle is augmented by a pronounced statistical/econometric viewpoint.

The remainder of the paper is organized as follows: In Section I.2, we outline the methodological basis for the systematic review. The main findings from relevant HBR research are summarized, and the results of our literature assessment are presented in Section I.3. Section I.4 provides a thorough discussion of the results. A short selection of concluding remarks in Section I.5 finalizes the paper.

I.2 Methods

I.2.1 Systematic search

With the aim of providing a comprehensive systematic review of the relevant scientific literature on HBR, we designed and implemented a sufficiently sensitive search and selection strategy. To optimize our search design, we relied on the expertise of an experienced librarian. Given the intrinsically multidisciplinary nature of HBR, we needed to extend our search to multiple databases covering the fields of medicine, health care, social work and economics. Thus, the search algorithms were applied in the digital databases Scopus, EBSCOhost, CINAHL Plus (with full text), MEDLINE, Academic Search Complete, SocINDEX, Social Work Abstracts, Business Source Complete and Econlit. The development of the search syntax reflects the terminological uncertainty concerning HBR as well as our goal to allow for the location of publications that assess the economic dimension of the care strategy studied. The search results discussed below are based on the string "(reablement OR re-ablement OR restorative W/3 (home OR care)) AND (economic* OR cost* OR evaluation*)", where the sub-command "restorative W/3 (home OR care)" indicated that we were looking for instances in which either the term "home" or the term "care" can be found within a three-word-neighbourhood of the term "restorative". No search filters were applied, and the same search string was used on all databases. The initial search was performed on 2016-03-17. It resulted in a total of 554 records. Consecutive updates were run on 2016-08-03, 2017-11-15 and 2019-09-04. All databases were searched on the same search date. Figure I.1 shows the main steps of our sequential search and selection process.

I.2.2 Eligibility criteria

While the first stage of the literature search relied on algorithms, the second stage involved the authors functioning as "human classifiers". Our work was guided by a predefined list of inclusion and exclusion criteria. A study qualified for inclusion if it (i-1) contained at least a partial evaluation on some quantifiable economic measure, both direct and indirect, of HBR, i.e., concepts like "effectiveness", "benefits" and "costs" of the treatment were considered, and (i-2) was published in a peer-reviewed academic journal. To maintain focus on the HBR intervention and to consider only original research satisfying reasonable design standards, two sets of exclusion criteria were defined. While the first set centres around characteristics of the intervention itself (related to particular branches of medicine and institutional/organizational aspects), the second set refers to features of the respective research study. Specifically, we agreed to exclude studies of reablement (e-1) closely linked to dental health or paediatrics or (e-2) provided by and in hospitals or nursing homes. Moreover, an article was excluded if (e-3) it could be classified as a "conceptual article", "review article" or "research protocol", or if (e-4) it did not assess well-defined comparator interventions, as traditional care or other. Titles, abstracts and full texts were checked against the inclusion and exclusion criteria by at least two authors independently.

I.2.3 Selection and categorization

One reviewer (TB) organized and carried out the initial search and eventually removed duplicates in sporadic coordination with the co-authors. Following this initial stage, a stepwise elimination procedure based on (e1)-(e4) was performed. First, two reviewers (TB, JJ) collaborated to filter records by keywords appearing in the title and the journal name. The keywords used for this purpose were "dental", "dentist", "caries", "children", "oral", and "surgery". For all matches, titles were screened and records removed if required. In the second stage, two reviewers (TB, JJ) independently screened the remaining titles. In almost 80% of those cases, the reviewers came to a unanimous decision. As a rule, a split decision lead to inclusion of the article in question. In the following stage, all reviewers independently screened the abstracts of all remaining records before discussing full-text eligibility. Subsequent searches and elimination exercises, i.e., those referred to as "updates" in Figure 1, followed an analogous procedure with different roles assigned to the reviewers. During the update(s), one reviewer (TB) performed the filtering process on all new titles. Subsequently, the reviewers (TB, EA) screened the remaining titles for abstract eligibility. While the first update lead to the inclusion of five new records, no additional articles could be identified during the second update. Next, each reviewer independently read and analysed the articles identified in the previous stages to decide on their full-text eligibility. Finally, following a discussion, the team of reviewers reached a consensus on the pool of studies to be included in this review.

We chose to categorize the final included studies. The categorization of Emmert et al., 2012 is constructive, since, in addition to providing an overview of the types of studies included in this review, it also generates the structure for the presentation of our results in Section I.3. Studies that focus on cost and other consequences regarding economic efficiency were grouped into Category 1. If studies analysed the impact of HBR on both cost and consequences but did not clearly differentiate between the two, they were still included in this category. Studies evaluating health benefits for patients were placed in Category 2. Category 3 includes articles that assess the consequences of HBR on health-service usage. One could argue that reduced service usage is beneficial for the patient and should therefore be a part of Category 2. However, an effect on service usage could also have a direct monetary effect. The latter argument motivates the consideration of an additional category. Studies with multiple outcome measures were categorized by their primary outcome measure.

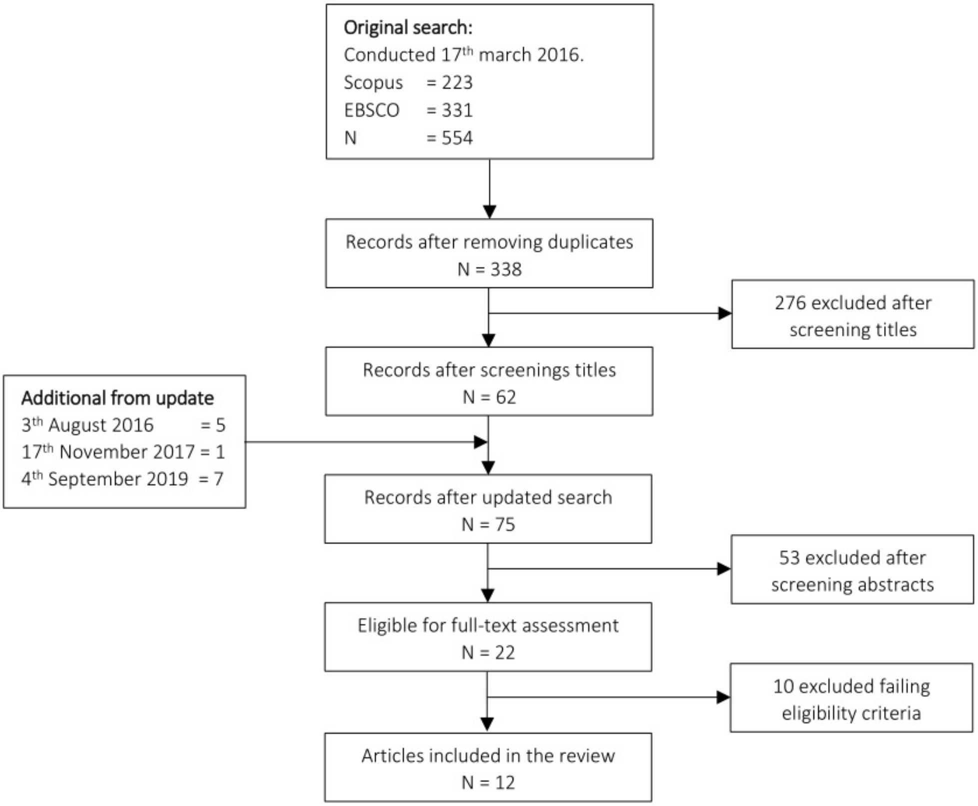

Figure I.1: Flow chart that captures the main steps of our sequential search and selection process

I.2.4 Quality assessment

A 35-point checklist from Drummond and Jefferson, 1996 is often used as a quality assessment tool for health-economic evaluations. These guidelines contain 10 sections under three headings: study design, data collection and analysis, and interpretation of results. A similar instrument has been proposed by Drummond et al., 2005, p. 27. In our view, these tools are particularly suited for the assessment of classic economic evaluation studies focusing on cost-effects analysis. Not all of the studies included in our review were designed for this purpose. Therefore, we saw the need to devise an instrument suitable for the assessment of research papers that are related to the complex topic of HBR. The resulting tool (c.f. Table I.1) is rooted in the research and publication culture of economics, econometrics, and statistics. Those areas of expertise are represented by two of the reviewers (TB, JJ). The questions in this instrument target four different aspects of a research effort: i) general introduction, ii) data sampling and description, iii) statistics and iv) external validity. By describing the scheme of this instrument to the reader, we will allow our review process to be transparent and comprehensible. Moreover, the instrument serves as the basis of a rough scoring scheme applied to the papers included in this review. Our checklist contains 15 different questions, each checking for a specific attribute of the study scrutinized. Given the binary scaling (yes (1) or no (0)), the total score of a paper will range from 0 to 15 points. The scores reflected in the remainder of the paper were generated in a 2-stage process. After independently scoring the HBR studies, two reviewers (TB, JJ) discussed their decisions and agreed on a final score.

Prior to presenting results, a caveat should be mentioned. The scores described should be interpreted in a sensible manner. A low score is not necessarily indicative of a low quality of the research reported; it may just reflect the fact that the article assessed by means of our instrument came from an area of science in which publication standards differ from those in economics, econometrics and statistics or was published in a journal that does not emphasize certain standards to reach a specific clientele. It should be evident that our tool is not intended to denigrate the valuable work of our colleagues but instead is of a constructive nature, which should be emphasized. Use of this tool is meant to support the transparency of our review process and help to reveal differences in the publication culture across disciplines. In that sense, the tool might produce a rough indication regarding the state of HBR literature.

Table I.1: The assessment scheme

	No	Question	Comment
General introduction	1	*Does the research question identify the outcome(s) of interest?*	A well-formulated question clearly identifies the type of effects the study seeks out to investigate. This clearly informs the reader about the aim of the study.
	2	*Does the research question identify the treatment alternatives being compared?*	The research question should identify the alternative treatment(s) being compared. It will inform the reader about the relevant sphere of the health services studied.
	3	*Are the important stakeholders identified?*	Health systems vary across countries. One factor that causes this variation is the incorporation of different stakeholders. The reference to the major stakeholders in the study context makes the reader aware of a key institutional characteristic of the health system under scrutiny. Important stakeholders include patients, health service providers, institutes, insurance companies, municipalities and governments.
	4	*Is the study context clearly defined?*	The motivation for and the background of the study should be understood by the reader. This includes the motivation for implementing and testing HBR.
	5	*Are the alternatives being compared clearly described?*	Detailed descriptions of the treatment alternatives will enable the reader to comprehend the typical service profiles provided. The natural baseline for HBR is 'usual care' which will vary across individuals, and therefore cannot be described in every case. We expect that the authors provide a clear description of the most common service provided to the reference group.
	6	*Are methods for evaluating health states and other benefits described?*	The readers should be able to understand all methods used for evaluation. Authors should not assume that every reader is familiar with all existing instruments for measuring health. It is therefore essential that a short description of each instrument is provided.

Continued on next page

Table I.1 – *The assessment scheme - Continued*

	No	Question	Comment
Data sampling and description	7	*Are the necessary scales for the methods used described?*	A description of an evaluation method is incomplete without information about the scales of the methods. Only a reader who knows the scaling will be able to fully comprehend and appreciate the results of the study.
	8	*Is the sampling procedure clearly described?*	The sampling procedure should be described in detail. If different instruments where used, then the interview setting should be described. Are data extracted from different databases, then the content of each source should be described. Authors should provide details of the dataset used and inform the reader on aspects such as timeframe, frequency, number of observations etc.
	9	*Does the paper provide a clear data description?*	All empirical economic papers should provide a table of descriptive statistics and describe the data based on the table. Providing a table of baseline descriptive statistics without describing data and findings is not sufficient.
Statistics	10	*Is the choice of statistical methods used discussed and justified?*	Statistical models are based on assumptions, that implies that they have strengths and weaknesses. Most of these models are designed for different settings and types of data. As the choice of statistical methods could have a direct influence on the results, authors should explain their choice of methods based on the sampling procedure and their research setting.
	11	*Are assumptions underlying statistical methods used discussed and addressed explicitly?*	Typically, the operational characteristics of statistical methods are known conditional on a set of assumptions being fulfilled. A violation of such assumptions might seriously affect statistical validity. Ramifications in the context of the study should be discussed and addressed whenever possible.
	12	*Are alternative statistical estimators discussed?*	The researchers should carefully motivate the statistical estimators used. Notably they should reflect the main drawbacks of potential alternative estimators in the specific research setting.
	13	*Is the data analytical part of the study replicable?*	Replicability is regarded as an important requirement for studies published in the field of economics. Given the dataset and the methodical description in the paper, an independent researcher should be able to replicate the results.

Continued on next page

Table I.1 – *The assessment scheme - Continued*

	No	Question	Comment
External validity	14	*Has the study a large degree of external validity?*	External validity is essential for a policy-maker who is considering the implementation of HBR. Several studies are not designed for providing information beyond their study setting. Studies with low degree of external validity should not be the basis of a policy-makers' decision. Studies lacking external validity may still be interesting in terms of learning about and developing HBR.
	15	*Is the study grounded in the relevant theory?*	Research procedures are often designed to reveal treatment efficacy rather than unsheathe the causes leading to the efficacy of a treatment. Deaton (2010) argues that RCTs focusing on "whether", are not informative about the mechanisms that cause a treatment to work. He suggests that learning about theory, or mechanisms, requires that the investigation should be targeted towards that theory. Studies, RCTs or non-experimental, that are not theoretically grounded are unlikely to provide any external validity.

I.3 Results

The search strategy identified 338 potentially relevant studies after discarding duplicates. The fulltexts of 22 studies were assessed after screening titles and abstracts. The 12 articles that met our eligibility criteria are presented in Table 2. Three studies were associated with Category 1. Two of these evaluate cost and consequences separately. Five studies were assigned to Category 2. Category 3 included four studies, but three of them had secondary outcome measures that fit Category 2. Of the 12 studies identified, five were conducted in Australia, three in New Zealand, three in Norway and two in the US. All studies included in this review had a maximum intervention period of three months. The Australian HBR model specifically targets patients with low to medium levels of need (Lewin et al., 2008), whereas the HBR interventions in other cases target frailer, older patients on the verge of residential care (Senior et al., 2014). The other studies included did not have a directly specified target group in terms of needs. In all reviewed studies, the multidisciplinary teams were composed of a physiotherapist, occupational therapist and a nurse. One of the team members functioned as a care manager for each client (Lewin et al., 2016).

For data synthesis, a narrative qualitative synthesis of the eligible studies was executed. HBR is a personalized intervention and studies included have a heterogeneous range of outcome measures, we were therefore unable to do a meta-analysis.

I.3.1 Category 1 - Costs and consequences

Kjerstad and Tuntland, 2016 carried out a cost-effectiveness analysis (CEA) of HBR using data from the first randomized controlled trial (RCT) on HBR conducted in Europe (Tuntland et al., 2015). The recruited sample consisted of 61 participants randomized to HBR (n = 31) or usual care (n = 30), but the CEA was conducted on a sample of 46 participants (HBR = 25 and control = 21). All participants were fully assessed at baseline, 3 months and 9 months. Self-perceived activity performance and satisfaction with performance were chosen as effectiveness measures. Cost data were based on individual registrations of the number of home visits, duration of each visit and profession of service delivered. There was no significant difference in the mean cost per participant during the intervention period (3 months), but the HBR group had, on average, fewer but longer visits compared to the control group. At the 9-month follow-up (6 months post-intervention period), the authors found a significant difference in mean cost per visit in favour of HBR. However, the difference of 1.5 €[1] (14.7 NOK) was modest. There was no statistically significant difference in mean cost per participant. The mean changes between baseline, 3- and 9-month follow-up for both effectiveness measures were significant. The incremental cost-effectiveness ratios for the intervention period were -89.5 €[1] (-868.2 NOK) for the activity performance measure and -68.7 €[1] (-666.3 NOK) in terms of satisfaction with performance.

Using data from an Australian RCT (Lewin et al., 2013a), Lewin et al., 2014 examined the use of healthcare services and the associated costs of HBR compared to conventional care. Participants were compared at baseline and after 1- and 2-year follow-ups. Seven hundred fifty participants were included, with 375 in each group for the intention-to-treat (ITT) analysis. For the actual treatment (AT) analysis, 310 participants were included in the HBR group, 395 were in the control group, and 45 participants were excluded. The mean homecare costs per participant were different over the first year and the overall 2-year period in favour of the HBR group. The differences were 959 €[2] (745 €[2]) and AU$1,511 (AU$1,174) for the AT (ITT) analysis after the first year and 1,886 €[2] (1,613 €[2]) and AU$2,971 (AU$2,541) overall. A significantly lower proportion of HBR participants compared to conventional care patients were approved for residential or equivalent homecare at the end of the study. The HBR group had a 30% reduced risk for emergency department presentation in the AT analysis. Over the 2-year period, the AT (ITT) analysis indicated that total costs per participant for all hospital admissions were 825 €[2] (194 euro[2]) and AU$1,299 (AU$306) lower for HBR participants than control patients. Additionally, the HBR group had a reduced risk for unplanned hospital admission in the AT analysis. Over the 2-year period, the mean aggregated cost per participant was lower for the HBR group, and the difference was 1,821 €[2] (AU$2,869) in the ITT analysis and 2,754 €[2] (AU$4,338) in the AT analysis. The HBR group was significantly less costly in the first year and over the total 2-year period in the AT analysis only. The results for the

[1]Exchange rates from 06.02.2018. Collected from the Norwegian national bank, 1 €= 9.7005 NOK.

[2]Exchange rates from 06.02.2018. Collected from the Reserve bank of Australia, 1 AU$ = 0.6348 €.

second year alone did not show a significant difference. Randomization of participants was compromised, and there was some measurement bias in hours of service.

In a retrospective study, Lewin et al., 2013b investigated whether individuals using HBR reduced their need for ongoing services and had lower homecare costs compared to those receiving usual care. By linking several data sources, the authors created a dataset with 10,368 individuals and a time period of 57 months. The individuals received usual care or either of two different HBR versions. In the first HBR version the patients were referred from the community, and in the second version patients were discharged from the hospital. For the second HBR version, the maximum intervention period was 8 weeks and not the standard 12 weeks. The need for ongoing services was measured by a binary yes/no variable and used as an outcome variable in a regression framework. HBR users referred from either the community or a hospital were less likely to use ongoing services over the first 3 years compared to those getting usual care. This effect persisted over the whole time period for HBR users who were referred from the community. Quantile regression was used when making cost comparisons at 3, 12, 24, 36, 48 and 57 months. The costs for both HBR groups were substantially less than that for conventional care over the observation period. The median savings per HBR participant after 57 months amounted to more than 7,935 €[2] (AU$12,500) in both HBR groups.

I.3.2 Category 2 - Health benefits

A cluster RCT conducted in New Zealand by King et al., 2012 examined the impact of HBR versus usual care and applied HRQoL as the primary outcome. The following secondary outcomes were included: functional mobility, sense of control and social support network. All outcome data were collected at baseline and at 4- and 7-month follow-ups with structured face-to-face interviews. In total, 186 participants were included at baseline with 93 participants in each group. At the final 7-month assessment, 157 participants remained, 76 in the HBR group and 81 in the control group. HRQoL was measured by the 36-Item Short Form Health Survey (SF36[3]), an instrument that generates an overall score between 0 and 100, with larger numbers indicating better HRQoL. The instrument also provides separate mental and physical subscores. After adjusting for baseline demographics, the SF36 overall score differences were statistically significant at the 10% level in favour of the HBR group. The mean difference in SF36 score from baseline to 7 months was 3.8. Splitting the SF36 into the two different components indicated significant results for the mental subscore only. This suggests that HBR may improve HRQoL. For all the secondary outcomes, no evidence for significant differences between the groups was found.

The study by Lewin and Vandermeulen, 2010, which used data collected from 2001 to 2003, is the first Australian evaluation study included in this review. Using a non-randomized design, they investigated whether HBR participants had better personal and service outcomes compared to those receiving usual care. Data were collected manually with standardized outcome measures of functional independence,

[3] https://www.rand.org/health/surveys_tools/mos/36-item-short-form.html

confidence and well-being. All participants were assessed at baseline, 3 months and 1 year. One hundred participants were included in each group at baseline. At the 1-year follow-up, there were 67 HBR participants remaining and 73 participants in the control group. The HBR group scored significantly worse on all measures at baseline compared with those getting usual care. At both follow-ups, the HBR group showed improvements in all measures, whereas the participants receiving usual care remained mostly the same. These differences were significant when examined by the Mann-Whitney U-test. The regression analysis also confirmed these results for all measures except the Philadelphia Geriatric Morale Scale[4]. HBR participants also had a statistically significant decrease in the probability of needing ongoing services. The latter analysis was adjusted for baseline differences between the two groups. The authors pinpointed three major limitations: some potential selection bias, a lack of independent data to confirm the service outcomes and a lack of clinical information.

Parsons et al., 2013 used a clustered RCT to determine whether HBR improves physical functioning and social support compared to standard care. The study included 205 participants at baseline, and 197 remained at the 6-month follow-up (106 HBR patients and 91 traditional care patients). The researchers who completed the assessments were experienced and were blinded to group allocation. Physical functioning was measured by the Short Physical Performance Battery (SPPB[5]). The SPPB test contains three elements: standing balance, timed walk and timed rising/sitting from a chair. The results were interpreted conservatively, because the p-values were not corrected for multiple testing. Therefore, a 1% significance level was used in the primary analysis, and all evaluations followed the ITT principle. The HBR group had a significantly greater mean increase in overall SPPB score and in the walk component over time compared to the usual care group. There was no difference between the two groups in the balance or chair-stand components. Social support showed no difference over time between the two groups. In addition, there was no evidence for a significant relationship between allied health referrals and improvement in physical functioning over time. The authors argue that there is considerable ambiguity in determining whether a clinically meaningful change in physical function can be associated with HBR.

Tuntland et al., 2015 carried out the first RCT on HBR in Europe. The goal was to evaluate the effect of HBR compared to usual care on self-perceived activity performance and satisfaction with performance. Secondary outcomes were physical functioning and HRQoL. Sixty-one participants were randomized to HBR or usual care, and assessments were done at baseline and at 3- and 9-month follow-ups. At the last follow-up, 25 participants remained in the HBR group and 26 in usual care. The main outcome was measured by the Canadian Occupational Performance Measure (COPM[6]), and all analyses followed the ITT principle and used a significance level of 5%. There was a significant mean difference in COPM-Performance at both the 3- and 9-month follow-ups. For COPM-Satisfaction, there was only a significant mean difference after 9 months. All results were in favour of HBR, and all analyses

[4] `https://www.ncbi.nlm.nih.gov/pubmed/1109399`
[5] `https://www.nia.nih.gov/research/labs/leps/short-physical-performance-battery-sppb`
[6] `http://www.thecopm.ca/`

were adjusted for baseline values. The 9-month mean difference was 1.4 for both performance and satisfaction, which is below the cut-off value of 2. According to the COPM manual, this indicates a clinically relevant change. The authors acknowledge this value but also argue that there is a lack of evidence supporting this cut-off value. All the secondary outcomes were insignificant after 3 and 9 months. The study constraints rendered it inevitable that the same healthcare personnel provided services to both groups.

Langeland et al., 2019 presents the results of a multicentre, clinical controlled trial involving 47 municipalities in Norway. Primary outcome was self-perceived activity performance and satisfaction with performance measured with COPM, and the ITT principle was followed for all analyses. At baseline, 707 participants were in the HBR group and 121 participants received usual care. The remaining participants at 12-month follow-up was 294 and 54 respectively. Significant mean effects were found in favour of HBR on COPM-Performance and COPM-Satisfaction, both at 10 weeks and 6-moth follow-up. The 6-month mean difference was 1.4 for both COPM-Performance and COPM-Satisfaction. A series of secondary outcomes was measured with different instruments. Physical function, measured with SPPB, showed significant treatment effect in favour for HBR at both 6- and 9-month follow up. Health-related quality of life was measured with The European Quality of Life Scale (EQ-5D-5 L[7]), which contain a questionnaire and a visual analogue scale. The latter test showed significant treatment effect in mobility, personal care, usual activities and current health at the 6-month follow-up. Sense of coherence, measured with Sense of Coherence Questionnaire[8], showed at 6-months follow-up significant effect in favour of the HBR group. Interestingly, all measures, except SPPB, were insignificant at the 12-month follow-up using a significance level of 5%.

I.3.3 Category 3 - Health services usage

An Australian RCT carried out by Lewin et al., 2013a investigated whether individuals receiving HBR had less need for ongoing services compared to those getting usual care. In the follow-up study by Lewin et al., 2014 that is based on the same RCT data, the main outcome was a binary variable (yes/no) representing the need for ongoing personal care services. Data were collected at baseline and at 3 months and 12 months. The study also included secondary outcomes by examining functional status and quality of life (QoL) in a subgroup recruited within the RCT group. For the AT (ITT) analysis, 294 (300) participants were recruited to this subgroup at baseline. At the 12-month follow-up, 192 (198) participants remained, and 100 (88) of these received HBR. Using logistic regression adjusted for baseline covariates, HBR was found to significantly reduce the probability of using ongoing services. These results hold for the ITT and AT analyses in both follow-ups. Regarding functional status, there was a significant difference between the groups at the 12-month follow-up. Functional dependency increased for the usual care group between the 3- and 12-month follow-ups but was maintained in the HBR group. The latter results were adjusted

[7] `https://euroqol.org/eq-5d-instruments/eq-5d-5l-about/`
[8] `https://link.springer.com/chapter/10.1007/978-3-319-04600-6_12`

for baseline covariates and were only significant in the AT analysis. QoL showed no significant difference between groups.

Using an RCT design, Senior et al., 2014 examined whether HBR participants reduced their need for permanent residential care over a 24-month period. The study also included secondary outcomes focusing on functional and social health, which were measured at the 18-month follow-up. Patients received HBR either at home or in a short-term facility, and results were not presented separately for the different locations. A total of 105 participants were recruited, with 52 in the HBR group and 53 that received usual care. Only 17 participants were included in the 18-month follow-up, with 11 in the HBR group and 6 receiving usual care. All patients included were at high risk of residential care placement. Research assistants performed randomization and data collection. Data were collected on a laptop either at the older person's residence or at the respective short-term care facility. The ITT principle was used in all analyses. A Cox proportional hazard model with covariates was used for the primary outcome. For the combined primary outcome of death or residential care, there were no statistically significant results. The insignificant result was a 24% reduction in favour of HBR regarding the probability of residential care or death. Additionally, all the secondary outcomes showed no statistically significant differences after 18 months. The authors also argued that the lack of blinding constituted a limitation.

Tinetti et al., 2002 investigated functional status and the likelihood of remaining at home for persons receiving HBR versus usual care in a real-world setting. This controlled clinical trial compared usual care with HBR in areas like functional status, likelihood of remaining at home, duration and intensity of the homecare episode, emergency visits to a physician and emergency department (ED) visits. There were 691 HBR users included, and from a pool of potential control participants, 691 pairs were created using an algorithm based on several covariates. A subset of 382 pairs was created for patients remaining at home after the completion of either HBR or usual care. Data on functional outcomes were only available for this subset. The descriptive and outcome data used were based on patient records. HBR patients were significantly more likely to remain at home after completion of the homecare episode. The study showed no significant difference in the likelihood of visits to a physician's office. HBR patients were less than half as likely to have an ED visit during the homecare episode. Patients in the HBR group had significantly shorter homecare durations compared to those getting usual care. All these results were adjusted for baseline covariates. Discharge scores for self-care, home management and mobility were all slightly significantly higher for HBR users.

Tinetti et al., 2012 aimed to analyse the frequency of hospital readmissions for HBR compared to usual care after an acute hospitalization. Based on data from the original clinical trial study (Tinetti et al., 2002), 864 participants were admitted to homecare after discharge from an acute hospital stay and were therefore eligible for this study. In total, 770 participants were included, comprising 341 matched pairs and 88 additional unmatched participants. Outcome variables were hospital readmission and length of homecare episode. The algorithm previously used in Tinetti et al., 2002 was applied to generate matched pairs. All descriptive and outcome data came from the patient records of the original study. The main outcome variable, hospital readmission, was measured using a binary (yes/no) scale. The mean length of homecare episodes was significantly different between the two groups. The HBR group mean length was shorter than that of the control group. According to a conditional logistic regression analysis, HBR participants were 32% less likely to be readmitted than participants receiving usual care in the matched pair analysis. For the unmatched analysis, the corresponding result was 29%. The statistical significance was only marginal, with p-values for the matched and unmatched analyses of 0.10 and 0.09, respectively.

Table I.2: Characteristics of the included studies

Author	**Country**	**Design**	**Year of study**	**Participant - inclusion criteria**	**Participant - inclusion criteria**	**Data source**	**Category**	**Score**
Kjerstad and Tuntland, 2016	Norway	RCT	May 2012 to Feb 2014	Applying or referred for homecare service (18+) and had functional decline in one or more daily activities of living (ADL)	Not able to understand Norwegian, were in need for residential care or rehabilitation, terminally ill, or moderately or severely cognitively reduced	Face-to-face interviews, time registration manually	1	7
Lewin et al., 2014	Australia	RCT	Jun 2005 to Aug 2007	Assessed and eligible for homecare service (65+) due to ongoing difficulties with ADL and referred for personal care	Not able to communicate in English, require acute or post-acute care, known diagnosis of dementia or terminal illness, complex care need requiring 15+ hours of homecare a week	Linking data from several databases	1	5
Lewin et al., 2013b	Australia	RLS	Jan 2004 to Dec 2008	Assessed and eligible for homecare service (65+) due to ongoing difficulties with ADL. Referred from the community or discharged from hospital	Not able to communicate in English, known diagnosis of dementia or receiving palliative care	Linking data from several databases	1	7
King et al., 2012	New Zealand	RCT-cluster	Dec 2005 to May 2007	Received homecare assistance (65+) from the agency included in the study	Unable to participate in the interview due to physical and mental health condition	Face-to-face interviews	2	6

Continued on next page

Table I.2 – *Characteristics of the included studies - Continued*

Author	**Country**	**Design**	**Year of study**	**Participant - inclusion criteria**	**Participant - inclusion criteria**	**Data source**	**Category**	**Score**
Lewin and Vandermeulen, 2010	Australia	PLS	2001 - 2003	Assessed and eligible for homecare service (60+) due to ongoing difficulties with ADL and referred for domestic or personal care	Not able to communicate in English, require acute or post-acute care, known diagnosis of dementia or other progressive neurological disorders	Client database and data collected at home	2	6
Parsons et al., 2013	New Zealand	RCT-cluster	Sep 2007 to May 2008	Community-dwelling, new referrals for homecare (65+)	Severe cognitive impairment and referral for assessment for admission to residential care, care support, or short-term service	Not specified	2	2
Tuntland et al., 2015	Norway	RCT	May 2012 to Feb 2014	Applying or referred for homecare service (18+) and had functional decline in one or more daily activities of living (ADL)	Not able to understand Norwegian, were in need for residential care or rehabilitation, terminally ill, or moderately or severely cognitively impairment	Face-to-face interviews in patient's home	2	7
Langeland et al., 2019	Norway	MCT	Apr 2014 to Dec 2015	Home-dwelling (18+), understood Norwegian and recently experienced functional decline	Were in needed for institution-based rehabilitation or nursing home placement, or terminally ill or cognitively impaired	Face-to-face interviews and questionnaire	2	7

Continued on next page

Table I.2 – *Characteristics of the included studies - Continued*

Author	**Country**	**Design**	**Year of study**	**Participant - inclusion criteria**	**Participant - inclusion criteria**	**Data source**	**Category**	**Score**
Lewin et al., 2013a	Australia	RCT	Jun 2005 to Aug 2007	Assessed and eligible for homecare service (65+) due to ongoing difficulties with ADL, and referred for personal care	Not able to communicate in English, require acute or post-acute care, known diagnosis of dementia or other progressive neurological disorders, or receiving palliative care	Client database and data collected at home	3	7
Senior et al., 2014	New Zealand	RCT	Nov 2003 to Jun 2006	Assessed by hospital clinical team or regional geriatric assessment and having a high risk for institutionalization (65+)	Requiring immediate permanent residential care or unable to communicate in English	Data collected at home or residence	3	0
Tinetti et al., 2002	USA	PLS	Nov 1998 to Apr 2000	Person at risk of functional decline after acute illness or hospitalization (65+), but with potential for maintaining or improving function, and receiving homecare	Severe cognitive impairment, requiring total assistance with care and not bedridden	Client database	3	6

Continued on next page

Table I.2 – *Characteristics of the included studies - Continued*

Author	**Country**	**Design**	**Year of study**	**Participant - inclusion criteria**	**Participant - inclusion criteria**	**Data source**	**Category**	**Score**
Tinetti et al., 2012	USA	PLS	Nov 1998 to Apr 2000	Person at risk of functional decline after acute illness or hospitalization (65+), but with potential for maintaining or improving function, and received homecare	Severe cognitive impairment, requiring total assistance with care and not bedridden	Client database	3	5

Abbreviation: RCT = Randomized controlled trail; RCS = Retrospective longitudinal study; MCT = Mulitcenter controlled trial; PLS = Prospective longitudinal study

I.3.4 Assessment

Next, we discuss the results rendered by the application of our assessment tool (cf. Table I.1, in Section I.2.4) to the papers reviewed above. The detailed scores reflecting our assessment are presented in Table I.A.1 of the Appendix I.A.

For our sample of 12 papers the range of observed total scores was 0 to 7. The largest score observed (7) also constitutes the modal score, which was assigned to 5 papers (Kjerstad and Tuntland, 2016; Lewin et al., 2013a; Lewin et al., 2013b; Tuntland et al., 2015; Langeland et al., 2019). Apparently, even the strongest papers reached only 47% of the maximum possible points. The mean and median total score were estimated as 5.4 and 6.0, respectively, while the standard deviation was 2.2. Approximately 83% of the papers received scores of 5, 6 or 7. Two papers were assigned 2 points or less.

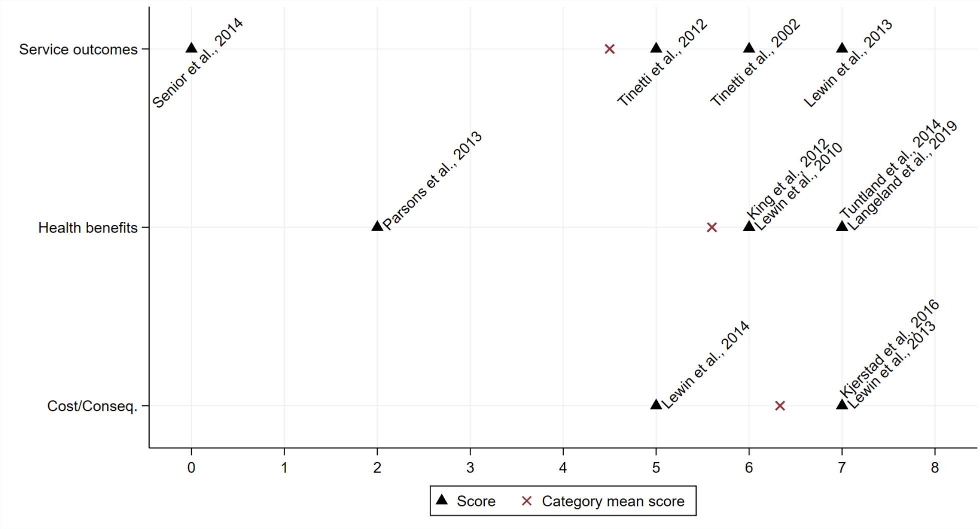

Figure I.2: Total scores for studies by category. Y-axis represents our defined study categories, and the X-axis represents total scores from our assessment scheme. Maximum possible score = 15. The triangular symbols represent the total score obtained for different studies, whereas the red multiplication sign indicates category means. Total scores is presented in Appendix I.A.

Figure I.2 illustrates the differences in the distribution of scores between the three categories defined in Section I.2.3. The highest average score (6.3) and the highest minimum score (5) were found in Category 1 – Costs and Consequences. In Category 2 – Health benefits and Category 3 – Service Outcomes, the average scores were 5.6 and 4.5, respectively. In those categories, the spread of the overall scores around the mean score exceeded the spread observed for Category 1.

When presenting the total scores of the studies by country of origin and year of publication as done in Figure I.3 , pronounced differences in means (x) and variances across countries became apparent. Two of the studies from Australia received a score of 7, but since two papers scored less, the Australian studies had an average score of 6.3. The mean score for the two studies carried out in the US was 5.5. Since two of the studies originating in New Zealand constitute the lower end of the range of total scores, the average score for New Zealand was only 2.7. In contrast, the three Norwegian studies received a score of 7.

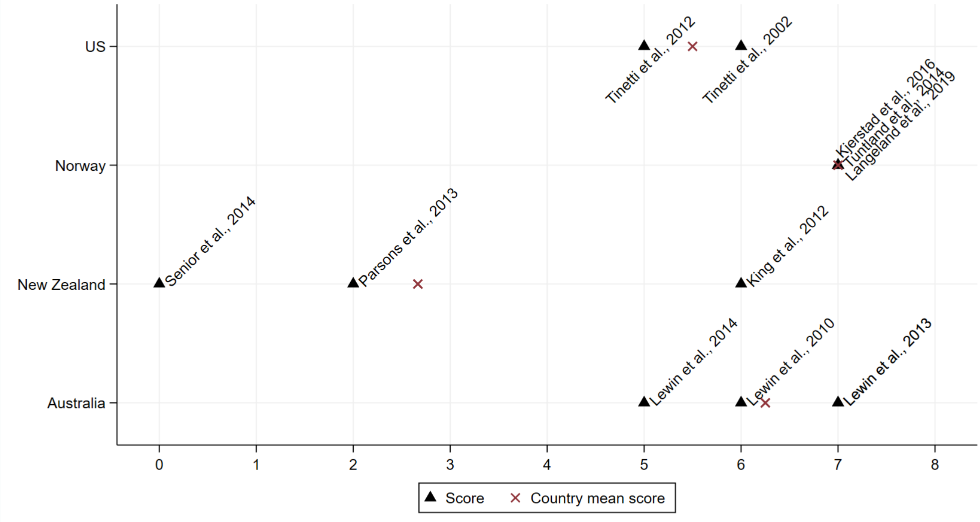

Figure I.3: Total scores for studies organized by country. Y-axis are countries of the included studies, and the X-axis represents total scores from our assessment scheme. Maximum possible score = 15. The triangular symbols represent the total score obtained for different studies, whereas the red multiplication sign indicates country means. Total scores is presented in Appendix I.A.

Figure I.3 does not provide evidence for a clear time trend. Nonetheless, it should be pointed out that the Norwegian studies represent fairly recent efforts. Thus, the authors of these studies had the chance to learn the shortcomings of earlier publications, and they effectively avoided them in their subsequent work.

Our descriptive analysis of the total scores attests that the existing HBR literature contains some heterogeneity but has an overall mediocre quality level. An inspection of the detailed scores given in Table I.A.1 (cf. appendix) readily reveals the reason for this assessment. All the papers accumulated scores of 0 on items 10-13 (Statistics) and items 14 and 15 (External validity). According to our assessment, virtually all the studies failed to be informative about key aspects of statistical modelling and lacked external validity. With respect to the former trait, we found only two exceptions. The study by Kjerstad and Tuntland, 2016 scored a 1 on item 10 for their explicit and careful motivation of the "bootstrap" strategy. The study by Langeland et al., 2019 scored a point for external validity.

A consequence of the previous observations is that all the existing heterogeneity can be attributed to differences in scores for items 1-5 (General introduction) and items 6-9 (Data).

The aggregated scores for the General Introduction (items 1-5) section ranged between 0 and 5. The average score was 3.5, which corresponds to 70% of the maximum possible score. Scaling the standard deviation of 1.4, we obtain a coefficient of variation of 41%. The modal score of '4' was received by 4 of the studies. On items 1 and 2, which relate to the explanation of the research question, 10 studies received a point on item 1 and 9 studies received a score of '1' on item 2. In contrast, only 6 studies were found to identify and describe important stakeholders in an adequate way. Summing the scores for items 6-9 resulted in scores varying between 0 and 3. The mean score of 1.8, equivalent to only 44% of the maximum score attainable, indicates that one can expect a paper from the HBR literature to show deficits in the data focus area of sampling and description. Note that on item 9, which relates to a clear data description, only the study by King et al., 2012 received a score of '1'. Moreover, only 5 studies were found to describe methods evaluating health states or other benefits (item 6) in a satisfactory way. Scaling the standard deviation of scores for this focus area (1.1) by the mean results in a coefficient of variation of 65%. Moreover, 3 studies (Lewin and Vandermeulen, 2010; Lewin et al., 2013a; Lewin et al., 2013b) with scores on items 1-5 that exceeded their respective mean scores also had scores larger than the overall mean for items 6-9. The correlation coefficient between the two sets of scores was 0.4.

Analysing the outcomes of the assessment process suggests that while a typical HBR paper describes the motivation and all aspects of the research question in a satisfactory manner, the documentation of data-related issues could clearly be improved. The latter issue also seems to contribute slightly more to the heterogeneity in quality.

The most striking outcome of the assessment so far is that the majority of the HBR papers under review failed to be informative about key aspects of the statistical modelling. This is surprising, since due to the nature of our selection process, all papers under review appear to rely on statistical methodology. One can group the techniques implemented into two groups, i) mean comparisons, both parametric and non-parametric, and ii) regression analyses. Table I.3 lists the different models and inferential techniques applied in the context of the primary outcomes.

Table I.3: Statistical methods for primary outcome in each study

			Regression:						
Author	**Bootstrap**	**Wilcoxon**	Linear	Logistic	GLM	Quantile	GLMM	MM	Cox-Hazard
Kjerstad and Tuntland, 2016	✓								
Lewin et al., 2014					✓				
Lewin et al., 2013b					✓	✓			
King et al., 2012							✓		
Lewin and Vandermeulen, 2010			✓	✓					
Parsons et al., 2013							✓		
Tuntland et al., 2015								✓	
Langeland et al., 2019								✓	
Lewin et al., 2013a			✓	✓					
Senior et al., 2014									✓
Tinetti et al., 2002		✓		✓					✓
Tinetti et al., 2012				✓					

Abbreviation: Wilcoxon = Wilcoxon signed rank test; GLM = Generalized linear model; GLMM = Generalized linear mixed model
MM = Mixed model, also called mixed effects models (Cameron and Trivedi, 2005, p. 774); Cox-Hazard = Cox proportional hazard model

Apparently, various types of regression models feature prominently in the HBR literature. According to our assessment, it is a prominent feature of the published HBR literature that the choice of such a model is virtually never justified. Alternative modelling approaches are not explicitly discussed. Models are not presented explicitly. Underlying key assumptions are not documented and it is typically not substantiated that they hold in light of the data collected. The 'path' leading from the data to the model is not made explicit. This, of course, has negative ramifications for the reader's ability to critically appraise the results as well as for the replicability of the research documented. To be clear on this point, we do not believe that the authors ignored the stated aspects of statistical modelling in the research process. We simply point out the fact that, for whatever reason, there is not sufficient space allocated to such considerations in the publications under scrutiny.

The almost homogenous '0' responses to item 14 regarding external validity suggest that the HBR studies existing so far still lack external validity. Three reasons for this drawback were typically identified and discussed: (i) small sample size and short timeframe, (ii) only one service provider or region and (iii) various methodical issues. These methodical issues arose from selection bias and manipulation of the randomization process. Langeland et al., 2019, who received a point for external validity, differs from other studies as it was conducted in several different areas of Norway and in a natural setting. However, one would ideally have more patients included in the control group. Finally, the fact that all studies were assigned a '0' score on item 15 regarding theoretical foundation does not come as a surprise. Here, we see the manifestation of a common trend stated by Deaton, 2010 (p. 425), *"Econometric analysis has changed its focus over the years, away from analysis of models derived from theory toward much looser specifications that are statistical representations of program evaluation."*

I.4 Discussion

The application of our assessment scheme revealed one specific common pattern among the studies. In our view, none of the papers scrutinized provided sufficient information about the data or the statistics employed. Almost all studies lacked external validity. For questions 9 – 15 (data description, statistics, external validity), only three points were given. We do not believe that this evidence is indicative of the quality of the underlying research process. More likely, our findings reflect an established publication standard idiosyncratic to the health and medical journals where these studies were published. Failure to provide full information on data and statistics can create uncertainty in an informed reader. For example, knowledge of the sampling procedure and the process of data generation is essential for choosing an identification strategy. Without this information, the reader will not be able to fully understand the data or the strengths and weaknesses of the study. Ten out of twelve studies used regression models. The models were not presented. None of the studies provided information regarding the estimation technique used or possible adjustments of the standard errors. Not providing this type of essential information leads to a lack of transparency that in turn will reduce the replicability of a study. Thus, our assessment detects a trait of the publication culture that runs counter to two values we believe should be promoted in social sciences in general and in health economics in particular.

Since seven of the included studies are RCTs, i.e., randomized experiments, it is interesting to discuss RCTs more explicitly. The ideal RCT is often considered the "gold standard" approach for establishing causality. In biostatistics, RCTs are often viewed as the only credible approach, while experimental evaluations have traditionally been less common in economics (Imbens and Wooldridge, 2009). One might, however, argue that a "gold standard" does not exist (Cartwright, 2007). The primary benefit of an experiment lies in the fact that it solves the selection bias problem, not by removing the bias but by balancing the bias between the experimental groups (Heckman and Smith, 1995). Experiments also provide a generalizable estimate of the treatment effect for the population when the sample size is large (MacLeod, 2017). The lack of sufficient sample size in the RCT studies reviewed contributes to their rather low external validity. In the design phase, all studies use power calculations for determining the target sample size. If one compares those targets to the numbers of participants included at the last follow-up, only Parsons et al., 2013 meets the number of participants indicated by their respective power calculation. Providing power calculations and meeting the indicated estimates does not necessarily translate into a possibly causal estimated treatment effect. Power calculations are also based on assumptions, and substantial guess work is needed (Duflo et al., 2007). Computing the results of an RCT is fairly straightforward, as it simply involves comparing the group means. However, for statistical inference one needs to estimate the standard errors, which is more complicated (Deaton, 2010). There are several alternatives for testing the significance of differences in means, but the workhorse for these computations is regression. As Table 3 indicates, most of the studies included used regression, and six of the seven RCT studies relied on regression.

Freedman, 2008a points out that it is common practice to adjust data from clinical trials using regression models and the like, which is also confirmed by the observations in this study. The standard way of performing a regression on data from clinical trials is to regress the outcome variable on one assignment variable, including a constant, and often control for baseline covariates. Freedman, 2008a analyses this model in detail and concludes that this standard way is nothing like a standard regression. He shows that the main issue is the dependence between the assignment variable and the error term, which violates key OLS assumptions. This could bias the estimated treatment effect substantially in small samples. The bias tends to decrease as the number of participants increases, but it is possible that a regression without covariates may render superior results. It is difficult to identify the studies in our analysis that use regression and OLS, but there are clues pointing at two studies (Lewin et al., 2013a; Parsons et al., 2013). Freedman, 2008b also discuss the issues of using logit/probit regression on experimental data. His key finding is that randomization does not justify the assumptions underlying these models because the outcome value is deterministic given the assignment value. Under a logit model, the outcome variable is interpreted as being random. Consequently, the usual maximum likelihood estimates could be inconsistent. The main problem here is not necessarily that these models have been used, as there are ways to solve the apparent problems, but rather the lack of discussing potential drawbacks. Freedman, 2008a (p. 13) states this issue quite sharply: *"Practitioners will doubtlessly be heard to object that they know all this perfectly well. Perhaps, but then why do they so often fit models without discussing assumptions?"* There are some non-technical problems with experiments, and these are more difficult to solve. Randomized experiments in the social setting often rest on unstated assumptions, especially considering the behavioural response of the participants, whose behaviour is often altered due to the randomization (Heckman, 1991). Randomization bias, or deviations from assignment, cannot necessarily be treated as random measurement error and can therefore influence the results (Deaton, 2010). None of the RCT studies discussed the latter aspects. The RCT technology may constitute a powerful tool in applied situations when the underlying assumptions are met. Often these assumptions are not arguably better than assumptions found in non-experimental econometrics and statistics (Heckman, 1991).

One of the objectives of this paper is to provide an overview of economic evaluations of HBR. Previous reviews either found no data evaluating the effects of HBR or only concentrated on a few outcome measures (Legg et al., 2016; "Effectiveness of Reablement: A Systematic Review"). Our review effort differs from earlier attempts, especially in terms of "wider" inclusion criteria with fewer limitations on study type and outcome measures. Each of the twelve studies found to be eligible for our review was assigned to one of three categories. Three studies estimated the cost differences between HBR and usual care after the intervention, and all showed lower costs for HBR participants (Kjerstad and Tuntland, 2016; Lewin et al., 2013b; Lewin et al., 2014). However, these results are not clearly significant. In the Norwegian study, the mean cost difference was not statistically significant (Kjerstad and Tuntland, 2016), and one study did not report significance (Lewin et al., 2013b). For one of the studies, the significant differences in mean cost differed between the ITT and AT analysis, with AT showing a significant difference (Lewin et al., 2014). Like the former, the latter study also includes cost for ED and hospital admissions along with homecare costs. If one focuses on homecare cost alone, then the potential yearly cost reduction for HBR seems to range from approximately 800 – 1,700 €per participant. Here, we have to stress that this is only a rough estimate, and the results of the study vary greatly with time. This might explain the wide range of potential cost savings. It seems that the potential savings increase with the length of the post-intervention period (Lewin et al., 2013b).

Table I.2 exhibits some general information about the studies included in this review. Scrutinizing columns five and six of the table, one finds similar inclusion and exclusion criteria defining the pool of participants in the various HBR studies. The main exclusion criterion was that participants are not in need of residential care and not significantly cognitively reduced. Most studies applied narrow inclusion criteria requiring that patients eligible for care are older than 65 years. An exception is the Norwegian studies, in which the minimum age was set to 18 years. There are, however, only small variations in the mean age of included participants. For the HBR group, the mean age was between 76 and 82, while for the usual care group it was between 77 and 83. An additional trait common to the studies reviewed is the length of the HBR intervention itself, which was a maximum of 12 weeks. In the New Zealand version, in which participants were referred from the hospital, the length was limited to 8 weeks. In the most recent Norwegian study, the intervention length varied between 4 to 10 weeks (Langeland et al., 2019). Two studies (King et al., 2012; Parsons et al., 2013), failed to be informative with respect to this aspect. The homogeneity with respect to this feature increases the comparability of the studies. In fact, according to our observations, the length of the intervention itself was hardly ever explicitly explained. Neither the actual amount of HBR administered to the participants nor the possible effects of a variation in treatment duration on treatment outcome was discussed. Studies examining potential health benefits from HBR do not use one standardized instrument. In fact, to establish different types of health benefits, one may require different instruments, but often there are several instruments used for the examination of the same type of benefit. Directly comparing the results then becomes difficult. We will therefore focus the discussion on whether there were some common trends in terms of statistical significance for potential health benefits.

Physical functioning or independence were the potential benefit categories where we found the most studies, and often these focused on ADL. The study by Lewin and Vandermeulen, 2010 is the first study to use functional gain as the primary outcome. This study produced some promising results. The HBR group scored significantly better on all physical measures after 3- and 12-month follow-ups. These results are consistent with earlier studies examining short-term effects (Tinetti et al., 2002). A more recent study also indicated improvements in physical functioning for the HBR group (Parsons et al., 2013). Less clear are the results of an Australian study, in which statistical significance in instrumental ADL could only be established in the context of the AT analysis. The latter study used a 12-month follow-up period. In contrast, three studies showed no statistical significance in either functional mobility or ADL (King et al., 2012; Senior et al., 2014; Tuntland et al., 2015). The follow-up periods in these three studies lasted between 7 and 18 months. This is longer than the respective period in the studies reporting positive statistical significance in favour of HBR. A common pattern for all the results is that there were no significant or clear effects on physical functioning. These studies all included physical gain as a secondary outcome. They were not originally designed for detecting any effect on physical functioning. This may influence the results. However, this argument partially also holds for the studies reporting a positive significant effect. The exception from the above finding was the study from Langeland et al., 2019, which found significant effect on the secondary outcome physical functioning. The study by Parsons et al., 2013 defines physical functioning as the primary outcome, but it relies on data collected from another study (Parsons et al., 2011). There is no clear evidence supporting the notion that HBR significantly increases physical functioning. Encouragingly, in the studies that produced no significant difference, HBR tended to lead to superior results on the selected instruments.

Increased HRQoL or QoL is often used as a measure of increased health benefits, and three studies in our review relied on this measure. However, only one study had change in HRQoL as the primary outcome (King et al., 2012). This study showed a promising result, with the HBR group scoring significantly better at the 7-month follow-up. The mental health component of SF36 was the main driver for the increased score for the HBR group. For the mean difference between baseline and 7-month follow-up, both the overall score and the mental component had a p-value of 0.05. The three remaining studies looking into HRQoL or QoL reported insignificant differences between HBR and usual care (Lewin et al., 2013a; Langeland et al., 2019; Tuntland et al., 2015) at the final follow-up period, which varies between 9 – 12 months. If one examines the result table in Tuntland et al., 2015 and Langeland et al., 2019, most of the HRQoL components are in favour of HBR. To summarize, there is no convincing long-term evidence that HBR increase HRQoL or QoL. Again, the studies that report insignificant differences had HRQoL or QoL as their secondary outcomes. With respect to other self-perceived health benefits, the results are also not definite. Two studies (King et al., 2012; Parsons et al., 2013) reported no significant difference in social support measured with the Duke Social Support Index (DSSI) (Koenig et al., 1993). The end follow-up periods for these studies only differed by 1 month, and both were conducted in New Zealand, making these studies highly comparable. The p-value associated with the change in the DSSI score between usual care and HBR was found to be 0.09 in the study by Parsons et al., 2013. However, the authors argue for the use of a significance level of lower than 0.05 because of the risk of type-II errors. Regression results from assessing the state of psychological well-being of older people also showed no significant difference at the 12-month follow-up (Lewin and Vandermeulen, 2010). Self-perceived activity performance and satisfaction with that performance was analysed in (Tuntland et al., 2015, Langeland et al., 2019). Both the performance and satisfaction measures were significantly better for the HBR group at the 6 or 9-month follow-up. However, the treatment effect was below the clinically relevant change cut-off.

Postponing the need for residential care is an obvious potential benefit for a patient and for the policy-maker, as residential care is, in general, costly. In an unadjusted analysis, it was demonstrated that HBR users were significantly less often assessed and approved for a higher level of care in a 2-year perspective (Lewin et al., 2014). Senior et al., 2014 observed that HBR reduced the probability of death or permanent residential care, but their observations lacked statistical significance. It was also shown that HBR users were less than half as likely to have an ED visit during the home care episode (Tinetti et al., 2002). Over a 2-year period, HBR recipients had significantly less ED presentations compared to individuals receiving the baseline treatment, though these results only hold for the AT analyses and were unadjusted (Lewin et al., 2014). The latter findings also hold for the number of hospital admissions. Moreover, an earlier study concluded that HBR participants were less likely to be readmitted to the hospital compared to subjects under usual care, a result that was only significant at a 10% level (Tinetti et al., 2012). In addition, HBR is showing some promising results with respect to reducing the need for specialist or residential care. As discussed earlier, HBR seems to reduce homecare costs, and therefore one would expect a decline in the volume of homecare services. In the first study included in this review, it was shown that HBR participants were significantly more likely to remain at home after a homecare episode (Tinetti et al., 2002). This effect seems to hold in a 12-month perspective, as it was shown that HBR participants were significantly less likely to need ongoing services (Lewin et al., 2013a). There is evidence for the fact that relative to usual care, HBR significantly reduces the number of homecare hours and visits as well as the general duration of homecare episodes in the long-term (Kjerstad and Tuntland, 2016; Lewin et al., 2014; Tinetti et al., 2002; Tinetti et al., 2012).

I.5 Conclusion

This review summarizes and assesses the currently available literature on empirical evaluations of the modern care concept of HBR. In short, the existing evidence regarding the effects of HBR is still inconclusive. The results are inconsistent, as some studies report a significant positive effect of HBR versus usual care, while others fail to establish such an effect. However, so far it has not been established that HBR renders negative effects. On one hand, this review is concerned with a concise summary of relevant existing findings generated by research focusing on HBR. On the other hand, it tries to provide a critical, constructive assessment of the associated publication process. Having worked on this project, we understand that HBR is a complex intervention implemented in an equally complex setting. Out of this understanding grows the utmost respect for all current research efforts aimed at estimating the effects of HBR. The research reviewed provides a basis to build on. With complex interventions in social settings, there might also be a need for different “eyes” to capture this complexity. To ensure successful future research efforts, the multidisciplinary spirit of HBR needs to be reflected in the diversity of the research teams taking on the challenge.

I.6 References

Aspinal, Fiona et al. (Sept. 2016). "New horizons: Reablement - supporting older people towards independence". In: *Age and Ageing* 45.5, pp. 574–578.

Baker, Dorothy I. et al. (Apr. 2001). "The Design and Implementation of a Restorative Care Model for Home Care". In: *The Gerontologist* 41.2, pp. 257–263.

Barnett, Karen et al. (2012). "Epidemiology of multimorbidity and implications for health care, research, and medical education: a cross-sectional study". In: *The Lancet* 380.9836, pp. 37 –43.

Cameron, Adrian and Pravin Trivedi (Jan. 2005). *Microeconometrics: Methods and Applications*. Cambridge University Press.

Cartwright, Nancy (2007). "Are RCTs the Gold Standard?" In: *BioSocieties* 2.1, pp. 11–20.

Cochrane, A. et al. (2013). "Home-care 're-ablement' services for maintaining and improving older adults' functional independence". In: *Cochrane Database of Systematic Reviews* 11.

Deaton, Angus (2010). "Instruments, Randomization, and Learning about Development". In: *Journal of Economic Literature* 48.2, pp. 424–455.

Drummond, M F and T O Jefferson (1996). "Guidelines for authors and peer reviewers of economic submissions to the BMJ". eng. In: *BMJ* 313.7052, p. 275.

Drummond, Michael F. et al. (2005). *Methods for the economic evaluation of health care programmes*. 3rd. Oxford University Press.

Duflo, Esther, Rachel Glennerster, and Michael Kremer (2007). "Chapter 61 Using Randomization in Development Economics Research: A Toolkit". In: ed. by T. Paul Schultz and John A. Strauss. Vol. 4. Handbook of Development Economics. Elsevier, pp. 3895 –3962.

Emmert, Martin et al. (2012). "Economic evaluation of pay-for-performance in health care: a systematic review". In: *The European Journal of Health Economics* 13.6, pp. 755–767.

Fersch, Barbara (2015). "Expectations towards home care re-ablement in Danish municipalities". eng. In: *International Journal of Sociology and Social Policy* 35.3/4, pp. 126–140.

Freedman, David A. (2008a). "On regression adjustments to experimental data". In: *Advances in Applied Mathematics* 40.2, pp. 180 –193.

— (2008b). "Randomization Does Not Justify Logistic Regression". In: *Statistical Science* 23.2, pp. 237–249.

Gruber-Baldini, Ann L. et al. (2011). "Adverse Events Associated With the Res-Care Intervention". In: *Journal of the American Medical Directors Association* 12.8, pp. 584 –589.

Heckman, James (1991). *Randomization and Social Policy Evaluation*. NBER Technical Working Papers 0107. National Bureau of Economic Research, Inc.

Heckman, James J. and Jeffrey A. Smith (1995). "Assessing the Case for Social Experiments". In: *The Journal of Economic Perspectives* 9.2, pp. 85–110.

Howard-Wilsher, Stephanie et al. (2016). “Systematic overview of economic evaluations of health-related rehabilitation”. In: *Disability and Health Journal* 9.1, pp. 11 –25.

Imbens, Guido W. and Jeffrey M. Wooldridge (2009). “Recent Developments in the Econometrics of Program Evaluation”. In: *Journal of Economic Literature* 47.1, pp. 5–86.

King, Anna I. I. et al. (2012). “Assessing the impact of a restorative home care service in New Zealand: a cluster randomised controlled trial.” In: *Health & Social Care in the Community* 20.4, pp. 365 –374.

Kjerstad, Egil and Hanne Tuntland (2016). “Reablement in community-dwelling older adults: a cost-effectiveness analysis alongside a randomized controlled trial”. In: *Health Economics Review* 6.1, pp. 1–10.

Koenig, Harold G. et al. (1993). “Abbreviating the Duke Social Support Index for Use in Chronically Ill Elderly Individuals”. In: *Psychosomatics* 34.1, pp. 61 –69.

Langeland, E. et al. (2019). “A multicenter investigation of reablement in Norway: a clinical controlled trial”. In: *BMC Geriatrics* 19.1, p. 29.

Legg, Lynn et al. (2016). “A systematic review of the evidence on home care reablement services”. In: *Clinical Rehabilitation* 30.8. PMID: 26374423, pp. 741–749.

Lewin, Gill, Karyn Concanen, and David Youens (2016). “The Home Independence Program with non-health professionals as care managers: an evaluation”. English. In: *Clinical Interventions in Aging* 11, pp. 807–817.

Lewin, Gill and Suzanne Vandermeulen (2010). “A non-randomised controlled trial of the Home Independence Program (HIP): an Australian restorative programme for older home-care clients”. In: *Health & Social Care in the Community* 18.1, pp. 91–99.

Lewin, Gill et al. (2008). “The Home Independence Project”. In: *Geriaction* 26 (3), pp. 13–20.

Lewin, Gill et al. (2013a). “A randomised controlled trial of the Home Independence Program, an Australian restorative home-care programme for older adults”. In: *Health & Social Care in the Community* 21.1, pp. 69–78.

Lewin, Gill et al. (2014). “A comparison of the home-care and healthcare service use and costs of older Australians randomised to receive a restorative or a conventional home-care service”. In: *Health & Social Care in the Community* 22.3, pp. 328–336.

Lewin, Gill F, Helman S Alfonso, and Janine J Alan (2013b). “Evidence for the long term cost effectiveness of home care reablement programs”. eng. In: *Clinical interventions in aging* 8, p. 1273.

MacLeod, W. Bentley (2017). “Viewpoint: The human capital approach to inference.” In: *Canadian Journal of Economics* 50.1, pp. 5 –39.

Oliveira Martins, Joaquim and Christine Maisonneuve (2007). “The drivers of public expenditure on health and long-term care”. In: *OECD Economic Studies* 2006, pp. 115–154.

Parsons, J.G. et al. (2013). “A randomized controlled trial to determine the effect of a model of restorative home care on physical function and social support among older people”. In: *Archives of Physical Medicine and Rehabilitation* 94 (6).

Parsons, John et al. (2011). "Goal setting as a feature of homecare services for older people: does it make a difference?" In: *Age and Ageing* 41.1, pp. 24–29.

Ryburn, Bridget, Yvonne Wells, and Peter Foreman (2009). "Enabling independence: restorative approaches to home care provision for frail older adults". In: *Health & Social Care in the Community* 17.3, pp. 225–234.

Senior, Hugh E. J. et al. (2014). "Promoting independence in frail older people: a randomised controlled trial of a restorative care service in New Zealand". In: *Age and Ageing* 43.3, pp. 418–424.

Tessier, A. et al. "Effectiveness of Reablement: A Systematic Review". In: *Healthc Policy* 11 (4), pp. 49–59.

Tinetti, Mary E. et al. (2002). "Evaluation of Restorative Care vs Usual Care for Older Adults Receiving an Acute Episode of Home Care". In: *JAMA* 287.16, pp. 2098–2105.

Tinetti, Mary E. et al. (2012). "Effect of a Restorative Model of Posthospital Home Care on Hospital Readmissions". In: *Journal of the American Geriatrics Society* 60.8, pp. 1521–1526.

Tuntland, Hanne et al. (2014). "Reablement in community-dwelling adults: study protocol for a randomised controlled trial". In: *BMC Geriatrics* 14.1, p. 139.

Tuntland, Hanne et al. (2015). "Reablement in community-dwelling older adults: a randomised controlled trial". In: *BMC Geriatrics* 15.1, p. 145.

Whitehead, Phillip J et al. (2015). "Interventions to reduce dependency in personal activities of daily living in community dwelling adults who use homecare services: a systematic review". In: *Clinical Rehabilitation* 29.11. PMID: 25587088, pp. 1064–1076.

W.H.O, Europe (2012). *Strategy and Action Plan for Healthy Ageing in Europe, 2012-2020.* Tech. rep. 122346. World Health Organization.

Wiles, Janine L. et al. (2011). "The Meaning of "Aging in Place" to Older People". In: *The Gerontologist* 52.3, pp. 357–366.

Yoon Y., Jean et al. (2014). "Costs Associated With Multimorbidity Among VA Patients". In: *Medical Care* 52 Suppl 2, ADVANCING THE FIELD: Results from the AHRQ Multiple Chronic Conditions Research Network.3, S31–S36.

I.A Appendix A: Assessment points given

Table I.A.1: Results of the quality assessment

	Question															
Author	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**Score**
Kjerstad and Tuntland, 2016	1	0	0	1	1	1	1	1	0	1	0	0	0	0	0	7
Lewin et al., 2014	1	1	1	0	1	0	1	0	0	0	0	0	0	0	0	5
Lewin et al., 2013b	1	1	1	1	1	0	1	1	0	0	0	0	0	0	0	7
King et al., 2012	0	1	0	1	1	0	1	1	0	0	0	0	0	0	0	6
Lewin and Vandermeulen, 2010	1	1	1	1	1	0	0	1	1	0	0	0	0	0	0	6
Parsons et al., 2013	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	2
Tuntland et al., 2015	1	1	0	1	1	0	1	1	0	0	0	0	0	0	0	7
Langeland et al., 2019	1	0	0	1	1	1	1	1	0	0	0	0	0	1	0	7
Lewin et al., 2013a	1	1	1	1	1	0	1	1	0	0	0	0	0	0	0	7
Senior et al., 2014	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Tinetti et al., 2002	1	1	1	1	0	1	1	0	0	0	0	0	0	0	0	6
Tinetti et al., 2012	1	1	1	1	0	1	0	0	0	0	0	0	0	0	0	5
Sum	10	9	6	9	8	5	8	7	1	1	0	0	0	0	0	

Paper II

Absenteeism and care provision: Evidence from a Norwegian cross-sectional dependent panel

Absenteeism and care provision: Evidence from a Norwegian cross-sectional dependent panel

Kjetil Andersson[1], Tore Bersvendsen[1*] and Jochen Jungeilges[1,2]

[1]Department of Economics and Finance, University of Agder, Norway
[2]Institute of Mathematics and Natural Sciences Ural Federal University, Ekaterinburg, Russian Federation

We examine the effects of increased sick leave on indirect expenses experienced by a public-health entity. On the basis of a unique 36-month panel from a micro environment in a large Norwegian municipality, we study how home-nurse hours provided by the public health entity vary in response to changes in absenteeism occurring in different organizational units of the care providing entity. While sick leave occurring in the central administration tends to increase the level of care provided, the opposite effect is established for the case in which sick leave increases in an operative unit of the organization. As a spin-off, we find that care recipients who are subjected to home-based reablement, a novel care treatment, tend to need fewer hours of personal services than comparable individuals subjected to a standard care treatment. Our approach to the problem renders a case study in applied econometrics demonstrating the roles knowledge about organizational features of a real-life system play on different stages of the econometric modelling process. In particular, we focus on a comparison of alternative state-of-the-art standard error correction techniques in the presence of the triad "heteroscedasticity, autocorrelation and cross-sectional dependence".

Keywords: Sick-leave, panel data, cross-sectional dependence, standard error correction

JEL classification: C18, I19, J39

**Corresponding author:* Tore Bersvendsen, tore.bersvendsen@uia.no

II.1 Introduction

Reflection on the phenomenon of absenteeism from the workplace readily produces a list of non-beneficial correlates that could be interpreted as costs. Such cost might be severe for the employee, family members, co-workers and the health system as such. According to Barmby et al., 2002, the absence rates due to sickness are quite substantial in several industrial countries, and they vary significantly across gender. Moreover, there is empirical evidence that workers in the public sector tend to have more days absent than employees in the private sector (Winkelmann, 1999). The direct cost of absenteeism is found to be substantial. According to Chatterji and Tilley, 2002 the direct cost of absenteeism in the UK has been estimated to exceed 1% of GDP. Barham and Begum, 2005 estimate the direct cost of days absent at £11.6 billion for 2003 which amounts to 0.928% of GDP[1]. According to Norway's national budget 2018 public sick pay amounts to 42.6 billion NOK which equals 1.2% of Norway GDP estimated for 2018[2]. A Norwegian study (Markussen, 2012) shows that on average, a 1% increase in an individual's sick leave rate reduces individual earnings 2 years later by 1.2 %.

In general, absence can either be valid, if due to sickness, or invalid, if due to shirking. So absenteeism is important since it provides information concerning the determinants of worker behavior (Brown and Sessions, 1996). There exist theoretical positions linking productivity to social activities in the workplace. For instance, Rotemberg, 1994 argues that more (less) opportunities to socialize promote (impede) altruism among co-workers which, in turn, tends to increase (decrease) productivity. Considering this relationship, and assuming that temporary employees cannot be hired, one predicts that sick leave reduces productivity: An increase in sick leave, reduces the number of potential participants in social activities. Simultaneously, the workload of the attendees typically increases. In consequence, those individuals can devote less time to socializing. As altruism is less likely to be promoted in such a situation, productivity will be affected negatively. Another strand of the theoretical literature emphasizes links between productivity and peer pressure. Kandel and Lazear, 1992, for example, argue that peer pressure plays a role in creating incentives for increased effort and thereby influences firm productivity. Significant spillover effects in absenteeism among peers at work have been established by Godøy and Dale-Olsen, 2018 on the basis of high-quality Norwegian matched employer-employee data. The effect is substantiated for self-certified absence as well as physician-certified absence from work. Conditional on model specification the authors find that a 1% increase in absence rate for focal workers increased the peer workers absence rate by as much as 0.41%. The significant role of peer effects for absenteeism has been documented for Sweden and Italy, where the findings lie within the range stated in the Norwegian study (Hesselius et al., 2009; Paola, 2010).

This study, first of all, examines the effects of increased sick leave on indirect expenses experienced by a public-health entity. That is, we scrutinize cost associ-

[1]The 2003 UK GDP in current market prices equaled £1.25 trillion (or 1250 billion).

[2]Estimated 2018 market BNP for Norway: 3,530,860 million, (42,600 M/3,530,860 M) * 100 = 1.2065%.

ated with absenteeism occurring over and above the direct costs as, for instance, the wage rate. A *unique 36-month panel from a micro environment in a large Norwegian municipality* is utilized to estimate the variation in the amount of additional home-nurse hours provided due to changes in administrative workers sick leave occurring in different organizational units of the care providing entity. Secondly, since we deal with idiosyncracies of the panel data set at hand, using unique insights into the administrative background of the municipality studied, the current effort constitutes an informative case study in applied econometrics. In particular by concentrating on the effects of techniques designed to correct standard errors of estimates for heteroscedasticity, autocorrelation and *cross sectional dependence* on inference, we touch upon a technical matter that has recently received considerable attention in applied statistics.

The findings concerning the effects of sick leave are in line with theoretical (Pauly et al., 2002) and empirical (Nicholson et al., 2006) findings, implying the cost of work loss will exceed wages if the firm has difficulties substituting the absentee with an equally productive worker. In the case at hand, the administration office has little personnel slack and the typical job responsibilities are fairly specialized. Therefore replacing an absent worker by one of equal competence and productivity is close to impossible. Interestingly, we find the opposite effects of increased sick leave in the home-nurse areas. In contrast to the administrative unit, the management of these areas can rely on a pool of potential substitute nurses temporarily replacing the nurse(s) on sick leave.

In addition, we examine the short-term effect of a fairly new health-care intervention referred to as home-based reablement (HBR). The main goal of HBR is to restore or increase a patient's level of functioning and self-reliance to decrease the individual's dependence on health-care services such as home-nurse services. HBR is time-limited, multidisciplinary, home-based, goal-oriented and person-centered. Typically, a multidisciplinary team works towards a patient-defined goal formulated in terms of everyday activities important to the patient. Seen from a public policy perspective such care strategies represent an attempt to provide support to the elderly in a community while keeping the strain on health/care budgets in check (Fersch, 2015; Bauer et al., 2019). The research area "economic evaluations of HBR" is still in its infancy. Yet two studies are of special relevance for us. Lewin et al., 2014 and Kjerstad and Tuntland, 2016 showed - on the individual level - that HBR recipients needed fewer hours of personal services, such as home nursing, than comparable individuals subjected to a standard care treatment. Our study supports these findings. Implementing HBR in the municipality has significantly reduced home-nurse hours.

Apart from health-care related results, we put forward a case study in applied econometrics. We demonstrate how knowledge about the basic "architecture" of the system designed to order and provide home-nurse hours helps to choose an adequate estimator in the context of our linear model of home-nurse service provision. We argue below that due to key features of the system, we can assume that unknown common factor(s) generating cross-sectional correlation in the variables are uncorrelated with those factors causing correlated errors. As a consequence, fixed effects estimation will guarantee unbiased and consistent parameter estimates. Moreover, we rule out endogeneity by considering the design/organization of the administrative system in place to facilitate home-nurse hours in combination with empirical evidence. And finally, in the context of correcting standard errors in the presence of the triad "heteroscedasticity, autocorrelation and cross-sectional dependence", our study reveals that "no-correction" is no reasonable alternative. We substantiate that simply relying on default correction options of standard software can lead to suboptimal inference with respect to the effects of our variables of interest on home-nurse hours.

Accordingly, the paper augments two fields of the empirical literature in economics: health-care economics and applied econometrics. With respect to the former, its "value added" extents to the role absenteeism plays for the provision of socially highly valued care services. With respect to the latter, the paper delivers a unique demonstration of the role(s) knowledge about organizational features of a real-life system can play on different stages of the econometric modelling process. To our knowledge this is the first paper that compares state-of-the-art techniques of standard error correction on real life data.

Section II.2 discusses the design and the organizational structures of the system of care provision in the Norwegian municipality and outlines some basic features of the sick leave dynamics measured throughout the observation period $[2014, 2016]$. Variable definitions along with information about the respective measurement processes, and a first descriptive analysis of the data panel are provided in Section II.3. Our model for the provision of home-nurse hours is specified in the subsequent Section II.4. In the following Section II.5 we establish cross-sectional dependence (CSD) as a phenomenon relevant in our situation. Alternative conceptualizations of the sources for CSD are discussed. Adopting the factor-structure approach for our own purposes, we finally justify our key identifying assumption. Section II.6 introduces a portfolio of currently available techniques suitable for the correction of standard errors if a panel data setting is potentially ridden by heteroscedasticity, autocorrelation and/or CSD. The performance of these methods will be assessed in the sequel. In Section II.7, we point at the features of the administrative system which allow us to rule out endogeneity. Our results are presented in Section II.8. To underline that our findings are robust we perform a robustness check by employing the SUR modelling approach in Section II.9. A discussion of our results, given in Section II.10, concludes the paper.

II.2 Institutional background and sick-leave dynamics

In Norway, unlike in many other OECD countries, the provision of health care to the population is an exclusive public sector responsibility. Primary care and specialist care are, however, organized differently. Primary services are administered and supplied under the auspices of the atomic unit of local government in Norway referred to as municipalities. While these units administer a wide spectrum of health services, e.g. rehabilitation, physiotherapy, preventive medicine, and health promotion, typically a main focus area is long-term care provided either in nursing homes or at the patient's home. Municipalities are mainly financed by grants from the national government and through local taxes. The allocation of funding for services provided lies in the responsibility of the highest legislative body, i.e. the municipality council.

We observe data from one of the most populous municipalities in Norway. In this municipality, one administrative office decides on applications for primary care mainly in long-term homes or nursing homes. In a sense, the office functions as the *gate keeper* for access to care services. The municipality operates under an organizational structure referred to as *order-execute model*. In particular, the administrative office *orders* home-nursing hours from home-nurse areas. The personnel of a home-nurse area *executes* these orders by providing the respective care services to the patients. Home-nurse areas are financed by the amount of home nursing provided in a pay-for-performance system[3]. Eleven home-nurse areas have been established within the geographic boundaries of the municipality considered.

The administrative office scrutinized manages more than 90% of the incoming *home-nurse applications*[4]. Those are approximately randomly allocated to the case workers via the applicants' birth dates. The special importance of this feature will become apparent in Section II.5. Within this group approximately half of the employees have additional responsibilities. They function as contact persons (CPs) to one or more of the eleven home-nurse areas[5]. In this function they hold regular meetings with representatives of the home-nurse areas, and handle the so-called "need-change" (NC) notes. Those notes are a key element of the *order-execute* structure. Monitoring the patients regularly, home nurses will notice changes in patient needs. By initiating NC notes the home-nurse area indicates that a patient's amount of home nurse hours ought to be modified. The rule is that the need for change is filed only if the absolute value of the relative change in home-nurse hours exceeds 25%. The NC notes are then sent to the CP's for approval. Given the experience of the case handler and the evident nature of many cases, the NC notes are typically approved. Should the CP's object the amount of hours indicated in a NC note, then they start a *special* professional *dialog* with the home-nurse areas to determine an acceptable level of hours. In this matter, the CPs have the sole decision making authority. The NC related flow of information between the administration office and the home-nurse areas is illustrated in Figure II.1. In *regular meetings* between the CP and representatives of the home-nurse areas, the parties discuss the NC-notes practise as well as complicated cases. This system has been has been in place for more than a decade. Since personnel turnover is low, each employee knows her/his role in this largely trust-based system.

[3]Although several Norwegian municipalities follow the same organization, they often differ with respect to the home-nurse financing mechanism.

[4]The office consists of two groups. In what follows we focus on Group 1 (GR1) which handles applications from patients with somatic health challenges. The second group is responsible for all non-somatic cases.

[5]In 2016, eight of the sixteen full time employees worked as contact persons.

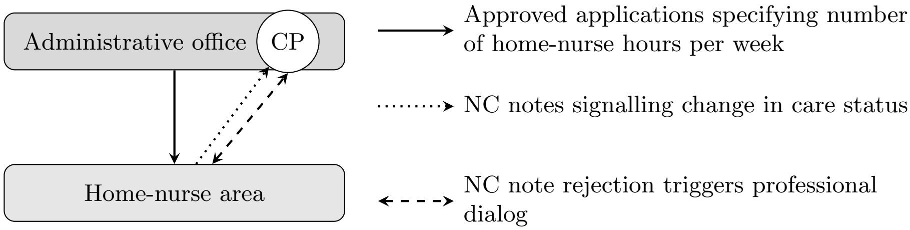

Figure II.1: Information flow related to NC notes

From an institutional point of view, two types of paid sick leave have to be distinguished in Norway: physician-certified sick leave and self-certified sick leave. General practitioners act as gatekeepers for paid sick leave, as all sickness related absence lasting longer than 3 days must be approved by a physician (Godøy and Dale-Olsen, 2018). If a sick employee is absent for less than 4 days, then a self-declaration is sufficient. Employees of the municipality studied have up to 52 days of self-certified sick leave at their disposal.

Throughout the investigation period the administration office was experiencing internal difficulties which were apparently reflected in an abnormal inflation of absence due to sickness. To be able to track this development, we operationalize "absence due to sickness" by the number of sick leave days as a fraction of work days *per month* that an individual has to deliver according to her/his employment contract. For groups, the individual measures are adequately aggregated. Prior to our sampling period, absenteeism did not seem to have been a problem, neither in GR1 nor among the CPs. The development of sick leave over the sample period for the home-nurse area, GR1 and its CP subgroup is illustrated in Figure II.2. Leaving aside the unsystematic "spike" observed in February 2015, from 2014-m1 to 2015-m6, the sick leave for GR1 and its subgroup CP was fairly stable below 4%[6]. From July 2015 on, one observes a pronounced cyclical motion - of period 4 and increasing amplitude - for the sick leave series among GR1 employees. The series for the CP group, exhibits the same type of dynamics, although for this group sick leave is more stable until it eventually escalates throughout the last quarter of 2016. In contrast to this pattern, Figure II.2 shows a fairly persistent downward trend in home-nurse-area sick leave throughout the observation period. A pronounced decrease in sick leave occurs over the first quarter of 2015.

[6] The spike occurring in February 2015 reflects the fact that three CP's were on partial leave simultaneously (50%, 25%, 15%).

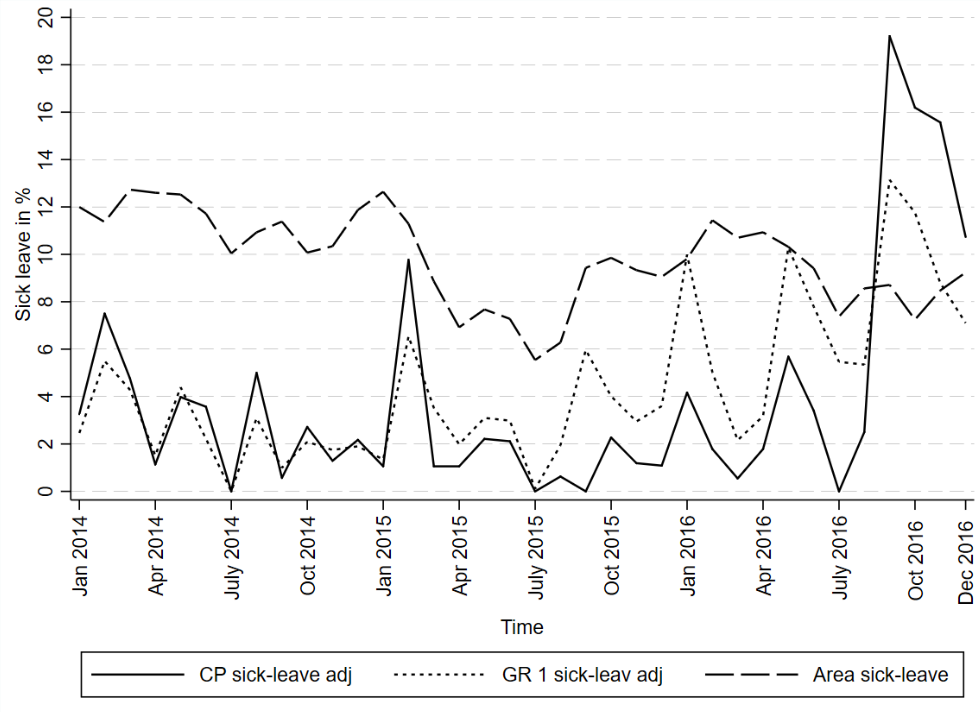

Figure II.2: Monthly sick-leave rates for GR1, CPs and Home-nurse areas (2014m1-2016m12)

Since it is typically difficult - if not impossible - to find substitutes for a CP on sick leave, the respective employee will face an inflated workload once she/he returns to the workplace. Since employees do not get paid overtime, they have an incentive to prioritize the time allocated to different tasks. Patients' needs always comes first, so treating new applications gets first priority. The CP's additional responsibilities, facilitating the regular meetings and treating the NC notes, are then put into second place. The NC notes will be treated, but more swiftly, less frequently, and less time is allocated to this task. During spring 2016, the regular meetings between CP and home-nurse areas also where dropped because of time constraints. Being pressed for time might increase a decision makers preference for those alternatives that minimize post-decision cost. In the case at hand, the acceptance of a NC note is less time consuming than a rejection, since there is less need for post-decision communication with the home-nurse areas after an acceptance. Moreover, as a consequence of less processing time per NC and the partial suspension of regular meetings, the CPs knowledge concerning individual NC notes will decay. In the light of these considerations one would expect that the acceptance rate for NC notes goes up as sick leave in the CP group increases. In line with the theory of Pauly et al., 2002, increased CP sick leave should lead to an inflation in home-nurse hours.

The municipality considered had implemented HBR already prior to 2014. A single HBR team operated as a pilot in two home-nurse areas. By January 2014 only two areas offered HBR type services. Subsequently, HBR was implemented gradually in all eleven home-nurse areas. This process was completed in November 2015. Only four different teams operated simultaneously in all of the eleven home-nurse areas. As mentioned in the introduction, due to the nature of HBR we expect that the new care treatment will tend to decrease home-nurse hours.

II.3 Data collection and description

Our study is based on data collected from different data systems used in the municipality. The observations on the dependent variable *amount of home-nurse hours per month* (HOURS), have been directly collected from the municipality's balanced scorecard system. This system uses a data format that is aggregated on a monthly level based on the electronic patient journal (EPJ) system. The amount of home-nurse hours is defined as the sum of all hours provided to patients during a given month. If a patient has been absent, and/or needed more than one employee to provided the indicated amount of hours *per* week, the hours are adequately adjusted. The same balance scorecard system provides the data source for *home-nurse-area sick leave*. Those observations were generated by the municipality's human resources (HR) system. A common account code was used to link the respective records.

All socio-demographic variables have been extracted from the EPJ system at an individual level. Originally provided in the form of a weekly panel, those observations had to be aggregated into monthly home-nurse area data. This is possible since there are defined service codes that are included in the definition of home-nurse hours. Only these codes where included in the individual level data to avoid unnecessary bias from other service types in the socio-demographic variables. Each individual is also registered at the home-nurse area under the same common account code mentioned above. Therefore it was possible to link the different data sources.

The variables HBR and CP sick leave had to be collected manually. To construct the HBR variable, one researcher met with the municipality's HBR project leader and retrieved the information about the timing of home-nurse areas' HBR implementation. Moreover, in collaboration with colleagues and leaders of GR1, the subgroup of employees serving as CPs was identified and the home-nurse areas served by each CP were recorded. Since personnel turnover was low, and only minor organizational changes occurred throughout the sampling period this procedure is likely to have produced reliable observations. The CP sick leave information was collected at the employee level directly from the HR system. In discussion with the *Norwegian Center for Research Data*[7] the CP's common account codes were rendered anonymous in line with the requirements of current data protection legislation.

In addition, data on a dummy variable *substitute leader* where generated in the course of a procedure similar to the one implemented in the HBR context. The two top managers of the home-nurse areas functioned as our source of information. Among the eleven different home-nurse areas, only two areas were managed by a substitute

[7]The *Norsk Senter for Forskningsdata* (NSD) is Norway's leading authority for privacy concerns arising in research data.

leader during 29 month. For a period of 2 months substitute leaders were in place in both areas simultaneously.

Moreover, three dummy variables are included to capture the potential effects of some minor organizational changes which occurred during the sample period. One variable captures a reduction in hours for two areas, since their responsibility for serving some small service apartments was moved to other institutions. In October 2014, a change in the municipality's traffic infrastructure affected the geographical size of two neighboring home-nurse areas. While one area was extended (by one street) the other shrank (by one street). To capture the effects of this change we introduce two binary variables, one for each of the respective home-nurse areas.

Table II.1 presents descriptive statistics for all time-varying variables included in our panel data set which includes observations for $N = 11$ home-nurse areas over $T = 36$ months. The monthly average of *home-nurse hours* per home-nurse area has increased by 3% from 2014 to 2016 after a slight drop in 2015. In contrast, the monthly average number of patients receiving home-nurse services grew monotonically at a rate of 3%. In 2016, the monthly home-nurse hours per user (15.08) where on average higher than in any other year, especially compared to 2015 (14.48). The home-nurse areas vary in general size, measured in amount of home-nurse hours and number of users. The within and between variation in home-nurse hours is reflected in Figure II.3.

Table II.1: Summary statistics

	2014	2015	2016	2014-2016		
	Mean	Mean	Mean	Mean	SD within	SD between
Homenurse hours	2,484	2,434	2,565	2,494	199.1	668.4
CP sick-leave in %	2.60	1.97	5.47	3.35	9.13	3.44
Area sick-leave in %	11.47	8.68	9.35	9.83	4.51	2.55
HBR	0.33	0.58	1	0.63	0.41	0.27
No. home-nurse patients	165	168	170	168	9.6	30.7
Substitute leader	0.11	0.09	0.02	0.07	0.20	0.17
Average age	74.8	74.7	74.8	74.8	0.60	3.75
Proportion of Men	0.36	0.37	0.38	0.37	0.03	0.01
Married or in partnership	0.21	0.23	0.26	0.23	0.03	0.06
Function score	2.56	2.63	2.66	2.62	0.12	0.20
Observations	132	132	132	396		
N	11	11	11	11		
T	12	12	12	36		

As can be seen in Table II.1, monthly average CP sick leave has more than doubled throughout the sample period. The average sick leave in the home-nurse area has decreased during the period, but there is a large within variation. There is a possibility for multi-collinearity between substitute leader and area sick leave. When the daily leader is on long-term sick leave, the leave is registered on the home-nurse area. Mainly, the reason for having a substitute leader is typically the long-term sick leave for the daily leader. The values of the binary substitute leader variable in Table II.1

indicate that the fraction of time during which provisional management was in place has been monotonically decreasing over the time period considered[8]. The average value on HBR was one in 2016, meaning that all home-nurse areas had implemented HBR by the end of 2015. Scrutinizing the fraction of male patients in the sample indicates that the gender composition of the patient population has been fairly stable across time as well as across areas. The latter observation also holds for the average age of approximately 75 years. A slight upward trend in the proportion of patients who do not live alone can be detected. The variable function score measures a patient's health status. Realizations of the index take values in the interval $[1, 5]$ with increasing scores reflecting a decay in the individual's health status. Due to data quality issues, the marginal increase in the means of the overall function score should be viewed with care[9]. Full variable list is provided in Appendix II.A

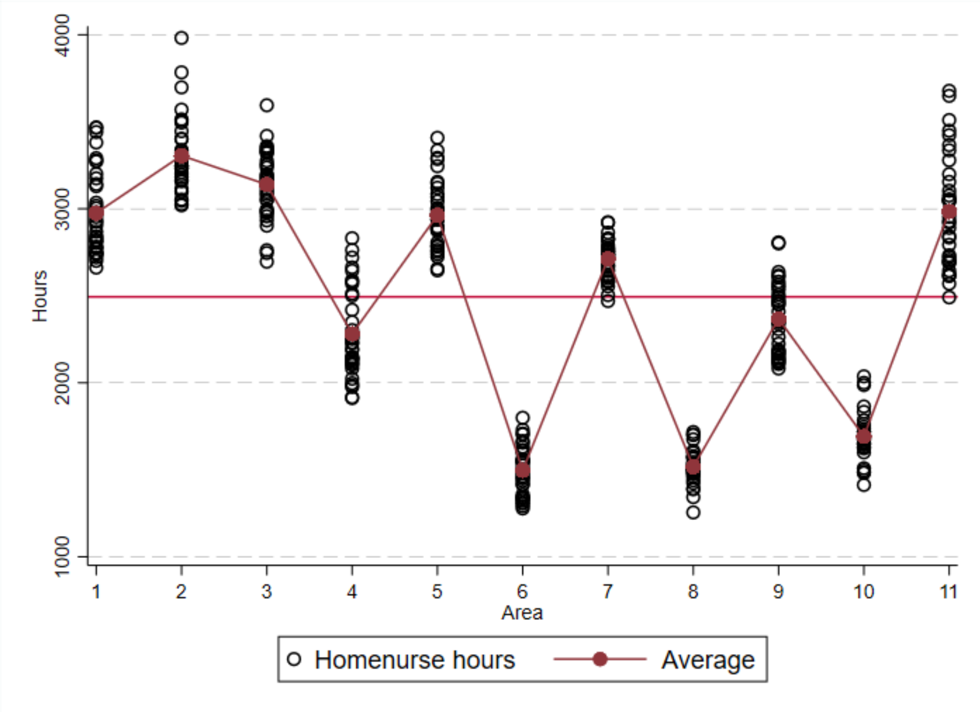

Figure II.3: HOURS$_{i\bullet}$ versus area i augmented by grand mean and time averages

[8]We have rerun our regressions including only one of these two variables in turn. There was no effect on the results.

[9]When trying to match function scores with patients registered at the home-nurse areas, we noticed that the scores for many patients were missing. So the aggregated scores on home-nurse areas do not include all patients that get home-nurse service. If this is a common trend between all home-nurse areas, then this weakness should not effect average scores.

II.4 A linear model for the provision of home-nurse hours

We assume the following relationship between home-nurse hours provided and our variables of interest

$$\text{HOURS}_{it} = \alpha + \beta_1 \,\text{HBR}_{it} + \beta_2 \,\text{CP-LEAVE}_{it} + \beta_3 \,\text{AREA-LEAVE}_{it} + \mathbf{z'}_{it}\boldsymbol{\theta} + a_i + \gamma_t + u_{it} \tag{II.1}$$

where $i = 1, 2, \ldots, 11$ indexes home-nurse areas and $t = 1, 2, \ldots, 36$ serves as the time index. In equation (II.1), HOURS_{it} denotes the number of home-nurse hours provided in area i in time period t. The binary variable HBR_{it} equals 1 if home-based reablement is offered in the i's area in period t and 0 otherwise. CP-LEAVE_{it} ($\times 100$) represents the sick leave for the contact person working with area i in period t and AREA-LEAVE_{it} ($\times 100$) denotes the sick leave in the respective area in period t. $\boldsymbol{\theta}$ and $\mathbf{z}_{it}$ are $l \times 1$ vectors, where $\mathbf{z}_{it}$ includes a set of up to six time-varying covariates listed in Table 1, and u_{it} denotes the random disturbance associated with area i at time t.

The parameters of interest are the β's in equation (II.1). The monthly effect of implementing HBR on the expected supply of home-nurse hours is captured by β_1. The effect of a *ceteris paribus* one percentage point change in the CP's sick leave on the expected amount of home-nurse hours is captured by β_2. The analog effect of a variation in sick leave in areas is reflected by β_3.

In (II.1), possible unobserved area heterogeneity is captured by a_i, while γ_t accounts for unobserved time effects. In our context, sources for potential unobserved heterogeneity could be, for instance, the CP's age, gender and/or education. The distribution of these characteristics is hardly independent of the regressor CP-sick leave. Consequently, we rely on the fixed-effects (FE) approach. Performing a Wald test on time dummies we reject the null that all time dummies are jointly equal to zero. We therefore estimate a two-way fixed effects model where γ_t in equation (II.1) will account for unobserved time effects (Baltagi, 2013, p. 39). The two-way fixed effects model (II.1) is estimated under the following standard restrictions

$$\sum_i a_i = \sum_t \gamma_t = 0.$$

The standard argument against the FE estimator is that one only uses the within variation. This objection is indeed valid if the within variation is smaller than the between variation. As substantiated in Table II.1, our three variables of interest all have larger within variation than between variation. Thus, in the case at hand, there is no strong argument against using the FE estimator.

II.5 Cross-sectional dependence

We can write the regression in (II.1) more compactly as

$$y_{it} = \alpha + \mathbf{x}'_{it}\boldsymbol{\beta} + a_i + u_{it}, \tag{II.2}$$

where $y_{it} = \text{HOURS}_{it}$, $\mathbf{x'}_{it}$ includes our variables of interest, $\mathbf{z'}_{it}$ and $T-1$ time dummies. The coefficients to be estimated are contained in the $K \times 1$ parameter vector $\boldsymbol{\beta}$, α denotes the constant, and area specific unobserved heterogeneity is captured by a_i as described in Section II.4. It is well known that if the strict exogeneity assumption holds,

$$E(u_{it}|\mathbf{x}_{i1}, \mathbf{x}_{i2}, ..., \mathbf{x}_{iT}) = 0, \tag{II.3}$$

holds, then the traditional FE estimator (or two-way FE estimator if we include a set of $T-1$ time dummies in $\mathbf{x'}_{it}$) will result in unbiased and consistent estimates of $\boldsymbol{\beta}$ even if the errors, u_{it}, are not identically independently distributed. Below we will argue that (II.1) is neither subject to common correlated effects, nor is it likely to be violated by reverse causality (cf. Section II.7).

It is now commonplace in applied empirical research to adjust the standard errors of the parameter estimates for heteroscedasticity and autocorrelation (HAC) of an unknown form by the HAC procedure[10]. Less attention has been directed towards a possible cross-sectional dependence in the errors however, even though, as pointed out by Chudik and Pesaran, 2013, p. 2

> ...the presence of some form of cross-sectional correlation of errors in panel data applications in economics is likely to be the rule rather than the exception.

Cross-sectional dependence (CSD) in our setting means that the error component in equation (II.1) is correlated in the cross-sectional dimension, implying that

$$cov(u_{it}, u_{jt}) \neq 0 \text{ for some } t \text{ and some } i \neq j. \tag{II.4}$$

That is, at each time t the errors may be correlated across some home-nurse areas. With a small N, large T sample, as in our case, we can use either the Breusch and Pagan, 1980 LM test, hereafter referred to as CD_{BP}, or the test developed in Pesaran, 2004, hereafter CD_P, to test for cross-sectionally correlated errors. As demonstrated in Pesaran, 2004, the CD_P has good small-sample properties, and is robust both to structural breaks and unit roots[11].

[10] This adjustment is facilitated by the standard `robust` option in STATA's `xtreg, fe` environment.

[11] The CD_P test is applicable also for large N moderate T samples. As shown in Pesaran, 2015, in this case the test is really a test of H_0: weak cross-sectional dependence.

Table II.2: p-values for various diagnostic tests

	Regression model[a]				
	(1)	(2)	(3)	(4)	(5)
Breusch-Pagan LM test[b]	0.000	0.000	0.000	0.000	0.000
Pesaran CD test[c]	0.000	0.000	0.000	0.000	0.000
Modified Wald test[d]	0.000	0.001	0.050	0.008	0.033
Wooldrige auto-correlation test[e]	0.000	0.000	0.000	0.000	0.000
Observations	396	396	396	385	385

[a] The 5 model specifications are given in Table II.5 in Section II.8.1.
[b] Breusch-Pagan LM test; H_0: cross-sectional independence
[c] Pesaran (2004) cross-sectional dependence test; H_0: independence
[d] Modified Wald test; H_0: homoscedasticity
[e] Wooldridge auto-correlation test; H_0: zero auto-correlation

We report the results of the diagnostic testing for the regression model (II.1) in Table II.2. The estimated coefficients will be presented and discussed in Section II.8. The evidence presented in the 3rd and the 4th row of the table suggests that at conventional α levels, we have to reject the null of homoscedastic errors as well as the null hypothesis of zero autocorrelation in the errors. Moreover, both the CD_{BP} and the CD_P test indicate the presence of CSD in the errors at the 0.01 significance level. To complement the results concerning Pesaran's test we report the CD_P statistic for each variable in $\mathbf{x}_{it}$ in Table II.3 to check whether the right-hand side variables in the regression are CSD. The results in Table II.3 suggest that apart from substitute leader and the first difference of the function score, all covariates included in our model exhibit CSD.

Table II.3: Variable specific Pesaran CD-test for cross-sectional dependence

	CD-test*	p-value
Homenurse hours	2.201	0.028
CP sick-leave	4.880	0.000
Area sick-leave	4.206	0.000
HBR	17.918	0.000
No. home-nurse patients	-1.752	0.080
Substitute leader	-0.229	0.819
Average age	7.128	0.000
Proportion of Men	5.754	0.000
Married or partnership	17.608	0.000
Diff. function score	1.102	0.270

* The CD-statistic is $\mathcal{N}(0,1)$ distributed.
* Extreme values indicates strong correlation

Given these findings, the choice of an appropriate estimation technique will largely depend on how we conceptualize the sources and the nature of CSD. As outlined in

Sarafidis and Wansbeek, 2012 and Chudik and Pesaran, 2013, two dominant approaches have recently evolved and are currently dominating the literature: The *spatial approach* and the *factor structure approach* (or *residual multifactor approach*). The spatial approach assumes that CSD is related to some measure of "distance" between the cross-sections. In our setting, for instance, it might be a common unmeasured shocks generating a spatial connection. In general, the spatial approach requires a pre-specified weight matrix to account for such distances (Anselin, 2001). The factor structure approach assumes that, at each time, there exist a finite number of unobserved common factors affecting the cross-sectional units with diverse unknown intensities. As noted by Sarafidis and Wansbeek, 2012 and Chudik and Pesaran, 2013 this general modelling approach is likely to capture cross-sectional correlation generated by a variety of spatial processes. For that reason we adopt the approach in this paper.

To illustrate, assume errors are subject to the following single-factor error structure:

$$u_{it} = \lambda_i f_t + \varepsilon_{it}. \tag{II.5}$$

Here, f_t is an unobservable common factor, λ_i is the unknown, unit specific factor loading and ε_{it} is a random disturbance distributed independently from $\mathbf{x}_{it}$, λ_i and f_t. Then $cov(u_{it}, u_{jt}) = \lambda_i \lambda_j \sigma_f^2 \neq 0$, where $\mathbb{V}[f_t] = \sigma_f^2$ and we have cross sectional dependence as defined in (II.4). A key question now is whether the cross-sectional correlation observed in the right-hand side variables reported above, is also generated by f_t (albeit with different factor loadings). If so, the strict exogeneity assumption in II.3 could be violated without some further restrictions on the factor structure[12].

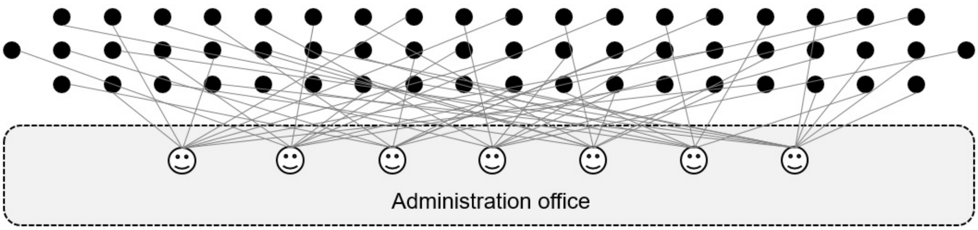

Figure II.4: Each dot represents a new case applying for home-care services. When entering GR1, a case is allocated to a caseworker (smileys) by patients's birth day. Thus the allocation appears to be random for the caseworker, and the decisions of each caseworker will be directed to several different home-nurse areas in an unsystematic way

To solve our estimation problem, one might consider the use of Pesaran's common correlated effects (CCE) estimator (Pesaran (2006)) which yields consistent and

[12]As shown in Chudik and Pesaran, 2013, the presences of common factors does not affect *consistency* of the conventional fixed effect estimator, as long as the common factors are not strong, meaning that the potential unobserved factor affects either only a fixed number of units, or a subset of units is affected by a factor growing slower than N (Sarafidis and Wansbeek, 2012). However, inference is affected since the asymptotic variance is affected by the factor structure of the error term.

asymptotically normal estimates of β even if f_t affects the regressors. Unfortunately, the estimator's asymptotic properties have been established under the assumptions that T is fixed, $N \to \infty$, and $(N, T) \to \infty$ jointly. Hence, this estimator is not applicable in our small N case. To escape the apparent dilemma, we set out to identify traits of the real-life *order-execute system* under scrutiny that may imply a unique (specific) factor structure. We found that the institutional nature of the system in which home-nurse hours are ordered and provided (cf. Section II.2) allows us to make the following identifying assumption: The unknown common factor(s) that generate the CSD in the errors are uncorrelated with the unknown common factor(s) that generate CSD in the regressors.

The institutional background described in Section II.2 motivates the conjecture that the CSD in the errors is primarily generated by each individual case worker allocating hours for different home-nurse areas during a given time period. If the propensity to allocate orders differs between case workers, then errors will be cross-sectionally correlated. However, as illustrated by Figure II.4, cases are distributed to the case workers by the patients' birth day, the pattern of these cross-sectional correlations are likely to be totally independent of the cross-sectional correlations in the regressors. The fixed effects estimator applied to (II.1) will then give unbiased and consistent parameter estimates. Moreover, for robust inference, we may make use of the nonparametric covariance matrix estimator developed by Driscoll and Kraay, 1998 that produces standard-error estimates robust to general forms of heteroscedasticity, autocorrelation, as well as cross-sectional dependence (HACSC).

II.6 Techniques for the correction of standard errors

One of the objectives of this paper is to demonstrate alternative adequate robust inferential procedures in an applied setting. In particular, we will focus on the p-values implied by different algorithms for standard error corrections. To this end, we review three procedures for the correction of standard errors. Proofs will be suppressed, as they can be found in the respective original papers as well as in graduate textbooks as, for instance, Baltagi, 2013, Green, 2003 and Wooldridge, 2002. Our notation is in line with the one established in these references. To obtain the model in equation (II.1) one typically includes $T - 1$ dummy variables if the number of variables is not too large (Green, 2003, p. 291). Performing the "within" transformation and adding the grand averages - to be able to estimate the intercept - on equation (II.2) we get,

$$\tilde{y_{it}} = \alpha + \tilde{\mathbf{x}}_{it}'\boldsymbol{\beta} + \tilde{u}_{it}, \tag{II.6}$$

where

$$\begin{aligned}
\tilde{y_{it}} &= y_{it} - \left(T^{-1}\sum_{t=1}^{T} y_{it}\right) + \left(NT^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T} y_{it}\right) \\
\tilde{\mathbf{x}}_{it} &= \mathbf{x}_{it} - \left(T^{-1}\sum_{t=1}^{T} \mathbf{x}_{it}\right) + \left(NT^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T} \mathbf{x}_{it}\right) \\
\tilde{u_{it}} &= u_{it} - \left(T^{-1}\sum_{t=1}^{T} u_{it}\right) + \left(NT^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T} u_{it}\right).
\end{aligned}$$

Performing OLS on equation (II.6) would provide the FE estimates from STATA's `xtreg, fe`. By stacking the panel as an equation for each individual, one obtains

$$\tilde{\mathbf{y}}_i = \alpha \mathbf{j_T} + \tilde{X}_i \beta + \tilde{\mathbf{u}}_i, \tag{II.7}$$

where $\tilde{\mathbf{y}}_i$ is $T \times 1$, $\tilde{X}_i$ is $T \times k$, $\tilde{\mathbf{u}}_i$ is $T \times 1$, α is a scalar and $\mathbf{j_T}$ is $T \times 1$ vector of ones. Under appropriate regularity conditions, the following asymptotic distribution for the within-group estimator of a linear regression is well known,

$$\sqrt{N}(\hat{\beta} - \beta) \sim \mathcal{N}(0, M^{-1} V M^{-1}), \tag{II.8}$$

where

$$\begin{aligned}
M &= \operatorname*{plim}_{N\to\infty} N^{-1}(\tilde{X}'\tilde{X}) \\
V &= \operatorname*{plim}_{N\to\infty} N^{-1}\sum_{i=1}^{N}(\tilde{X_i}'\mathbf{u}_i\mathbf{u}_i'\tilde{X}_i),
\end{aligned}$$

where $\mathbf{u}_i$ is not observed and needs to be estimated. Assuming homoscedasticity, the standard fixed-effect asymptotic variance-covariance ($Avar$) without any correction can be given as

$$Avar(\hat{\beta}_{FE}) = \hat{\sigma}_u^2 \left[\sum_{i=1}^{N} \tilde{X_i}'\tilde{X}_i\right]^{-1},$$

where

$$\hat{\sigma}_u^2 = [N(T-1) - K]^{-1} \sum_{i=1}^{N} \hat{\mathbf{u}}_i^2,$$

where $\hat{\mathbf{u}}_i$ is the fixed effects residual $\hat{\mathbf{u}}_i \equiv \tilde{\mathbf{y}}_i - \tilde{X}_i\hat{\beta}$.

II.6.1 Heteroscedasticity and autocorrelation

Arellano, 1987 presented a heteroscedasticity and autocorrelation consistent (HAC) standard error robust estimator, based on results due to White, 1984. This estimator is often used as it is generated by the standard `robust` option in Stata's `xtreg,` `fe` environment. Using a White estimator V, Arellano shows that the HAC robust asymptotic variance-covariance estimator of $\hat{\beta}$ is

$$Avar(\hat{\beta}_{HAC}) = (\tilde{X}'\tilde{X}) \left[\sum_{i=1}^{N} \tilde{X}_i{}' \hat{\mathbf{u}}_i \hat{\mathbf{u}}_i' \tilde{X}_i \right] (\tilde{X}'\tilde{X})^{-1}. \tag{II.9}$$

This HAC robust estimator will be valid in the presence of any type of heteroscedasticity and serial correlation (Wooldridge, 2002, p. 275). Since the result has been derived under the assumption of a fixed sample size T and $N \to \infty$, the result only holds for panels with fixed, moderate T and large N.

II.6.2 Heteroscedasticity, autocorrelation and cross-section dependence

As shown by Driscoll and Kraay, 1998, one can obtain standard errors in panel data that are robust to very general forms of correlation in the cross-section, combined with HAC robustness (HACSC). Driscoll and Kraay proposed a simple modification of the standard nonparametric time series covariance matrix estimator relying on large T asymptotics. Their estimator of V in equation (II.8) coincides with that of Newey and West, 1987

$$Avar(\hat{\beta}_{HACSC}) = (\tilde{X}'\tilde{X})\hat{S_T}(\tilde{X}'\tilde{X})^{-1},$$

where S_T follows the Newey and West structure,

$$\hat{S}_T = \hat{\Omega}_0 + \sum_{j=1}^{m(T)} w(j, m(T))[\hat{\Omega}_j + \hat{\Omega}'_j], \tag{II.10}$$

where

$$\begin{aligned} w(j, m(T)) &= 1 - j/[m(T) + 1] \qquad \text{(II.11)} \\ \hat{\Omega}_j &= T^{-1} \sum_{t=j+1}^{T} h_t(\hat{\beta}) h_{t-j}(\hat{\beta})' \\ h_t(\hat{\beta}) &= N(T)^{-1} \sum_{i=1}^{N(T)} h_{it}(\hat{\beta}). \end{aligned}$$

The residuals may be autocorrelated in $m(T)$ lags in equation (II.10), and equation (II.11) gives the modified Bartlett weights (Hoechle, 2007). The lag $m(t)$ has to be chosen. According to Green, 2003, p.200, one typically uses the closest integer to $m(T) \approx T^{1/4}$. Stata's `xtscc` program, which implements Driscoll and Kraay standard errors, uses the following rule when no lag is specified: $m(T) = floor[4(T/100)^{2/9}]$ (Hoechle, 2007). In our case, we use lag 3, the lag implied by these two methods. Results for 8 lags are presented in Appendix II.C and the p-values for 8 lags are presented in Section II.8. Alternatively, one could have used `xtivreg2`, a user-written program provided by Schaffer, 2005, with the `dkraay` option[13].

Note that the Driscoll and Kraay covariance matrix is general and not limited to linear panel models. For the linear case though, one finds

$$h_{it}(\hat{\beta}) = \tilde{\mathbf{x}}_{it}\hat{u}_{it} = \tilde{\mathbf{x}}_{it}(\tilde{y}_i t - \tilde{\mathbf{x}}'_{it}\hat{\beta}).$$

This estimator is precisely the Newey and West HAC consistent covariance matrix applied to the sequence of cross-sectional averages of $h_{it}(\hat{\beta})$ (Driscoll and Kraay, 1998). The result holds also for large $N > T$ as long as T is sufficiently large.

II.6.3 Fixed-b asymptotics

Kiefer and Vogelsang, 2005 introduced a new asymptotic theory for HAC robust tests. When providing results for the usual asymptotic behavior for time-series, it is assumed that the sample size T tends to infinity alongside with the truncation point M. This truncation point is typically defined as, $M = 1 + lags$ where $lags$ refers to the number of lags chosen for the autocorrelated variance adjustment. The standard assumption is that M goes to infinity at a slower rate, such that M/T tends to zero. As Kiefer and Vogelsang point out, in practice there is a given sample size, and even if the practitioner chooses a fraction that goes to zero with increased sample size, the fact is that a positive fraction is used.

Therefore Kiefer and Vogelsang derive the asymptotic distribution for nonparametric HAC estimators under the assumption that $M = bT$ where $b \in (0, 1]$ is a constant. They refer to the asymptotics obtained under this assumption as "fixed-b asymptotics". To make fixed-b asymptotics "user-friendly" they provided new critical values for the t-statistic which depend on b. The fixed-b framework is implemented for the HACSC robust standard errors by Vogelsang, 2012 in a linear panel regression estimated by fixed-effects. Vogelsang arrives at the result that the HACSC robust standard errors by Driscoll and Kraay, 1998 follow the fixed-b asymptotic distribution established in Kiefer and Vogelsang, 2005. One can therefor use fixed-b critical values for t-statistics and the associated confidence intervals in the HACSC setting. The main benefit attributed to fixed-b is, that it delivers test statistics depending on the choice of kernel and bandwidth required to implement the HACSC robust standard errors (Vogelsang, 2012)[14]. The fixed-b critical values (CV) for our regressions will be presented in Section II.8 together with our comparison of p-values.

[13]The `xtscc` and `xtivreg2` programs use different critical values for creating confidence intervals and the reflecting p-value. The standard asymptotic critical value of 1.96 for the 5% significance value is used in `xtivreg2` independent of the sample size. Therefore we prefer to use the `xtscc` program.

[14]The Stata code for calculating fixed-b critical values, associated confidence intervals, and p-values using HACSC robust standard errors due to Driscoll and Kraay, 1998, is available on Vogelsang's web page.

II.7 Ruling out endogeneity

At first glance, one could argue that our model suffers from endogeneity due to reversed causality. An increase in the amount of home-nurse hours could have an effect on the sick leave of CPs. Expert opinion, though, is at odds with this position. When interviewed on the issue, the new incumbent manager of the administration office, as well as his predecessor who had been in the position for 10 years, unanimously expressed the opinion that sick leave effects home-nurse hours but not *vice versa*. Below, we will present our argument ruling out such reversed causality. Our argument rests on organizational features of the administration office described in Section II.2, the NC notes system.

Two circumstances can lead to an unexpected increase in home-nurse hours: i) an unexpected increase in number of patients and ii) existing patients use more home-nurse hours than planned. As shown in Table II.1, the average number of patients has not increased over the sample period. Thus, (i) is not relevant. The cause for increased home-nurse hours in our case is ii). The hours *per* patient have increased over the sample period. As mentioned in Section II.2, a change in any patient's service hours has to pass through the NC-notes system. In the case of reversed causality, there should be a connection between sick leave and the NC notes. For an unexpected increase in hours *per* patient the NC notes could either be more frequent, or home-nurse areas demand more hours in each NC note. We are not able to test the latter. Still it seems reasonable to assume that a change in the content of the NC notes will not necessary influence CP sick leave. Just a change in content will not generate more NC notes, so the number of NC notes *per* patient would be the same. Unless the *content* is significantly more complicated, the NC note related workload for the CP's would not change. It is difficult to argue that NC notes would affect sick leave, if the workload is not changed. However, an increased *amount* of NC notes would change the workload and could therefore affect sick leave. We were able to obtain the yearly amount of NC notes *per* home-nurse area from the municipality at hand[15]: 6,976 NC notes were registered in 2014, 8,758 in 2015 and 9,075 in 2016. A fairly large increase occurs between 2014 and 2015, although the amount of hours actually dropped in this period. If NC notes where to influence CP's sick leave, we would expect an positive correlation between sick leave and NC notes.

Table II.4 presents different correlation coefficients based on yearly data from the eleven home-nurse areas. None of the coefficients is significant and the correlation coefficient in 2015 is negative, which runs counter to what we expected. The changes in NC notes from one year to the next is also not significantly correlated with CP sick leave, and here again one coefficient is negative. Meaning that, for example, increased NC notes compared to the year before will decrease the sick leave compared to the year before. This is opposite to what one would expect under reversed causality. Similar findings surfaced when we focused on the correlation coefficients between yearly changes in CP sick leave and NC notes. This evidence clearly supports our model and notions expressed by the CPs during interviews[16]. During turbulent times the

[15]Unfortunately, we are not able to check the data quality.

[16]To provide statistical evidence in line with our argument we performed variants of a Granger

employees tend to "pushed" away the NC notes to focus their resources on handling standard duties.

Table II.4: Correlation coefficients

	Sum NC notes			Δ Sum NC notes	
	2014	2015	2016	2014 - 2015	2015 - 2016
CP sick-leave 2014	0.218				
	(0.519)				
CP sick-leave 2015		-0.342		0.023	
		(0.304)		(0.947)	
CP sick-leave 2016			0.128		-0.068
			(0.708)		(0.842)
Δ CP sick-leave 2014 - 2015				-0.028	
				(0.935)	
Δ CP sick-leave 2015 - 2016					0.049
					(0.887)

p-values in parentheses
* p<0.10, ** p<0.05, *** p<0.01

II.8 Results

In line with the dual motivation underlying the paper, we present two sets of results. The first set reveals the findings with respect to the problem of absenteeism in care-service provision, while the second subsection focuses on the use of alternative HAC technologies in the case at hand.

II.8.1 Provision of care hours: Estimation results

The FE estimation results for equation II.1, with Driscoll and Kraay HACSC adjusted standard errors and p-values reflecting the fixed-b critical value, are presented in Table II.5. Before commenting on the results for our variables of interest, we observe that none of the coefficients associated with the socio-demographic control variables, average age, proportion of men and patients' partnership status, differ significantly from zero. These results are not surprising, since we do not use patient-level data. Our measurements are aggregated to home-nurse areas. There is little variation in the measurements for these variables and average age, number of males, and patients in partnership are stable in all of the home-nurse areas. Taking into account evidence hinting at the fact that the overall function score is $I(1)$, we include its first order difference which reduces the number of observations to 385. The results for unit-root tests on all variables are presented in Appendix II.B.

Only one of the three binary variables introduced to capture the potential effects of some minor organizational changes - the variable accounting for the reduction in

non-causality test for panels due to Dumitrescu and Hurlin, 2012. A Stata program written by Lopez and Weber, 2017 was used.

area size, cf. Section II.3 - is associated with a coefficient that differs significantly from zero.

Table II.5: Regression results - Fixed-b, lag 3

	(1) Home-nurse hours	(2) Home-nurse hours	(3) Home-nurse hours	(4) Home-nurse hours	(5) Home-nurse hours
HBR	-126.5***	-136.6***	-154.8***	-147.9***	-146.0***
	(0.000)	(0.004)	(0.003)	(0.005)	(0.005)
CP sick-leave	2.047**	1.807**	1.939**	2.277***	2.238***
	(0.013)	(0.030)	(0.017)	(0.008)	(0.008)
Area sick-leave	-5.882**	-6.546**	-9.581**	-9.153**	-9.584***
	(0.028)	(0.033)	(0.016)	(0.018)	(0.011)
No. homenurse	13.21***	12.62***	13.58***	13.98***	13.29***
patients	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Average age		43.20**	29.36*	18.90	18.16
		(0.013)	(0.066)	(0.216)	(0.254)
Proportion men		-573.8	-604.8	-849.0	-1264.8
		(0.530)	(0.505)	(0.355)	(0.169)
Married or		242.2	-256.7	-118.2	334.3
partnership		(0.762)	(0.757)	(0.883)	(0.691)
Substitute leader			190.5**	174.1**	199.8**
			(0.030)	(0.026)	(0.017)
Δ Function score				348.4**	337.2**
				(0.028)	(0.035)
Org. dummies					Yes
Constant	308.6**	-2667.4**	-1649.8	-784.9	-559.5
	(0.035)	(0.032)	(0.183)	(0.506)	(0.643)
Time dummies	Yes	Yes	Yes	Yes	Yes
Observations	396	396	396	385	385
Within R^2	0.495	0.509	0.537	0.547	0.553

p-values in parentheses
* p<0.10, ** p<0.05, *** p<0.01

As expected, we find a negative significant relationship between amount of home-nurse hours and HBR. The coefficient changes marginally when introducing different covariates and the p-value always stays below 0.01. Our estimates suggest that HBR decreased the expected amount of home-nurse hours, approximately by 127 - 155 hours *per* month *per* home-nurse area from the month HBR where implemented. As seen in Table II.5, the monthly average hours *per* area during 2014 - 2016 was 2,494 hours, so the HBR reduction is in the region of 5.5% - 6.3%. Moreover, we find a positive significant effect of the CP sick leave variable with stable p-values and coefficients across alternative specifications. Estimates indicate that a one percent point *ceteris paribus* variation in CP sick leave, leads to a variation of the expected amount of home-nurse hours by 2 hours in the same direction. Based on the dynamics described in Section II.2, this finding confirms the expected effect. At a first glance the coefficient seems to be fairly low. But one has to consider that we only assess the sick leave for CPs. Several employees are actually working with verdicts for home-nurse hours. As was pointed out in Section II.2 the CPs hold the direct contact with the home-nurse areas and work on NC notes. The overall CP sick leave during 2014 - 2016 was 3.4%, so the large jumps during some months will increase the amount of hours substantially.

Interestingly, the sick leave in the home-nurse area has a negative significant effect on home-nurse hours. Everything else fixed, a one percentage point increase (decrease) would decrease (increase) the expected total hours by 9-10 hours *per* month. The main explanation for this finding is that with increased sick leave home nurses have less time to write NC notes. This leads to less adjustments in home-nurse hours. Since most NC notes aim at an upward adjustment, less time to spend on NC notes will reduce the total amount of notes issued. For the home-nurse area this is quite costly. High sick leave will increase cost, but also as shown here, reduce potential income. It is also logical that most of the NC notes are upward adjustments, since one would expect a decrease in health status as patients gets older. Finally, the first difference of the average total function score also has the expected positive significant effect on home-nurse hours. Increasing function scores reflect a patient's decaying health status. A positive (negative) first difference in the average total function score indicates a negative (positive) change of health states in the patient population of the respective home-nurse area. Marginal changes in the first difference should trigger changes in the expected home-care hours in the area in the same direction.

II.8.2 A comparison of HAC correction techniques

After reviewing our results concerning the role of absenteeism for the supply of a care service, we present the related secondary findings demonstrating the effect of alternative HAC techniques on the inference in the case at hand.

Our results are summarized in Tables II.6 - II.8. The information contained in each table refers to the parameter estimate associated with a specific variable. In particular, each table exhibits the p-values associated with the alternative standard error correction methods reviewed in Section II.6: Arellano's HAC, HACSC due to Driscoll and Kraay, 1998, and Fixed-b which refers to HACSC adjusted standard errors with fixed-b critical values proposed by Vogelsang, 2012. For the HACSC version two columns have been included, one for each lag order selection (3 and 8). The rows of the tables refer to the five model specifications corresponding to the columns of Table II.5.

The detailed regression results for the different standard error correction methods are presented in Appendix II.C. To facilitate an easy comparison the tables in the appendix have the same structure as Table II.5. Given one table for each variable of interest, we can now compare p-values for a given parameter across different standard error correction techniques.

As indicated in Table II.3, there is cross-sectional dependence in *CP sick leave*. The CD-statistic is not extreme, but high enough to be significant at the 1% significance level. Given the fact that some CPs work with several home-nurse areas, one would expect some dependence. Nevertheless, the cross-sectional dependence found, is not severe. Since all p-values are less than 0.05, all correction techniques unanimously imply a "reject H_0" decision. But we find convincing evidence for increased precision with HACSC adjusted standard errors. The HACSC adjustment leads to smaller standard errors and p-values, even with 8 lags and fixed-b critical value. Of course the p-values increase in the fixed-b framework because of a higher critical value. The cross-sectional dependence is not severe in this variable, and the HAC works reasonable well in this case. But it still is associated with a loss of precision. The base line strategy of "no-correction" leads to the largest standard errors on this variable.

Table II.6: p-values and standard errors for contact person sick-leave

	No correction	HAC	HACSC, 3 lags	HACSC, 8 lags	Fixed-b, 3 lags	Fixed-b, 8 lags
(1)	0.029	0.020	0.005	0.003	0.013	0.024
	(0.931)	(0.737)	(0.689)	(0.632)	(0.689)	(0.632)
(2)	0.053	0.029	0.015	0.013	0.030	0.057
	(0.930)	(0.710)	(0.709)	(0.691)	(0.709)	(0.691)
(3)	0.033	0.039	0.007	0.005	0.017	0.035
	(0.905)	(0.817)	(0.682)	(0.654)	(0.682)	(0.654)
(4)	0.012	0.027	0.003	0.001	0.008	0.015
	(0.903)	(0.881)	(0.715)	(0.636)	(0.715)	(0.636)
(5)	0.014	0.014	0.003	0.001	0.008	0.013
	(0.902)	(0.750)	(0.698)	(0.607)	(0.698)	(0.607)
Fixed-b CV					2.284	2.737

Standard errors in parentheses

The pattern in p-values for the *area sick leave* closely resembles the one for CP sick leave just described. This does not come as a surprise, once we take into account the fact that the CD statistics given in Table II.3 for area sick leave (4.21) and CP sick leave (4.88) only differ by 0.67.

Table II.7: p-values and standard errors for area sick-leave

	No correction	HAC	HACSC, 3 lags	HACSC, 8 lags	Fixed-b, 3 lags	Fixed-b, 8 lags
(1)	0.002	0.081	0.014	0.010	0.028	0.048
	(1.859)	(3.031)	(2.276)	(2.148)	(2.276)	(2.148)
(2)	0.001	0.024	0.018	0.015	0.033	0.063
	(1.916)	(2.461)	(2.623)	(2.571)	(2.623)	(2.571)
(3)	0.000	0.019	0.007	0.009	0.016	0.045
	(1.981)	(3.425)	(3.351)	(3.436)	(3.351)	(3.436)
(4)	0.000	0.022	0.008	0.008	0.018	0.044
	(1.992)	(3.368)	(3.257)	(3.232)	(3.257)	(3.232)
(5)	0.000	0.023	0.004	0.004	0.011	0.032
	(2.017)	(3.580)	(3.148)	(3.124)	(3.148)	(3.124)
Fixed-b CV					2.284	2.737

Standard errors in parentheses

The p-values associated with the cases [specification (1), HAC] and [specification (2), Fixed-b 8 lags] exceed the threshold value of 0.05. All other p-values lie below 0.05. The HACSC adjusted standard errors still tend to increase precision relative to HAC adjusted errors. If no correction is carried out, the standard error seems to be underestimated - the strength of evidence appears to be inflated. The opposite was true for the case of CP sick leave. This shows how problematic regular FE model standard errors are when standard assumptions are not fulfilled. HACSC with the fixed-b values do produce slightly higher p-values compared with HACSC due to the new critical value. Still with fixed-b and three lags adjustment, the values are lower than those for HAC.

Compared to the CD statistic on the other two variables of interest, the CD-statistic for *home-based reablement* turns out to be high. The value of 18.1 indicates strong cross-sectional correlation. Such strong correlation is in line with the information concerning HBR, its implementation and practice (four teams work across several home-nurse areas), outlined in Section II.2. Thus, by design, the HBR variable will exhibit quite severe cross-sectional dependence. One would therefore expect some variation in the p-values once we control for the phenomenon.

Table II.8: p-values and standard errors for HBR

	No correction	HAC	HACSC, 3 lags	HACSC, 8 lags	Fixed-b, 3 lags	Fixed-b, 8 lags
(1)	0.000	0.211	0.000	0.000	0.000	0.004
	(30.17)	(94.66)	(27.51)	(28.95)	(27.42)	(29.41)
(2)	0.000	0.217	0.001	0.003	0.004	0.027
	(30.94)	(103.73)	(38.46)	(43.30)	(38.40)	(43.30)
(3)	0.000	0.151	0.001	0.003	0.003	0.026
	(30.36)	(99.55)	(43.07)	(48.63)	(43.93)	(49.92)
(4)	0.000	0.173	0.002	0.004	0.005	0.030
	(30.57)	(100.69)	(43.09)	(47.54)	(43.72)	(48.29)
(5)	0.000	0.168	0.001	0.003	0.005	0.028
	(31.04)	(98.11)	(42.08)	(46.38)	(42.35)	(4.58)
Fixed-b CV					2.284	2.737

Standard errors in parentheses

The standard errors estimated with no-correction are considerably lower than the ones obtained under some type of correction. One detects the same pattern in the case of the p-values. Economists - almost by default - implement the HAC adjusted standard errors, i.e. they report the output from STATA's `robust` option. In the case at hand, this alternative results in the largest standard errors and p-values among all of the alternative correction strategies. If an analyst had based the inference on the HAC option, her/his conclusion would have been completely at odds with the decisions implied by all other alternatives. When one uses the HACSC correction, standard errors and p-values drop, even with an 8 lags adjustment. It seems that the HAC option struggles to handle the strong cross-sectional dependence in the HBR variable. A second reason for the clear difference between HAC and the others can be related to the fact that the HAC is based on large N and small T asymptotics. As with others, one can see how the fixed-b critical values increases when the lag adjustment increases. This is reflected in the higher p-values for the HACSC standard error with the fixed-b critical values implemented.

II.9 Robustness check

Most graduate econometric textbooks like, Wooldridge, 2002 and Cameron and Trivedi, 2005, focus on short panels, where $N \to \infty$ and T is fixed. The asymptotic properties of standard panel estimators, such as FE, hold for these short panel dimensions. Asymptotic analysis is valid for arbitrary time dependence alongside distributional heterogeneity across t with fixed T (Wooldridge, 2002, p. 250). Wooldridge, p. 250 points out that it is suitable to view the cross-section observations as independent and identically distributed with large N asymptotics, i.e. one would not have to consider CSD. As argued above, CSD could be controlled for using HACSC adjusted standard errors in the standard FE model. Based on simulated data exhibiting traditional HACSC problems, Vogelsang, 2012 presents results for an FE model with HACSC adjusted errors, with and without fixed-b critical values. According to his findings, the FE model with HACSC adjusted errors alongside fixed-b critical values provided superior results. This holds for long panels ($T > N$), and even for fairly small samples. As described by Vogelsang, with fixed T there is not enough information in the time dimension relative to the cross-section dimension for Driscoll and Kraay's approach to work.

Some econometricians would argue that choosing the FE estimator for a long panel is not appropriate, even if the standard errors are corrected using a technique based on long-panel asymptotics. An alternative estimator, which is based on fixed N and $T \to \infty$, is Zellner's *seemingly unrelated regression (SUR)* model (Zellner, 1962; Zellner, 1963; Zellner and Huang, 1962). The SUR model specifies one linear regression equation for each of the N home-nurse areas, each with T observations and an individual error term. For our case that would imply 11 equations, with 36 observations. One stacks these N equations to obtain the SUR model. The equation specific error terms are assumed to have a zero mean, to be i.i.d and homoskedastic for each (across) home-nurse area. The individual errors, however, are assumed to be correlated across home-nurse areas (Cameron and Trivedi, 2010, p. 163). Typically, the feasible generalized least-squares (FGLS) principle is invoked to estimate the parameters of the SUR model.

SUR models are often estimated with heterogeneous treatment effect, but in Sections II.2 and II.4 we argue for and present a regression model with a homogeneous treatment effect. To ensure comparability between the model presented in Section II.4 and the SUR model used in the robustness check, we impose linear coefficient constraints on the SUR model. In the SUR context one cannot not use time dummies, because the full rank assumption would not be fulfilled. Our FE model has time dummies to capture potential seasonal effects. To allow for some seasonal effects in the SUR model, heterogenous quarterly dummies have been included.

The estimation results for alternative specifications of the SUR models are presented in Table II.9. All parameters of interest are significant on the 5% level. The estimated effects of both sick leave parameters are of the same magnitude as the results presented in Section II.8.1. The estimated effect of HBR is, however, smaller in the SUR model. Still the parameter estimate from the SUR model suggests that HBR has lead to a reduction in expected home-nurse hours. Our findings from the SUR model are in line with our results presented in Section II.8.1. They clearly support our initial findings. The SUR model will not necessary capture the potential auto-correlation within each home-nurse area and as mentioned, it assumes homoskedasticity within equation variance. This assumption is not necessary unrealistic in our case, but in case this assumption is not fulfilled on can bootstrap the standard errors for the SUR model.

Table II.9 also presents our SUR model with bootstrapped errors. The significance level associated with HBR drops for specification (1), and the respective levels drop for CP sick leave for regression models (2) and (3). By and large the results presented replicated and all parameters of interest are still significant at the 10% level.

Table II.9: SUR model results - with linear constraints

	(1) Home-nurse hours	(2) Home-nurse hours	(3) Home-nurse hours	(4) Home-nurse hours
HBR	-45.13***	-60.04***	-55.95***	-60.00***
	(0.000)	(0.000)	(0.000)	(0.001)
	[0.017]	[0.002]	[0.002]	[0.000]
CP sick leave	2.226***	1.746**	1.764**	1.841**
	(0.000)	(0.017)	(0.012)	(0.013)
	[0.035]	[0.071]	[0.068]	[0.036]
Area sick leave	-5.002***	-4.934***	-5.519***	-5.943***
	(0.000)	(0.000)	(0.000)	(0.000)
	[0.006]	[0.005]	[0.003]	[0.001]
No. home-nurse patients	12.23***	11.38***	11.54***	12.03***
	(0.000)	(0.000)	(0.000)	(0.000)
	[0.000]	[0.000]	[0.000]	[0.000]
Average age		47.15***	38.68***	29.61***
		(0.000)	(0.000)	(0.002
		[0.000]	[0.002]	[0.009]
Proportion of Men		-2.508	94.97	97.70
		(0.995)	(0.816)	(0.825)
		[0.997]	[0.872]	[0.860]
Married or partnership		879.3***	841.3***	762.8**
		(0.002)	(0.004)	(0.011)
		[0.019]	[0.028]	[0.036]
Substitute leader			86.36***	62.94**
			(0.009)	(0.048)
			[0.052]	[0.144]
Diff. function score				338.4***
				(0.001)
				[0.008]
Heterogeneous constant	Yes	Yes	Yes	Yes
Quarterly dummies	Yes	Yes	Yes	Yes
Observations	396	396	396	385

p-values in parentheses

p-values in brackets based on bootstrapped errors, 400 repetitions, seed = 1234

* p<0.10, ** p<0.05, *** p<0.01, based on standard errors, not bootstrapped

II.10 Discussion and conclusion

In the course of their theoretical investigation of absenteeism, Pauly et al., 2002 find for the case of team production and substantial team-specific human capital, that the value of output lost due to absence will exceed the wage per day of the absent employee. I.e. the employer faces a cost component to absenteeism which exceeds that of wages lost. Our empirical findings support this prediction. For the case of an administrative office/agency in the Norwegian health sector which organizes its care service production in teams relying in a crucial manner on team-specific-human capital, we find that due to a *ceteris paribus* increase in CP sick leave one can expect an inflation in the amount of home-nurse hours allocated.

Our findings with respect to CP sick leave are in line with empirical results presented by Nicholson et al., 2006. The authors approximate job-specific multipliers that reflect cost of sick leave which goes beyond the employees daily wage. For instance, a registered nurse at a physician's office and a receptionist at a physician's office, have estimated 3-day cost of absence multipliers of 1.28 and 1.27 respectively. That is, 3 days of absence cost 28% more than daily wages alone.

Even though our CP's job profiles do not match the profiles for the jobs considered in Nicholson et al., 2006 exactly, we may calculate a CP specific multiplier. For the CP's, 3 days of absence in one month is equivalent to 13.6% sick leave. If the expected monthly sick leave value is zero, based on result in Table II.5, 3 days extra absence for one CP would translate into 27 extra home-nurse hours for the specific month *ceteris paribus*. With cost per home-nurse hour of approximately 600 NOK, the extra month specific cost over and above daily wages would be 16,400 NOK, which translate into a multiplier of 1.30[17]. The order of magnitude of this estimate closely matches the multipliers for health related jobs reported in Nicholson et al., 2006.

When interviewed on the topic, the management of the administrative office pointed out that the unit had operated under a stationary, yet barely sufficient staffing level throughout the sample period (despite the challenges coming from the demographic development). This fact alone may have contributed to the observed level of sick leave. In addition, in a setting where it is difficult to substitute absentees due to the relevance of team-specific-human capital, the negative effects of sick leave will be severe. In such a situation, even a slight increase in manpower could actually reduce sick leave, and allow the unit to respond in a robust manner to additional stressors as, for instance, sick leave.

The latter results are especially interesting in the light of our finding that increased sick leave in the home-nurse area tends to reduce the amount of home-nurse hours. Unlike in the case of the administrative unit, the management of the home-nurse area can rely on a pool of back-up employees in the case of unexpected sick leave notices.

Even if the stand-in is a close substitute for an experienced nurse, (s)he might still have to operate under more severe time constraints (e.g. due to the time elapsing between the registration of a sick leave note and the time at which the substitute becomes available). Given the reduced time frame, a nurse will most likely prioritize the care function over the administrative effort associated with writing the note. In any case, the substitute may not be perfect. Often stand-ins are younger home-nurses with little prior exposure to the NC note system. They might therefore act extremely careful, in addition to being even more time constrained than a perfect substitute

[17]Based on an annual salary of 500,000 NOK, which is multiplied with 30% to account for public pension and Norwegian payroll tax. After adjustment, annual cost equals 650,000 NOK, and monthly 54,167 NOK. Cost due to extra hours $(16,400/54,000) + 1 = 1.3$. All cost are based on 2018 NOK.

simply due to a lack of familiarity with the cases (s)he is assigned to. As a result such substitutes will be biased against writing NC notes. If no substitute is available, then the work schedules of nurses currently on duty will be readjusted such that the appropriate level of care can be received by all patients. Also in this case, it is likely that the administrative chore of writing NC notes will not be high on the list of a home nurse's priorities. Thus, one should expect a reduction in home-nurse hours as a consequence of sick leave.

The findings presented in Table II.5 also show that, everything else being equal, the introduction of HBR leads to a reduction in expected home-nurse hours. One might think of this result as being indicative of the fact that the new care strategy (treatment) works as intended, and therefore patients need less home-nurse hours. However, we cannot rule out the reduction in home-nurse hours overstates the reduction in need for home-care services. The multidisciplinary HBR team typically consists of a physiotherapist, an occupational therapists, a nurse and a home helper. Once such teams become operational in a home-nurse area they might absorb some functions/chores that previously had been administered by home-nurses. For instance, a physical therapist might pay an extra visit to see how the patient is doing. Thereby (s)he fulfills a function typically performed by a home nurse. Since certain occupations in the HBR teams are not financed by the pay-for-performance system introduced in Section II.2 but operate under a regular budget, the available data do not allow us to account for such phenomena.

The use of HAC adjusted standard errors seems to be the "norm" in applied research using FE estimation. Table II.2 clearly shows that our standard error is CSD alongside with HAC. Failing to take this into account and base inference on the "norm", would yield less convincing results. As clearly shown in Tables II.6 - II.8, all HACSC adjusted standard errors produce lower p-values compared to the "norm" HAC. This usually also holds when applying HACSC adjustment with more restrictive fixed-b critical values. For the most restrictive fixed-b setup, lag 8, one could see a slight tendency for higher p-values, but overall they are still lower than those of the HAC version. Using the standard "norm" would not yield different inference for the two sick leave variables, but the evidence would be less convincing. Using the "norm" on HBR would, however, yield totally different inference. Seen in Table II.3, the HBR variable has an extreme CD statistic and the two others have not. The "norm" apparently fails to take this strong dependence into account. We are able to identify the above because of the randomness in case allocation illustrated in Figure II.4. Without this key feature, one could not necessarily rule out that the factors causing CSD in the variables and the standard errors, are not correlated. If such a relationship existed, the above method would yield inconsistent coefficients. The case at hand demonstrates that, in the context of applied econometrics, the knowledge of the underlying system is crucial for the choice of an adequate statistical technique. Intricate system knowledge is essential for the modelling process as well as for optimal inference.

To our knowledge, there does not exist any applicable method handling CSD without the randomness for a small N large T case. This is something we would like to see in the future, but also acknowledge the difficulties developing such methods. For future research it would be intriguing to develop theoretical models for such administration offices based on traditional *principal-agent-theory*. Extending theoretical work from Pauly et al., 2002 to also include the effects of non-comparable substitutes would be appealing and relevant for the "real world".

II.11 References

Anselin, Luc (2001). "Spatial Econometrics". In: *A Companion to Theoretical Econometrics*. John Wiley & Sons Ltd. Chap. 14, pp. 310–330.

Arellano, M. (1987). "Computing Robust Standard Errors for Within-groups Estimators." In: *Oxford Bulletin of Economics & Statistics* 49.4, pp. 431–434.

Baltagi, Badi H (2013). *Econometric analysis of panel data*. eng. 5th ed. Chichester: Wiley.

Barham, Catherine and Nasima Begum (2005). *Sickness absence from work in the UK*.

Barmby, Tim A., Marco G. Ercolani, and John G. Treble (2002). "Sickness Absence. An international comparison". In: *The Economic Journal* 112.480, pp. 315–331.

Bauer, Annette et al. (2019). "Cost-minimisation analysis of home care reablement for older people in England: A modelling study". In: *Health & Social Care in the Community* 27.5, pp. 1241–1250.

Breusch, T. S. and A. R. Pagan (1980). "The Lagrange Multiplier Test and its Applications to Model Specification in Econometrics". In: *The Review of Economic Studies* 47.1, pp. 239–253.

Brown, Sarah and John G. Sessions (1996). "The economics of absence: Theory and Evidence". In: *Journal of Economic Surveys* 10.1, p. 23.

Cameron, A. and Pravin Trivedi (2010). *Microeconometrics Using Stata, Revised Edition*. StataCorp LP.

Cameron, Adrian and Pravin Trivedi (Jan. 2005). *Microeconometrics: Methods and Applications*. Cambridge University Press.

Chatterji, Monojit and Colin J. Tilley (2002). "Sickness, Absenteeism, Presenteeism, and Sick Pay". In: *Oxford Economic Papers* 54.4, pp. 669–687.

Chudik, Alexander and M Pesaran (2013). *Large panel data models with cross-sectional dependence: a survey*. Globalization and Monetary Policy Institute Working Paper 153. Federal Reserve Bank of Dallas.

Driscoll, John C. and Aart C. Kraay (1998). "Consistent Covariance Matrix Estimation with Spatially Dependent Panel Data". In: *The Review of Economics and Statistics* 80.4, pp. 549–560.

Dumitrescu, Elena-Ivona and Christophe Hurlin (2012). "Testing for Granger non-causality in heterogeneous panels". In: *Economic Modelling* 29.4, pp. 1450–1460.

Fersch, Barbara (2015). "Expectations towards home care re-ablement in Danish municipalities". eng. In: *International Journal of Sociology and Social Policy* 35.3/4, pp. 126–140.

Godøy, Anna and Harald Dale-Olsen (2018). "Spillovers from gatekeeping – Peer effects in absenteeism". In: *Journal of Public Economics* 167, pp. 190 –204.

Green, William H. (2003). *Econometric analysis*. 5th. Prentice Hall. Chap. 5 and Appendix E.

Hesselius, Patrik, Per Johansson, and Peter Nilsson (2009). "Sick of Your Colleagues' Absence?." In: *Journal of the European Economic Association* 7.2/3, pp. 1–15.

Hoechle, Daniel (2007). "Robust standard errors for panel regressions with cross-sectional dependence". In: *Stata Journal* 7.3, pp. 281–312.

Im, Kyung So, M.Hashem Pesaran, and Yongcheol Shin (2003). "Testing for unit roots in heterogeneous panels". In: *Journal of Econometrics* 115.1, pp. 53 –74.

Kandel, Eugene and Edward P. Lazear (1992). "Peer pressure and partnerships." In: *Journal of Political Economy* 100.4, p. 801.

Kiefer, Nicholas M. and Timothy J. Vogelsang (2005). "A New Asymptotic Theory for Heteroskedasticity-Autocorrelation Robust Tests". In: *Econometric Theory* 21.6, pp. 1130–1164.

Kjerstad, Egil and Hanne Tuntland (2016). "Reablement in community-dwelling older adults: a cost-effectiveness analysis alongside a randomized controlled trial". In: *Health Economics Review* 6.1, pp. 1–10.

Levin, Andrew, Chien-Fu Lin, and Chia-Shang James Chu (2002). "Unit root tests in panel data: asymptotic and finite-sample properties". In: *Journal of Econometrics* 108.1, pp. 1 –24.

Lewin, Gill et al. (2014). "A comparison of the home-care and healthcare service use and costs of older Australians randomised to receive a restorative or a conventional home-care service". In: *Health & Social Care in the Community* 22.3, pp. 328–336.

Lopez, Luciano and Sylvain Weber (2017). *Testing for Granger causality in panel data*. IRENE Working Papers 17-03. IRENE Institute of Economic Research.

Markussen, Simen (2012). "The individual cost of sick leave". In: *Journal of Population Economics* 25.4, pp. 1287 –1306.

Newey, Whitney K. and Kenneth D. West (1987). "A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix". In: *Econometrica* 55.3, pp. 703–708.

Nicholson, Sean et al. (2006). "Measuring the effects of work loss on productivity with team production". In: *Health Economics* 15.2, pp. 111–123.

Paola, Maria De (2010). "Absenteeism and peer interaction effects: Evidence from an Italian Public Institute". In: *The Journal of Socio-Economics* 39.3, pp. 420 –428.

Pauly, Mark V. et al. (2002). "A general model of the impact of absenteeism on employers and employees". In: *Health Economics* 11.3, pp. 221–231.

Pesaran, M. Hashem (2004). *'General Diagnostic Tests for Cross Section Dependence in Panels'*. Cambridge Working Papers in Economics 0435. Faculty of Economics, University of Cambridge.

— (2015). "Testing Weak Cross-Sectional Dependence in Large Panels". In: *Econometric Reviews* 34.6-10, pp. 1088 –1116.

Rotemberg, Julio J. (1994). "Human relations in the workplace." In: *Journal of Political Economy* 102.4, p. 684.

Sarafidis, Vasilis and Tom Wansbeek (2012). "Cross-Sectional Dependence in Panel Data Analysis." In: *Econometric Reviews* 31.5, pp. 483–531.

Schaffer, Mark E (2005). *XTIVREG2: Stata module to perform extended IV/2SLS, GMM and AC/HAC, LIML and k-class regression for panel data models*. Statistical Software Components, Boston College Department of Economics.

Vogelsang, Timothy J. (2012). "Heteroskedasticity, autocorrelation, and spatial correlation robust inference in linear panel models with fixed-effects". In: *Journal of Econometrics* 166.2, pp. 303 –319.

White, H. (1984). *Asymptotic theory for econometricians*. Economic Theory, Econometrics and Mathematical Economics Series. Academic Press.

Winkelmann, Rainer (1999). "Wages, firm size and absenteeism". In: *Applied Economics Letters* 6.6, pp. 337–341.

Wooldridge, Jeffry M. (2002). *Econometric analysis of cross section and panel data*. Cambridge, Mass: MIT press. Chap. 20.

Zellner, Arnold (1962). "An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests for Aggregation Bias". In: *Journal of the American Statistical Association* 57.298, pp. 348–368.

— (1963). "Estimators for Seemingly Unrelated Regression Equations: Some Exact Finite Sample Results". In: *Journal of the American Statistical Association* 58.304, pp. 977–992.

Zellner, Arnold and David S. Huang (1962). "Further Properties of Efficient Estimators for Seemingly Unrelated Regression Equations". In: *International Economic Review* 3.3, pp. 300–313.

II.A Appendix A: Variable list

Table II.A.1: Variable list

Variable	**Comment**
Home-nurse hours	Sum allocated home-nurse hour per month for each area
CP sick leave	Contact-persons sick leave, originally measured in percentage, but multiplied with 100 for easier interpretation
Area sick leave	Area sick leave, adjusted as above
Number of home-nurse patients	The number of home-nurse patients per month for each area, whom generated the sum hours
Average age	Average age for patients in each area
Proportion of Men	The percentage of male patients for each area
Married or in partnership	The percentage of patients whom are married or in a partnership for each area
Substitute leader	Binary variable equal 1 in the periods for the areas using substitute leader
Function score	Mean function score for each area per month
Organizational dummy - zone1	Binary variable capturing increase in geographical size
Organizational dummy - zone2	Binary variable capturing decrease in geographical size
Organizational dummy - apartments	Binary variable reduction in responsibility for service apartments

II.B Appendix B: Unit-root tests

Table II.B.1: Unit-root test results, reported p-values

	Levin-Lin-Chu test[a]		Im-Pesaran-Shin test[b]	
	BIC†	AIC††	BIC†	AIC††
Homenurse hours*	0.0191	0.0086	0.0587	0.0315
No. home-nurse patients*	0.0007	0.0013	0.0021	0.0029
CP sick-leave**	0.0000	0.0000	0.0000	0.0000
Area sick-leave	0.0000	0.0000	0.0000	0.0000
Average age	0.0004	0.0004	0.0009	0.0009
Proportion of Men	0.0281	0.0168	0.0439	0.0255
Married or partnership*	0.0021	0.0023	0.0305	0.0275
Function score	0.3335	0.2630	0.2810	0.2069
Diff. function score	0.0000	0.0000	0.0000	0.0000

[a]Recommended for moderate sample size, T grows faster then N, H_0 = contains unit root
[b]Justified by using sequential limit theory, works best with a large T and at least moderate N
†Bayesian information criterion (BIC) for specifying lags
††Akaike information criterion (AIC) for specifying lags
*Linear time trend included
**CS averages subtracted from the series to relax the CS independence assumption
CS = Cross-sectional

Both test, Levin et al., 2002 and Im et al., 2003, have the null hypothesis that all panels contain a unit root. We clearly can not reject the null for function score, and will therefor use the first difference of function score in our analysis. All other have low p-values for all different version the two unit-root tests, and we reject the null and conclude that all other variables are stationary.

II.C Appendix C: Results - different standard error correction

Table II.C.1: Regression results - no SE adjustment

	(1) Home-nurse hours	(2) Home-nurse hours	(3) Home-nurse hours	(4) Home-nurse hours	(5) Home-nurse hours
HBR	-126.5***	-136.6***	-154.8***	-147.9***	-146.0***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
CP sick-leave	2.047**	1.807*	1.939**	2.277**	2.238**
	(0.029)	(0.053)	(0.033)	(0.012)	(0.014)
Area sick-leave	-5.882***	-6.546***	-9.581***	-9.153***	-9.584***
	(0.002)	(0.001)	(0.000)	(0.000)	(0.000)
No home-nurse patients	13.21***	12.62***	13.58***	13.97***	13.29***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Average age		43.20***	29.36**	18.90	18.16
		(0.004)	(0.049)	(0.207)	(0.226)
Proportion men		-573.8	-604.8	-849.0	-1264.8**
		(0.280)	(0.242)	(0.105)	(0.030)
Married or partnership		242.2	-256.7	-118.2	334.3
		(0.588)	(0.567)	(0.796)	(0.512)
Substitute leader			190.5***	174.1***	199.8***
			(0.000)	(0.000)	(0.000)
Δ Function score				348.4***	337.2***
				(0.006)	(0.008)
Org. dummies					Yes
Constant	308.6**	-2667.3**	-1649.8	-784.9	-559.5
	(0.043)	(0.017)	(0.136)	(0.481)	(0.620)
Time dummies	Yes	Yes	Yes	Yes	Yes
Observations	396	396	396	385	385
Within R^2	0.495	0.509	0.537	0.547	0.553

p-values in parentheses

* p<0.10, ** p<0.05, *** p<0.01

Table II.C.2: Regression results - HAC adjusted (robust)

	(1) Home-nurse hours	(2) Home-nurse hours	(3) Home-nurse hours	(4) Home-nurse hours	(5) Home-nurse hours
HBR	-126.5	-136.6	-154.8	-147.9	-146.0
	(0.211)	(0.217)	(0.151)	(0.173)	(0.168)
CP sick-leave	2.047**	1.807**	1.939**	2.277**	2.238**
	(0.020)	(0.029)	(0.039)	(0.027)	(0.014)
Area sick-leave	-5.882*	-6.546**	-9.581**	-9.153**	-9.584**
	(0.081)	(0.024)	(0.019)	(0.022)	(0.023)
No home-nurse patients	13.21***	12.62***	13.58***	13.97***	13.29***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Average age		43.20	29.36	18.90	18.16
		(0.189)	(0.188)	(0.389)	(0.472)
Proportion men		-573.8	-604.8	-849.0	-1264.8
		(0.643)	(0.614)	(0.480)	(0.404)
Married or partnership		242.2	-256.7	-118.2	334.3
		(0.791)	(0.780)	(0.909)	(0.749)
Substitute leader			190.5**	174.1**	199.8**
			(0.020)	(0.016)	(0.028)
Δ Function score				348.4**	337.2**
				(0.049)	(0.045)
Org. dummies					Yes
Constant	308.6	-2667.3	-1649.8	-784.9	-559.5
	(0.136)	(0.263)	(0.302)	(0.617)	(0.753)
Time dummies	Yes	Yes	Yes	Yes	Yes
Observations	396	396	396	385	385
Within R^2	0.495	0.509	0.537	0.547	0.553

p-values in parentheses

* p<0.10, ** p<0.05, *** p<0.01

Table II.C.3: Regression results - HACSC, lag 3

	(1) Home-nurse hours	(2) Home-nurse hours	(3) Home-nurse hours	(4) Home-nurse hours	(5) Home-nurse hours
HBR	-126.5***	-136.6***	-154.8***	-147.9***	-146.0***
	(0.000)	(0.001)	(0.001)	(0.002)	(0.001)
CP sick-leave	2.047***	1.807**	1.939***	2.277***	2.238***
	(0.005)	(0.015)	(0.007)	(0.003)	(0.003)
Area sick-leave	-5.882**	-6.546**	-9.581***	-9.153***	-9.584***
	(0.014)	(0.017)	(0.007)	(0.008)	(0.004)
No home-nurse patients	13.21***	12.62***	13.58***	13.97***	13.29***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Average age		43.20***	29.36**	18.90	18.16
		(0.006)	(0.041)	(0.175)	(0.213)
Proportion men		-573.8	-604.8	-849.0	-1264.8
		(0.504)	(0.476)	(0.317)	(0.131)
Married or partnership		242.2	-256.7	-118.2	334.3
		(0.750)	(0.746)	(0.878)	(0.675)
Substitute leader			190.5**	174.1**	199.8***
			(0.015)	(0.013)	(0.007)
Δ Function score				348.4**	337.2**
				(0.014)	(0.018)
Org. dummies					Yes
Constant	308.6**	-2667.3**	-1649.8	-784.9	-559.5
	(0.019)	(0.017)	(0.144)	(0.477)	(0.623)
Time dummies	Yes	Yes	Yes	Yes	Yes
Observations	396	396	396	385	385
Within R^2	0.495	0.509	0.537	0.547	0.553

p-values in parentheses
* p<0.10, ** p<0.05, *** p<0.01

Table II.C.4: Regression results - HACSC, lag 8

	(1) Home-nurse hours	(2) Home-nurse hours	(3) Home-nurse hours	(4) Home-nurse hours	(5) Home-nurse hours
HBR	-126.5***	-136.6***	-154.8***	-147.9***	-146.0***
	(0.000)	(0.003)	(0.003)	(0.004)	(0.003)
CP sick-leave	2.047***	1.807**	1.939***	2.277***	2.238***
	(0.003)	(0.013)	(0.005)	(0.001)	(0.001)
Area sick-leave	-5.882***	-6.546**	-9.581***	-9.153***	-9.584***
	(0.010)	(0.015)	(0.009)	(0.008)	(0.004)
No home-nurse patients	13.21***	12.62***	13.58***	13.97***	13.29***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Average age		43.20***	29.36**	18.90	18.16
		(0.004)	(0.039)	(0.205)	(0.241)
Proportion men		-573.8	-604.8	-849.0	-1264.8
		(0.474)	(0.467)	(0.310)	(0.110)
Married or partnership		242.2	-256.7	-118.2	334.3
		(0.753)	(0.745)	(0.878)	(0.686)
Substitute leader			190.5**	174.1**	199.8***
			(0.019)	(0.018)	(0.008)
Δ Function score				348.4***	337.2***
				(0.002)	(0.003)
Org. dummies					Yes
Constant	308.6***	-2667.3**	-1649.8	-784.9	-559.5
	(0.002)	(0.021)	(0.185)	(0.541)	(0.665)
Time dummies	Yes	Yes	Yes	Yes	Yes
Observations	396	396	396	385	385
Within R^2	0.495	0.509	0.537	0.547	0.553

p-values in parentheses

* p<0.10, ** p<0.05, *** p<0.01

Table II.C.5: Regression results - Fixed-b, lag 8

	(1) Home-nurse hours	(2) Home-nurse hours	(3) Home-nurse hours	(4) Home-nurse hours	(5) Home-nurse hours
HBR	-126.5***	-136.6**	-154.8**	-147.9**	-146.0**
	(0.004)	(0.027)	(0.026)	(0.030)	(0.028)
CP sick-leave	2.047**	1.807*	1.939**	2.277**	2.238**
	(0.024)	(0.057)	(0.035)	(0.015)	(0.013)
Area sick-leave	-5.882**	-6.546*	-9.581**	-9.153**	-9.584**
	(0.048)	(0.063)	(0.045)	(0.044)	(0.032)
No. home-nurse patients	13.21***	12.62***	13.58***	13.98***	13.29***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Average age		43.20**	29.36	18.90	18.16
		(0.029)	(0.107)	(0.301)	(0.336)
Proportion men		-573.8	-604.8	-849.0	-1264.8
		(0.539)	(0.533)	(0.398)	(0.203)
Married or partnership		242.2	-256.7	-118.2	334.3
		(0.778)	(0.771)	(0.890)	(0.722)
Substitute leader			190.5*	174.1*	199.8**
			(0.070)	(0.070)	(0.043)
Δ Function score				348.4**	337.2**
				(0.022)	(0.025)
Org. dummies					Yes
Constant	308.6**	-2667.4*	-1649.8	-784.9	-559.5
	(0.018)	(0.075)	(0.280)	(0.598)	(0.703)
Time dummies	Yes	Yes	Yes	Yes	Yes
Observations	396	396	396	385	385
Within R^2	0.495	0.509	0.537	0.547	0.553

p-values in parentheses

* p<0.10, ** p<0.05, *** p<0.01

Paper III

Creating a unique panel from Norwegian health register data: Technicalities and difficulties

Creating a unique panel from Norwegian health register data: Technicalities and difficulties

Tore Bersvendsen[1*] and Jochen Jungeilges[1,2]

[1]Department of Economics and Finance, University of Agder, Norway
[2]Institute of Mathematics and Natural Sciences Ural Federal University, Ekaterinburg, Russian Federation

To facilitate an economic assessment of the innovative care intervention *home based reablement* (HBR) a large unique dataset constituting a unique individual monthly panel covering the time-period from 2011 to 2015 was constructed on the basis of patient information drawn from Norwegian health registers. We describe the complex process of transferring the data from the sources to the researcher and discuss the choice of data imputation strategies needed to obtain a research dataset that was informative with respect to the cost assessment problem at hand. Apart from enhancing the replicability of our HBR research, we discuss institutional matters, legal aspects, organizational issues, IT-problems and statistical issues that constitute the main challenges for researchers trying to construct large research datasets of individual data.

Keywords: Imputation, panel data, register data

JEL classification: C81, C89

Acknowledgment: We would thank Eirik Abildsnes for valuable comments. This work was performed on the TSD (Tjeneste for Sensitive Data) facilities, owned by the University of Oslo, operated and developed by the TSD service group at the University of Oslo, IT-Department (USIT). (tsd-drift@usit.uio.no).

Disclaimer: Data from the Norwegian Patient Register has been used in this publication. The interpretation and reporting of these data are the sole responsibility of the authors, and no endorsement by the Norwegian Patient Register is intended nor should be inferred.

**Corresponding author:* Tore Bersvendsen, tore.bersvendsen@uia.no

III.1 Introduction

Over the past decades we have witnessed an explosion in the quantity and quality of data bases holding information about individuals. The public discourse resonating this development has generated a wide spectrum of views on the phenomenon. Portrayed by some as a curse, others see a blessing in our ability to amass individual-level data. For science in general, and for social scientists/economists in particular, such rich datasets provide additional opportunities to answer new relevant economic research questions. In response to the growing availability of individual data, distinct subfields of statistics/econometrics have developed. This is substantiated in the following quotation taken from Cameron and Trivedi, 2005, p. 4

> Processing and econometric analysis of such large microdatabases, with the objective of uncovering patterns of economic behavior, constitutes the core of microeconometrics.

The panel-data methodology developed in microeconometrics, for example, provides the methodological backbone of a project which tries to asses an innovative care intervention (home based reablement HBR) offered in Norwegian municipalities. Our effort to address a new research problem in the context of health economics involves a large dataset. The insight into the economic aspects of the care strategy that we provide today might be considered by policy makers in designing an economically viable and sustainable future care sector. We believe that society might benefit from the use of a large set of individual patient data in the case at hand.

This paper documents the construction of the dataset we used in the HBR study. Focussing on the various dimensions of the construction process - institutional matters, legal aspects, organizational issues, IT-problems and statistical issues - we hope to provide a guideline/orientation for researchers planning this type of work in the future.

Moreover, by means of this effort we try to increase the replicability of the research performed under the HBR project. Reviewing the relevant literature, we found that many papers do not give sufficient attention to the process in which data were collected/generated (Bersvendsen et al., 2019). Probably due to constraints set by journals and/or editors the process of data generation often remains a black box for outsiders. By providing an informative description of the data production process, we aim at eradicating this typical deficit.

Finally, we intent to provide a unique view of the statistical technique of data imputation as a solution to an apparent dilemma experienced by many researchers. Owners of sensible private information (individual data) have to adhere to legal rules protecting the private sphere of individuals when data are made available to researchers. Therefore data are anonymized. The datasets received by the researcher often allow, at best, for suboptimal statistical inference. Thus the information available for the policy maker will not be optimal. The subsequent decision making may not be in the best interest of society. We demonstrate how statistical imputation techniques can be used to reconstruct a research dataset with the goal to improve the precision of statistical inference.

Among economist and econometricians, the extra benefits of precise and rich panel data are well known. Being able to control for individual heterogeneity, increased efficiency and the possibility to study dynamics are all desirable benefits of panel data (Baltagi, 2013, p. 6). The panel data literature has predominantly been focussed on micro panels with a large number of cross-sections (N) and a small time dimension (T). With an increased availability of data, there has been a movement towards models with both large N and T. The fairly new models by Pesaran, 2006, Bai, 2009 and Chudik and Pesaran, 2015, which adjust for unobserved common factors potentially correlated with the regressors, are good examples for such large N and T models. The properties of both old and new statistical estimators, are typically only known in the asymptotic case (Green, 2003; Baltagi et al., 2008).

The dataset we constructed is large in terms of both N and T. It was created based on information held in several separate Norwegian health registers. Norway is in a rather rare situation that key services as schooling, social welfare, and health, are exclusively provided by the public sphere. All national register data is therefor often mandatory, of high quality, and covers almost the entire Norwegian population. These features definitely characterize the four health registers that delivered the data which, after being merged, constituted an unique individual monthly panel covering the time-period from 2011 to 2015.

Although initially created to answer cost questions regarding a new rehabilitation intervention referred to as HBR, the data could certainly be used to find solutions to a variety of research problems. Even though Norway has amassed interesting data, the extraction and merging processes are cumbersome. In our case, not all necessary support systems were in place. Thus new routines had to be designed in order to generate the dataset. Certain data operations/transformations and manipulations could only be carried out by the organizations owning the data. The activities of those organizations have to be coordinated. In essence, the researcher relies to a large degree on the willingness and the ability of the data owning organizations to cooperate and share their resources.

The time elapsed between the first application to the availability of the final workable dataset amounted to almost three years. There is definitely room for organisational improvements. After all, in the current institutional setting characterizing the research landscape, such a time-consuming process might be viewed as prohibitive (or at least as a disincentive) for a researcher. The latter issue has recently been elaborated on in Norwegian print media[1].

A part of finding the right balance, legal requirements needed the dataset to be anonymized. In the course of this process some observations had to be deleted or *suppressed*. An algorithm based on four variables was used to mark observations for suppression, and thereafter the associated service data where deleted for the suppressed observations. Creating a univariate missing data pattern. Since one can observe the four variables used in the algorithm, the missing value process is defined as *missing at random* (MAR). Meaning that, conditional on observed data, the probability of data missing does not depend on unobserved data. From a theoretical point of view MAR

[1] www.aftenposten.no/meninger/debatt

is a desirable property and most imputation techniques and their implementations (statistical software) assume such a pattern (Buuren et al., 1999; Morris et al., 2014; *Stata 15 Base Reference Manual* 2017).

We tried to counter the negative statistical effects of data suppression by applying imputation techniques. Below we will document the accuracy of conventional imputation techniques in a big-data context. In a comparison of various imputation techniques Kleinke et al., 2011 found that standard methods like *fully conditional specification* (FCS), with the likes of *Predictive Mean Matching* (PMM), where hardly outperformed by new imputation techniques specifically designed for panel data. Thus we choose this method as our principal imputation strategy. Depending on the nature of the variables being imputed adequate variants of the PMM have been employed (Raghunathan et al., 2001).

The operational characteristics of the imputation techniques to be employed were carefully assessed. All test imputation proved accurate for the mean and standard deviation. The accuracy is so high that for instance on binary variables, the mean and standard deviation is identical up to the fourth decimal. However, all techniques underestimate the frequency of service usages, resulting in an overestimated "within" standard deviation. With a large dataset, it is argued, that it is beneficial to have only one complete dataset. Interestingly, with great variable insight, simple interpolation proved to work remarkably well for ceratin variables. Test results for the latter, showed that one is able to perfectly estimate the true value in 97% of the case with a 50% missing rate. When applied, the imputation strategy is used on 1.9% of the total 7,951,682 observations on 154,839 patients.

Documenting the process of constructing the dataset, we naturally focus on the HBR context. However, the process could be replicated for other new interventions of special interest, like new home-based e-health services. Apparently the value-added of our contribution can be derived from the fact that the process described below can be implemented in a wide range of health care related research problems. Future researchers can benefit from the experiences we outline below.

In Section III.2, both the data collection and the anonymization process are explained in detail. Section III.3 documents how adequate imputation strategies were identified. In each case a careful discussion of test results allows for an assessment of the quality of the imputation process leading to the complete dataset. A short summary of the complete dataset is given in Section III.4. A short discussion in Section 5 finalizes the paper.

III.2 Data collection and anonymization

III.2.1 Bird's eye view of the data collection process

Norway has several health registers containing individual health data on both the primary- and specialist care level. These data repositories have been assembled at high cost by the society. Data from selected registers constitute the basis of our investigation into the short- and long term effects of HBR, which in Norway is provided by municipalities. To be specific, we rely on data from the following health registers, (i) Registry for Individual-based Nursing and Care Statistics (IPLOS), (ii) Norwegian Patient Registry (NPR), (iii) Register for Control and Payment of Primary care Reimbursement Scheme (KUHR) and (iv) Statistics Norway (SSB). By connecting multiple data records from these sources, a unique panel-data set is created. Since measurements are taken at an individual level, the analyst can keep track of an individual's use of various health-care services over time. The different health registers as well as the information extracted from them is described in detail in Section III.2.1.1.

The extracted population, contains data points collected over 5 years, from 2011 up to and including 2015, in selected municipalities. Since none of the registers listed above holds information concerning a patient's use of HBR, the Norwegian personal number (P.no) of persons who had received HBR had to be collected "manually" by the author. Due to the apparent difficulties/cost caused by collecting HBR data in such a way for all Norwegian municipalities, only observations from selected administrative entities have been included in the study.

Apart from feasibility considerations, a set of explicit criteria governed the municipalities selection process. In particular we applied (i) a binary treatment criterion, (ii) a size criterion as well as (iii) an experience criterion. Under (i) only municipalities that have implemented the HBR strategy alongside a conventional care strategy qualify for inclusion. Under the size criterion (ii) fairly populous municipalities become candidates for selection. Thereby we aim at a high potential variation on the municipality level. Finally, to avoid capturing predominantly transient phenomena, the experience criterion (iii) favors the inclusion of municipalities that have gathered sufficient experience with administering the care strategy itself and that have accumulated "know how" concerning the management of the respective HBR units. Taking these criteria as well as feasibility considerations into account, the following 10 Norwegian municipalities were selected: Arendal, Bærum, Bergen, Drammen, Kristiansand, Oslo, Sandefjord, Trondheim and Tromsø.

III.2.1.1 Health registers and data

Various types of health related data flow from different stakeholders to various health related data registers. These registers operate under the auspices of the Norwegian Directorate of Health (NDH). They serve different purposes and, to some degree, play different roles in the Norwegian health system. Each register has its own data processing unit[2]. The three registers that became central data repositories for our project will be described below. In each case, we present a short profile of the institution and describe the data we received from the respective register[3].

Registry for Individual-based Nursing and Care Statistics (IPLOS): The purpose of this register is to collect and manage data from Norwegian municipalities on persons who have applied for, receive or have received health and care services. The reporting to IPLOS is mandatory for municipalities. In our context, the essential to point out

[2] For the relevant information see Norwegian Parliament proposition 106 L (2015-2016)

[3] A detailed introduction to the registers - written in Norwegian - can be found here

that the register relies on pseudonyms rather than on personal numbers (P.no) to identify individuals[4].

IPLOS uses data to monitor, to assure quality of service, to plan development and overall management of the health and care services. Data is also collected to provide a basis for research. In short, IPLOS contains three types of information on individuals: (i) background information such as gender, age, municipality of residence etc, (ii) information to assess service needs, mainly living situation and function score, and (iii) service information. For (iii), IPLOS contains 26 different service types. Due to anonymization issues, we could not access specific data on all the 26 different types. Instead, the information has been aggregated (grouped) into 3 variables. The first variable, measured in hours, contains service type 1, 2, 3, and 15. The *first* three types are different versions of home care and type 15 is traditional home service such as home-nurse services. All service types measured in amount of days are summarized into the *second variable*. It includes three different short-term institution service types 18, 19, 20 as well as service type 21 referring to stays in a long-term institution. The *third* variable is binary, indicating whether an individual received one or the other service type.

Norwegian Patient Registry (NPR): According to the NPR regulation, the motives for creating the registry coincide with those outlined in the IPLOS context. The main difference, however, is that NPR is designed for specialist care. That is, NPR mainly collects and interprets data provided by Norwegian hospitals. In addition, NPR data is used in a national activity based funding mechanism for somatic services. The hospitals are financed by the amount of Diagnosis-related Groups (DRG) points, which are administered by NPR.

The information held in the register resembles the information contained in IPLOS to some degree. But, as one would expect, there is a stronger emphasis on detailed medical information. Our project relied on NPR data concerning hospital stays (number of days spend in a hospital per month) and visits to polyclinics (number of visits per month) differentiated according to somatic or psychiatric. Moreover, we drew on information about a patient's main diagnosis group.

Register for Control and Payment of Primary care Reimbursement Scheme (KUHR): While NPR and IPLOS are both defined as health registers, KUHR is not. KUHR has been established to administer the national finance system for general practitioners (GP) and other health professionals, who provide services and receive reimbursements founded on the Norwegian National Insurance act. The compensation for GPs and physiotherapists is based on a set of detailed activities defined by sperate codes. The service provider links these codes to a patient's ID and transfers the data to KUHR which is responsible for storing and managing this information. We requested information concerning different activities of GPs and physiotherapists from KUHR. The categories were adequately aggregated per individual on a monthly basis.

HBR is targeted toward recipients of home care services, which in Norway is mainly provided by municipalities. IPLOS is the health register containing all data from services provided by municipalities, and therefore it is the natural "starting point" for

[4]In 2018, IPLOS has been replaced by a new register called *municipal patient and user register*.

creating the dataset. Data from NPR and KUHR need to be connected to the sample drawn from IPLOS. Connecting the registers in this order, and merging data volumes characteristic for our case, is not common and posed substantial difficulties. Studies based on health register data in Norway typically take a data sample from NPR, and thereafter merge/match the sample with data from other registers. Several routines and technical solutions are available to support such a protocol. Similar solutions do not exist for the case in which the initial sample is extracted from IPLOS. The root of the complications lies in the fact that NPR and KUHR both use the Norwegian P.no as the identification key, whereas IPLOS uses pseudonyms.

In addition to data from IPLOS, NPR, and KUHR some socio-demographic variables were collected from Statistics Norway (SSB). As will be pointed out below, SSB also played a key role in the process of merging the data coming in from IPLOS, NPR, and KUHR.

As pointed out above, the three national data registers have been designed to serve different purposes, and they clearly differ in many respects. One of those differences becomes apparent to the researcher as (s)he applies for data. Since KUHR is not classified as a health register in the strict sense, researchers applying for data from the registry face ethical requirements that are less stringent than those formulated by IPLOS and NPR.

III.2.1.2 Application process

When confronted with the idea of combining data from the said registers, a senior adviser in one of the health registers initially responded by stating *...that is not possible*[5]. Thus, from the outset of the project phase, it was clear that by trying to combine data from IPLOS, NPR, and KUHR we were moving into unknown territory. To our knowledge there was no precedence for the type of data merger we were trying to facilitate. Complications were anticipated but the spectrum of problems we encountered was, by far, broader than what we expected. The complications we experienced were basically caused by institutional factors and by technical modalities.

With respect to the former, we understand that there may be good legal and/or political reasons to design and maintain three different health related registers. However, seen from the perspective of the researcher relying on the registers as a data source, the heterogeneity of the institutions poses a hinderance. There is, for instance, no uniform application protocol. Administrative aspects of application procedures vary, so do legal and ethical requirements. Moreover, also our *post*-application experiences suggests that the rules regulating the different registers do not necessarily facilitate the cooperation between the registers. To the outsider it appears that the registers were not primarily designed to work together.

To exemplify the effect of such conditions on the progress of the project, we document the time line of the process from the first application to the final delivery of the workable dataset in Table III.1. A glance at the table reveals that the entire process, took almost three years.

[5]Taken from phone conversation.

Typically a significant amount of administrative work had to be completed in periods between the key events listed in Table III.1. The whole process could have been shortened substantially, if NPR had not used 11 months to transfer the data to SSB. According to the invoice from NPR, the institution used exactly 24 hours to complete the project during the time period marked by the signing the contract and the time of data delivery. Because of this delay, the data merger and anonymization process assigned to SSB had to be postponed.

The details given in the table should allow for a sufficient estimate of the cost in terms of waiting time and frustration experienced by the researchers. Evidently, these cost are not attributable to administrative aspects alone. Rather, they arose out of the interplay between administrative and technical factors. Since it maybe instructive/constructive for researchers trying to merge data from the said registers in the future, we devote the subsequent section to technical challenges and their solutions.

Table III.1: Time line of the application process

Date	Event
22.09.15	Application send to *The Regional Committees of Medical and Health Research Ethics* (REK)
11.11.15	REK verdict: application not approved
19.02.16	Application to *Norwegian Center for Research data - data protection service* (NSD) and updated REK application submitted
07.04.16	REK application approved
04.05.16	NSD application approved and forwarded to *The Norwegian Data Protection Authority* (NDPA) for further consideration
10.06.16	Application for KUHR data send
14.07.16	NDPA application approved
17.07.16	Application for NPR data send - submission of a missing document delayed the NPR process by 2 monhts
21.07.16	Application for IPLOS data send
29.07.16	Application for SSB data send
22.08.16	KUHR application approved
25.08.16	IPLOS application approved
03.10.16	SSB assigns the case worker responsible for data merger and anonymization
07.12.16	Signed agreement send to SSB
15.12.16	NPR application approved
05.01.17	Signed agreement send to NPR - waiting for NPR case-worker
09.02.17	KUHR data send to SSB
13.02.17	HBR data send to SSB
08.11.17	NPR data send to SSB
09.03.18	SSB delivers complete dataset to researcher (1st time)
04.04.18	SSB delivers data (2nd time), after correcting an error discovered by researcher
16.04.18	SSB delivers data (3rd time), after correcting an error in the second version discovered by resarcher
31.05.18	Last meeting with SSB; discussion of potential suppression errors in latest dataset

III.2.1.3 Data linkage

When data are transferred from the municipalities to IPLOS, they are sent with P.no as the identification key. Before IPLOS can receive the data, the P.no has to be converted into pseudonyms. This is done by a credible pseudonym administrator (CPA), which is The Norwegian Tax Administration in the case of IPLOS. According to the Norwegian health register law, the CPA shall be independent of both the data processor and the data owner. The Norwegian Directorate of Health is the owner of IPLOS data, but the data process responsibilities have been delegated to SSB. NPR and KUHR do not operate with these pseudonyms. Since the conversion of pseudonyms to IPLOS is not reversible, the P.no's in the data records from KUHR and NPR need to be substituted by adequate pseudonyms before connecting with IPLOS.

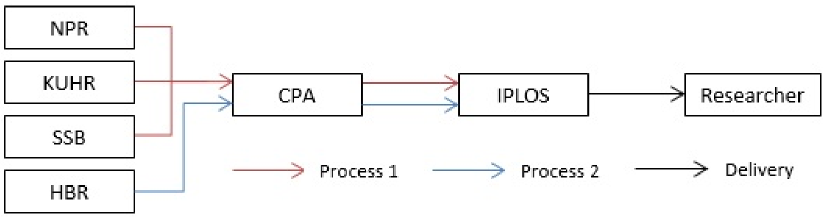

Figure III.1: Data flow and connection process

As Figure III.1 illustrates, the data connection and flow is decomposed into two processes. In process 1, NPR, KUHR and SSB send data on the respective requested variables for all registered inhabitants directly to the CPA for replacing P.no with pseudonymes. This is done only for the designated municipalities. The records from these registers that do not match with the IPLOS population are dropped. Process 2 only differs in that the researcher sends the HBR data to the CPA because it has been collected manually.

Process 1: IPLOS uses an encryption software called *RegKrypt*. This software should always be used when connecting other data sources with IPLOS. RegKrypt will encrypt a comma separated file with register data and generate an *xml* file that is sent to IPLOS for connecting the data. NPR, KUHR and SSB will run RegKrypt on a computer in a safe zone, i.e. a zone that is isolated from the internet. RegKrypt will then encrypt the data on the individual level with an encryption key from IPLOS. This encrypted file is then transferred to a computer connected to the internet via the sender module *RegBestilling* through a web-address at CPA. The part of the file that contains the P.no will then be decrypted by CPA and replaced with pseudonyms. The CPA does not have the possibility to decrypt other data than P.no. Thereafter the file is sent to IPLOS where an administrator will decrypt the file and merge the NPR, KUHR and SSB data with IPLOS data. This administrator is working for SSB since SSB has the data process responsibilities on the behalf of The Norwegian Directorate of Health. A more detailed description in Norwegian can be found at SSB's webpage[6].

[6] `https://www.ssb.no/innrapportering/offentlig-sektor/_attachment/127052?_ts=`

Process 2: To deliver the collected HBR data, we had to follow a protocol similar to the one described above. However, some details concerning the collection of HBR data should be outlined. There where three variables related to HBR users that had to be collected directly from the municipalities' patient journals: i) P.no for all the HBR user and ii) the date on which the patient's HBR treatment began and iii) the date on which the respective treatment was terminated. One researcher travelled to all municipalities included study. In each location, he met with the contact person for the respective municipality. Each contact person hade prepared a data file with the necessary information. This file was uploaded to the researcher's computer which uses BitLocker encryption. Here the files where encrypted with the *VeraCrypt* software for safe storing. The resulting files where deleted once delivered to IPLOS and the anonymization dialog with IPLOS was finalized.

III.2.2 Data storage

Eventually, the dataset was stored at the data base service for sensitive data 2.0 (TSD). This IT-platform is used to collect, store and analyze sensitive research data in a secure environment. TSD has been developed and is operated by the University of Oslo (UiO). It is a part of NorStore, the national infrastructure for handling and storing of scientific data. For more information concerning TSD, the reader is referred to UiO's webpage[7]. In particular, TSD provides a two-factor log-on system and is used by different research environments, including university hospitals. But before the final workable dataset could be transferred to TSD, it had to be transformed to satisfy crucial legal requirements aiming at the anonymity of those individuals whose data were to enter the research process.

III.2.3 Anonymization

The verdict of The Norwegian Data Protection Authority requires the dataset to be anonymized prior to delivery to the researchers. SSB handled the anonymization task as a part of the data process agreement with IPLOS. SSB operates on the basis of a pragmatic rule of thumb: If each cluster of identifiable variables contains at least five individuals then the dataset fulfills the anonymity requirement.

All data send from the registers had the identical panel structure that had been specified by the researchers and accepted by SSB. The data were delivered on a monthly time scale. Meaning each observation (line) in the dataset represents one month in one specific year for one individual. Prior to applying an anonymization algorithm, SSB had to merge the data it received from the registers and the researcher. As data realizations coming from other registers are not necessarily concurrent in time with IPLOS data, SSB generated empty observations for all time periods where an individual did not have IPLOS services. As a result, there are 60 observations (5 years) associated with each individual independent of the number of observed IPLOS periods. Then the additional columns holding data from the other registers could easily be

14a2e2417c8

[7]`https://www.uio.no/english/services/it/research/sensitive-data/`

added. Prior to this process the HBR group was indicated via dummy variables and HBR days during a month were added. Observations following an individuals death month were deleted, and therefore exceptions from the 60 observation rule may occur.

Once these tasks were completed, an anonymization algorithm could be applied. This algorithm proceeds in two steps. First, it identifies clusters of less then 5 individuals in the space spanned by four identifiable variables: HBR group (yes/no), Gender, Hospital area and Age. Apparently it was impossible to achieve anonymity with the raw detailed age information. Therefore, the informational content of the original variable had to be reduced by imposing a coarse scheme of 15 age groups. The cluster step was replicated for each of the 60 time periods. Each of the observations that fell into a cluster was marked by a dummy variable. In the second step, all service data where deleted for all observations marked on the previous step.

As mentioned above, ”empty observations” had been generated to facilitate the process of merging data from the various sources. If these observations still were empty after merging, they were interpreted as being associated with individuals who were not in need of health services at the given point in time. In some incidences such ”empty” records were found to lie in clusters of less than 5 individuals. Thus, they were labeled accordingly. SSB considered the knowledge that an individual coming from the IPLOS population is healthy in a given month as extra information and therefore insisted on labeling such observations for deletion. Out of the 8,561,088 total observation, 180,036 (2,1%) observations where marked and service data where deleted in 144,979 cases. This implies that there are 35,057 marked observations that originally were ”empty” observations. In the dataset handed to the researcher, the pseudonym-identification key had to be replaced with an ID variable.

III.2.3.1 An exemplification of the anonymization process

The anonymization process is best illustrated by help of an example. For ease of notation and readability, the example is restricted to two individuals observed in the IPLOS register, for two years (2014 - 2015). Moreover, we assume quarterly data. In the first step, SSB creates a panel data structure from the IPLOS observations, as presented in Table III.2. The entries in the different service columns are just chosen for illustrative purposes.

Table III.2: Initial panel

ID	YEAR	QTR	IPLOS
1	2014	2	3
1	2014	3	3
1	2015	4	20
2	2015	1	10
2	2015	2	10

Observe that individual 1 had three quarters of IPLOS services, while individual 2 used services in two quarters during 2014-2015. Individual 1 had two separated spells of IPLOS services, with one quarter in the end of 2015 and the rest in 2014. Once a panel data structure has been created from IPLOS data, data from NPR and KUHR can be added. Table III.3 illustrates how this has been done by SSB.

Table III.3: SSB production step

(a) Panel with surrogate observations

ID	YEAR	QTR	IPLOS
1	2014	1	-
1	2014	2	3
1	2014	3	3
1	2014	4	-
1	2015	1	-
1	2015	2	-
1	2015	3	-
1	2015	4	20
2	2014	1	-
2	2014	2	-
2	2014	3	-
2	2014	4	-
2	2015	1	10
2	2015	2	10
2	2015	3	-
2	2015	4	-

(b) Full panel after merger

ID	YEAR	QTR	IPLOS	NPR	KHUR
1	2014	1	-	-	2
1	2014	2	3	-	-
1	2014	3	3	-	1
1	2014	4	-	-	-
1	2015	1	-	-	-
1	2015	2	-	1	-
1	2015	3	-	4	-
1	2015	4	20	-	-
2	2014	1	-	-	-
2	2014	2	-	-	1
2	2014	3	-	-	-
2	2014	4	-	2	-
2	2015	1	10	-	-
2	2015	2	10	-	-
2	2015	3	-	3	-
2	2015	4	-	-	-

Since NPR and KUHR observations can potentially occur in other time periods than IPLOS, SSB first had to create the full time line for all individuals included in the IPLOS sample. This is seen in Table III.3a, where the gray observations illustrate the SSB generated observations/rows. Once these rows were generated, SSB could easily add data from NPR and KUHR into new columns since they can use ID and time as common keys. This is illustrated in Table III.3b where imaginary NPR and KUHR columns have been added. Once these columns have been added, SSB can run its anonymization algorithm described above. The main idea behind the algorithm is to identify clusters with less then 5 individuals on the basis of four observed variables. Every observation which is identified by the algorithm is marked by the help of the binary variable "suppressed". A value of 1 on this variable indicates that this record was "scheduled for deletion". Eventually all entries except those on variables used by the algorithm, were deleted. After deletion, the dataset could be transferred to the researcher.

Table III.4 shows the state of the data upon completion of the anonymization process, i.e. the state in which the dataset was handed over to the researcher. The nature of the outcome generated by the anonymization process can be assessed by comparing Table III.3b and Table III.4. For example, the suppression indicator for individual 1 in 3rd quarter of 2014 is set equal to 1 in Table III.4. Referring to the same observation in Table III.3b, one can see that IPLOS data were deleted. As pointed out above, 180,036 observations are marked with "suppressed = 1" in the final dataset. Exactly 144,979 of these observations resemble the case of individual 1 in the example. The rest of the suppressed observations, 35,057 in total, are observations for which no data are deleted. To illustrate this case, we focus on individual 2 in the 3rd quarter of 2014 in Table 2b and Table 3. Here the individual is still suppressed, although originally no service entries had been observed in the given period. The researcher only sees Table III.4, and therefore the researcher does not know with certainty whether or not some data entries were deleted in these two examples. A value of 1 on the suppressed indicator only means that the respective observation was marked by the algorithm. This event is independent of the health state of the respective individual.

Table III.4: Panel dataset after anonymization

ID	YEAR	QTR	IPLOS	NPR	KHUR	SUPPRESSED
1	2014	1	-	-	2	0
1	2014	2	3	-	-	0
1	2014	3	-	-	-	1
1	2014	4	-	-	-	1
1	2015	1	-	-	-	1
1	2015	2	-	1	-	0
1	2015	3	-	4	-	0
1	2015	4	20	-	-	0
2	2014	1	-	-	-	0
2	2014	2	-	-	-	1
2	2014	3	-	-	-	1
2	2014	4	-	2	-	0
2	2015	1	10	-	-	0
2	2015	2	-	-	-	1
2	2015	3	-	3	-	0
2	2015	4	-	-	-	0

The dataset delivered by SSB contained three flaws that had to be corrected. First, the observations of six unique patients occurred twice, since they had been registered with two different genders. For these observations, one set was deleted. The selection was randomized, controlling for the general gender distribution within the patients age group. Secondly, there were duplicated observations for nine patients. They had been registered with two different death months. The records of these patients had to be checked to find out which of the two different death months showed registered service data. The version in which some service data had been registered in the death month was kept. Cases in which some service data were registered in both death months did not occur.

The third problem was caused by a programming error. It had been agreed that SSB would generate observation lines only for the months up to and including a patient's month of death. No records of the patient should exist in the aftermath of her/his death. The dataset delivered showed clear evidence that this concept had not been properly realized. For example, one patient could have died in April 2012. As agreed, there were no observations during the remainder of 2012. However, the same patient was registered again from 2013 to 2015 with empty observations from January to April, or until the month of registered death. Thus, there where additional observations for every deceased patient in every year following her/his death. The number of extra months *per* year depended on the month in which the individual had died. All of these records had no service data attached to it. They where merely an artifact due to an unfortunate programming error on behalf of SSB. Evidently, they had to be deleted. After eradicating these errors, the total number of observations was reduced to 7,981,709. This includes the "empty observations" with the interpretation as "healthy observations" explained in Section III.2.3.

III.3 Imputation strategy

III.3.1 Delete extreme suppressed

The special characteristics of IPLOS data will be presented in Section III.3.3. Here, we point at a key characteristic. All patients' IPLOS data are characterized by a slow moving dynamics, i.e. the IPLOS data only vary slightly in time. Actually, just by scrolling trough patient records that have more then 90% of observations suppressed, one can see a clear pattern. If data have been observed early in the 60-month time horizon and some at a later stage, then it is possible to impute missing data with fairly great accuracy. Usually, the former and the latter observations indicate identical IPLOS services. Moreover, patients do not tend to drop out of the IPLOS service. Knowing this, all suppressed observations should most likely have the same IPLOS value. If the two observations are different, then it is highly likely that there has only been one change. To predict the time at which the change occurs is, however, more difficult to predict. The latter aspect constitutes the key weakness of the imputation strategy in general.

After careful inspection of the dataset and consideration of the findings in Subsections III.3.3.1 and III.3.3, all patients that have at least 93% of their observations suppressed were dropped. Proceeding in this way, we secured at least 4 observations for all suppressed patients, which will be sufficient to impute sensible estimates of function scores and IPLOS service. The number of potentially problematic patients fairly small, and it should not affect the average mean and variation. It should be emphasized that the 93% cut-off value is a subjective assessment based on first hand knowledge and insight into the data. Table III.5 shows the changes in sample size, i.e. with respect to observations and patients, as the extremely suppressed cases are deleted.

Table III.5: Deleting patients with 93% or more observation suppressed

	N	NT
Before deleting extreme	155,380	7,981,709
HBR data deleted	489	29,340
Non-HBR data delete	52	687
HBR data remaining	1,741	104,460
Non-HBR data remaining	153,098	7,847,222
Total after deleting	154,839	7,951,682

As documented by Table III.5, the HBR group is heavily affected by the deletion strategy outlined above. Most patient data have been deleted from this group. In total, 541 patients constituting 30,027 observations have been dropped. Of those, a subset of 489 patients belongs to the HBR group. The total amount of patients drops by 0.35% and observations are reduced by 0.38%. After deletion, we were left with a total of 7,951,682 observations. Of those 150,584 (1.89%) were suppressed and had to be imputed.

III.3.2 Creating test data

For the purpose of choosing and testing different imputation techniques, we created a "test dataset". It was created in two steps. In the first step, all patients with complete records in the original dataset were extracted. All patients with either one or more suppressed observations were deleted. In all, 6,800 patients were deleted, which left us with a test dataset containing 7,559,867 observations based on 148,039 patients. In the second step, 50% of the observations (i.e. 3,779,933) where randomly marked to simulate suppression. Thereafter, service data for all observations where duplicated into a set D_C, and those service data in D_C that had been randomly marked as suppressed were deleted. To assess the performance of a given imputation technique, we applied it to D_C and compared the respective imputation outcomes to the true values found in D_T.

For the group of 6,800 patients with at least one observation, the mean suppression rate is 38.4%. The chosen missing rate of 50% may therefore seem a bit excessive. However, for the HBR patients, who form a sub-group of those 6,800 patients, the missing rate is 58.3%. As seen in Section III.3.1, the overall percentage of suppressed observations is only 1.89%. Under a conventional approach these observations would simply be dropped prior to data analysis. The extra effort of developing or choosing an imputation strategy would thus be avoided. However, in the current case, the 40.5% of the suppressed observations occur in the group of interest: the HBR patients. All HBR patients have at least one observation suppressed. To just drop all suppressed data would lead to an excessive loss of HBR related information. Therefore an imputation strategy that aims at minimize the loss of HBR data is needed. The missing rate of 50% chosen for the test data, constitutes a reasonalbe compromise between the true overall and the HBR specific missing rate.

III.3.3 IPLOS data

The anonymization algorithm was applied each month t, independent of s where is $s \in T$ and $s \neq t$. Since patients may die the sample size can vary each month. Therefore, running the algorithm each month could lead to a variation in whether a patients' observation is suppressed or not. So it is apparently possible that for a given patient a series of suppressed observations is followed by a sequence of non-suppressed observations, which again maybe followed be a series of suppressed observations. That is, we might find "gaps" in observed data. Since a patient can have several sequences of suppressed observations, one would like to study the average length of these "gaps".

Lets define j different spells, with the spell pattern being a) *first spell* is not suppressed, b) *second spell* is suppressed, c) *third spell* is not suppressed, and d) *fourth spell* is suppressed and so on. In such a series of spells the patient's first observation, January 2011, cannot be suppressed. But this may not necessarily be true. Therefore, two datasets have to be created which are similar in the sense that they contain only patients who have at least one suppressed observation. One set contains patients whose initial observation has been suppressed while the other is constituted by patients whose initial observation has not been suppressed. When January 2011 is suppressed the spell pattern will be i) *first spell* is suppressed, ii) *second spell* is not suppressed, iii) *third spell* is suppressed, and iv) *fourth spell* is not suppressed and so on.

Table III.6: Key statistics for spells

(a) Initial observation NOT suppressed

	All	HBR	Other
Average no. spells	4.85	7.20	3.61
Average length	5.54	4.02	7.37

(b) Initial observation suppressed

	All	HBR	Other
Average no. spells	4.67	5.03	5.53
Average length	4.35	4.91	4.08

Basic estimates of the mean number of spells and the mean length of a spell are presented in Table III.6. The data under scrutiny have been split into two sets, as explained above. Table III.6a presents statistics for the case in which the first observation is not suppressed while Table III.6b refers to the opposite situation. As seen in Table III.6a there is a clear difference in the number of spells between HBR patients and other patients. Indicating that, HBR patients who are not suppressed in January 2011 have a greater tendency to move back and forth between not suppressed and suppressed. This is as expected, since the probability of becoming suppressed should increase if a patient has received HBR. In the opposite case seen in Table III.6b, the difference between the two groups are minor.

When the first observation is not suppressed, the expected spell length for non-HBR patients exceeds that for HBR patients by 3.35 months. In particular, this means that for a HBR patient whose initial state is "not suppressed" we can expect that suppressions will not occur in the subsequent 4 months. Suppressions will be observed for the next 4 months, followed again by a period of 4 month period in which suppression is relevant. This cyclical pattern extents to the end of the sample period. When a HBR patient's first observation is suppressed, then we can expect a spell to last one month longer.

In general, the number of months suppressed and observed, is not worryingly large, especially since the variation of IPLOS data in time is low. To illustration this, we study complete IPLOS observations only. Meaning, that out of the 7,559,867 test data observations, 4,203,157 are not-suppressed IPLOS observations registered on a total of 145,230 patients. Those are used to assess the dynamics of IPLOS data. Information about the underlying dynamics is essential for choosing an adequate imputation technique.

First, is the absolute frequency of IPLOS sequences, where IPLOS sequence means the time interval for which a patient has been continuously registered as a IPLOS patient. If such a time interval exists, then we know that a continuous history of IPLOS observations exists for the patient during her/his existence in the sample. If two of such intervals, which are by definition non-overlapping, exist, then we know that the patient has been registered in IPLOS for two uninterrupted periods. For the time between the two intervals we know that the patient was not registered in IPLOS and therefore no observed IPLOS data exist. If l non-overlapping IPLOS sequences exist for an individual, then we can conclude that we have l consecutive sub-histories of function scores. Since the observation period is finite, these sub-histories will become shorter and shorter as l increases.

The concept of "IPLOS sequence" can be illustrated in the following way. Lets say a patient is registered in IPLOS at time t_r where $0 < t_r < T$. By sample definition, all patients included in the dataset should be registered as an IPLOS observation at least once, but t_r might vary. Let ρ denote number of registrations and δ denote the number of de-registration event occurring in the sample period. Both ρ and δ is $\in \mathbb{Z}$ where is $\rho > 0$ and $\delta \geq 0$. Possible IPLOS sequences are listed in Table III.7 below.

Table III.7: Explanation of IPLOS sequence

Sequence no:	ρ	δ	**Comment:**
1	1	0	Patient was registered at time t_r and continued to be registered until T.
1	1	1	Registered at time t_r the patient was subsequently de-registered prior to T.
2	2	1	Registered at t_r the patient was de-registered, then registered again to stay in IPLOS until T.
2	2	2	Registered at t_r, left IPLOS once, was registered again, left again and was out prior to T.
3	3	2	...
3	3	3	...
$\vdots$	$\vdots$	$\vdots$	...

Using the definitions illustrated in Table III.7, for 88% (127,574) of the patients we observe IPLOS sequence 1. The mean time-span of sequence 1 for patients whom do not die, is 25.7 months. Actually, sequence 1 and 2 combined covers 98% of the patients. This finding indicates that changes in a patient's IPLOS registration status are a relatively rare event. For the intended imputation exercise, this feature could be beneficial.

III.3.3.1 Function scores

As seen above, there is little variation in IPLOS observations when a patient is initially registered at the IPLOS register. Studying the frequency of change in function score, also reveals an interesting pattern. On average, for more than 94% of the patients, the function scores change at most two times throughout the observation period. This measure differs across the different function scores. Considering all function scores, the respective ratio ranges between 91% - 99%. The latter highlights the fact that the velocity of function scores is extremely low. One might even doubt the informational value of these scores with respect to changes a patient's health state.

Based on the above findings, and keeping in mind that each missing "gap" is on average between 4 - 5 month long, the choice of a simple linear nearest-neighbour interpolation routine for the function scores seems reasonable. The time variable is used as the independent variable. In the nearest-neighbour approach, one assigns a known value of the function score - either the last one available before or the first one known after the missing value - to the case that has to be imputed. The value that is closest in time to the record that needs to be imputed is chosen. With ties, the last score known is used. Applying this simple technique to the test data described in Section III.3.2, produces remarkably good results. The interpolated function scores actually match the respective true values one-to-one at each point in time in 96% of all cases. The 4% failure rate can be attributed to cases in which the timing of the imputed change in the function score was off by one period. Table III.8, compares the monthly mean and within standard deviation (SD) of interpolated cases to their true values. As expected the differences are minor.

Table III.8: Comparing imputed function scores with true values (test data)

	Function score								
	1	2	3	4	5	6	7	8	9
True mean	1.551	1.974	1.572	1.896	1.424	2.261	1.655	1.515	1.829
Imputed mean	1.554	1.979	1.575	1.900	1.427	2.265	1.659	1.518	1.833
True within SD	1.676	1.811	1.698	1.726	1.671	1.871	1.699	1.694	1.820
Imputed within SD	1.667	1.803	1.690	1.718	1.662	1.863	1.691	1.686	1.811

(a) Score 1 - 9

	Function score							
	10	11	12	13	14	15	16	17
True mean	1.186	1.404	1.732	1.243	1.190	1.435	1.298	1.316
Imputed mean	1.188	1.407	1.735	1.245	1.193	1.438	1.300	1.319
True within SD	1.647	1.693	1.718	1.639	1.640	1.659	1.637	1.649
Imputed within SD	1.638	1.685	1.710	1.630	1.632	1.651	1.628	1.641

(b) Score 10 - 17

III.3.3.2 IPLOS service data

IPLOS service data, referring for example to the use of either home-service hours or institution days, follow a pattern similar to those of the functions scores discussed in Section III.3.3.1. Month-to-month variation for IPLOS service variables is often simply due to the apparent difference in the number of days per month. Thus intrinsic minor differences have to be expected, for instance, in the context of stays in a long-term institutions.

Using the interpolation routine described in Section III.3.3.1 on IPLOS service days and hours per day - thereafter multiplying imputed values with days per month - provides remarkably good results. Using the same test data, the comparison between the imputed home-service hours and the respective true values reveals exact matches in 82% of all cases. This is 14% lower than the result for function scores. However, the mean mismatch among the non-perfectly matched observations (28% of the cases), is just 16 minutes per month. Performing the same comparison, but allowing the imputed value to match within 2 months difference with the true value, increases the accuracy to 88% and the mean overestimation drops down to 6 minutes.

When interpolated institution days are compared with their true counterparts the relative frequency for an exact match equals 96%. The mean difference is an underestimation of less than a day per month. For the binary variable indicating "other" IPLOS services, the same interpolation routine has been used - just not on a per-day basis. Implementing this strategy lead to an outcome in which 99.9% of all imputations were exact. These high precision test results are also mirrored in Table III.9. Only minor differences surface in the comparison focussing on the monthly mean and the within SD. In fact, for home-service hours, the monthly mean difference between the true values D_T and the imputed values D_C, amounts to less then 2 minutes.

Table III.9: Comparing imputed IPLOS service use with true values

	Home-service hours	Institutions days	Other services
True mean	10.450	3.044	0.329
Imputed mean	10.474	3.064	0.329
True within SD	20.268	4.817	0.267
Imputed within SD	20.164	4.792	0.266

The imputation technique described above is also used for HBR patients, however not during the HBR treatment periods. If the routine is used for periods of HBR provision, one would risk to impute zero hours in some of these months. Such an imputation outcome would not be reasonable, because a HBR patient, *by definition*, has to receive some home-service in the treatment period. Under the imputation technique considered so far, such an unreasonable outcome could materialize if a HBR patient has zero observed hours in the month before or after the treatment period. This problem is solved by using various interpolation rules for different cases. If hours are observed before and after the HBR period, then the rule outlined above is used. However, if one observes zero hours before and non-zero hours after the HBR period, then *backward interpolation* is used. In the opposite case, we rely on *forward interpolation*. One could argue that this strategy would underestimate hours during the HBR period, since the expected increase in treatment intensity during the HBR period. This is a valid argument. Although, a check run on the non-suppressed HBR periods suggests that this is not the case. However, due to the lack of complete HBR period specific observations, there is no good way of testing this potential weakness. For the last case - zero hours before and after the period - *predictive mean matching* (PMM) introduced by Little, 1988, inspired by Rubin, 1986, was applied. Above all, there exist 23 patients receiving HBR from whom no information has been suppressed and zero service hours have been recorded. Again, this should not be possible and is most likely due to incorrect IPLOS registration. These patients have also been imputed by PMM. The variable list used in PMM can be found in Table III.11.

III.3.4 NPR and KUHR data

The favorable pattern seen in IPLOS variables does not hold for NPR and KHUR variables, and therefore different imputation techniques are needed. Before deciding on imputation techniques two factors plays a key role, *missing-data mechanism* and *pattern of missing-data*. Rubin, 1976, developed a typology for describing the missingness mechanism regarded as a probabilistic phenomenon. One of the terms introduced, missing at random (MAR), allows the probabilities of missingness to depend on observed data, but not on missing data (Schafer and Graham, 2002). Such a mechanism is desirable as most imputation techniques assume MAR. In the current case, MAR could be assumed as the missingness is solely due to four observed variables and the time variable, which is also observed.

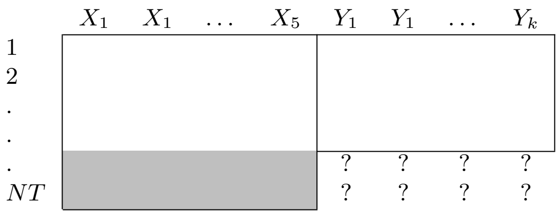

Figure III.2: Illustration of the univariate missing pattern. Variables X_1 - X_5 are all observed and used in the anonymization algorithm. All observations marked as suppressed, due to variables X_1 - X_5, can be structured in the bottom (gray area). Suppression automatically deletes service data Y_1 - Y_k.

Arranging the dataset such that rows correspond to observations and columns correspond to variable, one can illustrate the missing pattern (Schafer and Graham, 2002). Figure III.2 presents such an illustration for the current case and shows a univariate missing pattern. Such a pattern is beneficial, as one can avoid more tedious multivariate imputation techniques. Time dummies are included in the imputations, and that would capture any potential seasonality pattern. In the case of perfect prediction during imputation of categorical data, the augmented regression approach by White et al., 2010 is implemented. For all imputations, seed is set to 1,234. Not suppressed observations from test data is shortened with *No supp.* in tables.

III.3.4.1 NPR indicator variable

A patient can receive specialist care if and only if (s)he is registered as an NPR observation. The imputation of NPR data independent of the binary NPR observation variable, could lead to constellations in which NPR service variables would contain some data even if the indicator equals zero. Such an unreasonable outcome needs to be avoided. Attempts were made to impute some of the NPR service variables irrespective of the restriction that the NPR indicator variable equals one. This solution overestimated the amount of observations, with the overestimation being in the region of half a million observations. The strategy is therefore to first impute the binary NPR indicator variable, and thereafter impute NPR service data conditional on the indicator variable being equal to one. This approach is feasible due to the univariate missing pattern.

For the binary NPR indicator variable, the most intuitive choice of imputation technique is the *logistic regression imputation* method. An alternative method, PMM, heavily depends on standard linear regression assumptions. For instance, one has to assume that the missing variable is continuous. However, as described in the Stata manual (*Stata 15 Base Reference Manual* 2017), PMM uses linear regression to obtain linear predictions. These linear predictions are used as a distance measure to form a set of possible donors (often called *nearest neighbors*) consisting of complete values. PMM then randomly draws a value from these donors into the missing. One has to specify how many nearest neighbors (knn) to include in the donor set. Since all potential donors to the NPR indicator variable only have the value 0 or 1, PMM will not be able to impute other values. Therefore, PMM should produce logical estimates.

PMM is generally a cumbersome procedure, and can therefore be time consuming. As an example, we tried to run PMM for the NPR indicator variable on the complete test dataset (Section III.3.2). After running for eight days, the procedure did not produce results. The indicator variable had therefore to be tested on a smaller dataset. Such a smaller test dataset is created just as described in Section III.3.2, however one randomly draws 20,000 patients and their associated observations. The smaller test dataset includes 1,022,145 observations. As before, 50% are randomly deleted to mimic suppress observations.

Table III.10: Test results for NPR indicator variable imputations

	All			Equal 1		
	Mean	SD	Within SD	NT	N	T-bar
True value, small	0.2640	0.4408	0.3823	269,819	18,800	14.35
Logistic regression	0.2635	0.4406	0.4190	269,377	19,629	13.72
PMM, knn=1	0.2630	0.4403	0.4184	268,860	19,626	13.70
PMM, knn=5	0.2638	0.4407	0.4190	269,644	19,633	13.73
True value	0.2643	0.4410	0.3831	1,998,251	139,081	14.37
Imputed, logistic	0.2643	0.4410	0.4195	1,998,221	144,928	13.79

Covariates included: HBR days, gender, age group, marital status, function scores, IPLOS services, education level and time dummies
knn: number of nearest neighbors

Table III.10 compares imputed values from different imputation techniques with the true value on key statistical measures. The PMM version is presented with two different nearest neighbor options (knn), 1 or 5. All three alternatives produce approximately identical mean estimates which in turn are virtually identical to mean of the actual values. There is a slight variation on the fourth decimal level. Findings on the mean can directly be translated into estimated standard deviation, where all three imputation alternatives produces estimates of the true value that only varies on the fourth decimal. All imputation techniques overestimated the estimated within standard deviation, which is also reflected in the imputed number of patients (N) receiving NPR services. The sum of NPR observations (NT) is fairly close to the true value. Estimation of T-bar (N/NT), or the average number of times a patient visits specialist care, is underestimated. The number of patients receiving NPR services is overestimated by more than 4%. The same tendency to overestimate holds for the within standard deviation. This overestimation is observed for all three alternatives. Therefore, capturing the within patient changes seems to be the main weakness of the techniques considered. All techniques perform equally well. Since the logistic-regression imputation technique is less costly from a numerical/ computational point of view, it became our method of choice in the context at hand.

The second last row in Table III.10 presents the true NPR observation indicator variable for the large test dataset. Test results for the imputation by logistic regression on the large test data are presented in the last row of Table III.10. The imputed number of observations, mean and standard deviation are all remarkably close to the actual values. The within standard deviation as well as the number of patients are overestimated. This outcome is inline with the previous presentation. The NPR variables cannot take any value unless the indicator variable is equal to one. Thus, all NPR service variables are set to zero whenever the imputed predictor variable is equal to zero.

III.3.4.2 NPR variables

As mentioned previously, all NPR imputation are conditioned on the indicator variable. All NPR service variables are ordered by definition, and their values are positive integers. Therefore, ordered logistic regression imputation method is used for these variables. Poisson regression has also been tested, but ordered logistic regression outperformed the latter on all measures. For the NPR sector variable, a categorical variable, multinomial logistic regression imputation is used. For all variables, M imputations (complete datasets) are generated under the same imputation model. The imputations are indexed by $m = 1, 2, ..., M$. In the remainder, M will be typically set to 5, and test results are presented for $m = 1, 3, 5$.

The outcome of the imputation experiment is documented in Table III.11. Apparently the previously observed pattern in Section III.3.4.1 is also visible here for all other NPR variables. The imputation techniques produce satisfactory estimates for mean, standard deviation, minimum/maximum and number of observations. For some variables the techniques produce a maximum value which is lower than the true value. The reason for this is that the maximum values have been randomly deleted when creating test data. All imputed maximum values are equal to the maximum of the non-suppressed observations in our test data. This mismatch will not have any affect when applying the imputation techniques to the original dataset. The maximum values will then be included and one does not expect that any of the suppressed observations will exceed these values. The same drawbacks discussed earlier also occur. We detect overestimation for within standard deviation and number of patients receiving service.

Table III.11: Imputation test result for NPR variables

	Mean	SD	Within SD	NT	N	T-bar	Min.	Max.
NPR number of episodes[a]								
True value	2.1075	2.2977	1.8026	1,998,251	139,081	14.37	1	130
No supp.	2.1089	2.2973	1.7583				1	122
Imputed m=1	2.1124	2.2942	2.0566	1,998,211	144,928	13.78	1	122
Imputed m=3	2.1126	2.2984	2.0619	1,998,211	144,928	13.78	1	122
Imputed m=5	2.1129	2.2967	2.0594	1,998,211	144,928	13.78	1	122
NPR number of episodes - only somatic[b]								
True value	1.8575	2.0572	1.6218	1,362,250	132,825	10.26	1	57
No supp.	1.8577	2.0571	1.5693				1	46
Imputed m=1	1.8999	2.0859	1.8588	1,385,601	144,065	9.62	1	46
Imputed m=3	1.9002	2.0831	1.8571	1,385,102	144,039	9.62	1	46
Imputed m=5	1.9013	2.0882	1.8907	1,385,614	144,039	9.62	1	46
NPR number of policlinic visits[a]								
True value	2.0949	2.3372	1.8124	1,709,573	132,496	12.90	1	130
No supp.	2.0954	2.3357	1.7626				1	122
Imputed m=1	2.0953	2.3254	2.0694	1,710,057	142,166	12.03	1	122
Imputed m=3	2.0965	2.3264	2.0703	1,710,251	142,180	12.03	1	122
Imputed m=5	2.0955	2.3249	2.0703	1,710,214	142,194	12.03	1	122
NPR number of policlinic visits - only somatic[b]								
True value	1.8566	2.1690	1.6838	1,097,045	123,141	8.91	1	57
No supp.	1.8561	2.1686	1.6225				1	46
Imputed m=1	1.9030	2.2156	1.9457	1,115,804	139,964	7.97	1	46
Imputed m=3	1.9047	2.2137	1.9429	1,115,394	139,952	7.97	1	46
Imputed m=5	1.8979	2.1992	1.9307	1,116,883	139,941	7.98	1	46
NPR sum days length of stay[a]								
True value	8.8044	8.9174	6.3364	497,518	111,317	4.47	1	31
No supp.	8.7783	8.9079	5.8192				1	31
Imputed m=1	8.7661	8.9177	7.1096	496,250	131,943	3.76	1	31
Imputed m=3	8.7511	8.9059	7.1060	497,498	131,866	3.77	1	31
Imputed m=5	8.7502	8.9020	7.1070	497,793	132,020	3.77	1	31
NPR sum days length of stay - only somatic[b]								
True value	6.4748	6.4989	5.1547	403,610	105,657	3.82	1	31
No supp.	6.4639	6.5011	4.6571				1	31
Imputed m=1	6.3320	6.2410	5.0385	408,253	126,632	3.22	1	31
Imputed m=3	6.3426	6.2546	5.0519	409,518	126,561	3.24	1	31
Imputed m=5	6.3260	6.2342	5.0298	410,259	126,798	3.24	1	31

Continued on next page

Table III.11 – *Imputation test result for NPR variables - Continued*

	Mean	SD	Within SD	NT	N	T-bar	Min.	Max.
NPR main diagnosis group[c]								
True value	8.5922	6.5219	5.2224	1,361,819	132,819	10.25	1	27
No supp.	8.6000	6.5238	4.9872				1	27
Imputed m=1	8.5576	6.5423	5.8846	1,355,441	144,029	9.41	1	27
Imputed m=3	8.5569	6.5456	5.8907	1,356,315	144,008	9.42	1	27
Imputed m=5	8.5540	6.5374	5.8811	1,356,326	144,002	9.42	1	27
NPR sector[d]								
True value	1.3865	0.7881	0.6340	1,691,110	136,208	12.42	1	4
No supp.	1.3856	0.7873	0.6091				1	4
Imputed m=1	1.4028	0.8019	0.7271	1,698,309	144,679	11.74	1	4
Imputed m=3	1.4029	0.8024	0.7277	1,698,736	144,673	11.74	1	4
Imputed m=5	1.4028	0.8020	0.7276	1,698,845	144,696	11.74	1	4

[a] Ordered logistic regression imputation method
[b] Same as [a], also include imputed values for total amount of service type as factor variable
[c] Ordered logistic regression imputation method, poisson also tested and m-logit not feasible
[d] Multinomial logistic regression (m-logit), with NPR main diagnosis ad factor variable
Covariates included: HBR days, age group, material status, function scores, IPLOS services, gender, education level and time dummies

Studying the m different imputations in Table III.11, one finds marginal differences in estimated mean, standard deviation and within standard deviation. When working with multiple imputed data, *Rubin's combination rule* is used to combine M results to one set of results. According to this rule, the total point estimate is the average of the M complete datasets. However, the sampling variance associated with the combined parameter estimates is not that simple. The sampling variance is estimated as a combination of within-imputation variability and between-imputation dataset variability. To be specific let $U^{(m)}$ denote the point variance from the mth ($m = 1, 2, ..., M$) dataset[8]. The within-imputation variance component, W is the average of complete-data variance estimates,

$$W = \frac{1}{M} \sum_{m=1}^{M} U^{(m)}.$$

The between-imputation variance, B, is produced in the following way,

$$B = \frac{1}{M-1} \sum_{m=1}^{M} (Q^{(m)} - \bar{Q})^2,$$

where $Q^{(m)}$ is the mth specific point estimate and $\bar{Q}$ the average point estimate over M complete datasets. According to Rubin's combination rule, the total variance, TV, is given as

$$TV = W + \left(1 + \frac{1}{M}\right) B. \tag{III.1}$$

[8]The exposition relies on the notation established in Carlin et al., 2003.

Thus the total variance takes the form of a weighted sum of the within-imputation variance (W) and the between-imputation variance (B), where the weight on B decreases in the number of complete datasets available. Therefore, in the standard approach one would generate as many complete datasets as possible to keep the effect of B on the total variance as small as possible. In our case, we have seen that there is hardly any variation in the imputed means and standard deviations (cf. Table III.11). Thus the within-variation B is already close to zero. Considering this argument and recalling the extraordinary numerical/computational imputation costs in our case, one could hardly justify the generation of more than $M = 1$ datasets.

Reducing the number of imputations to only one also comes with two extra benefits. Firstly, it greatly reduces the need for computation power. The second benefit comes from studying Table III.11 a bit more in detailed. As can been seen, the outcomes of the imputation experiments for all somatic specific service types appear to be problematic. Number of observations, mean and standard deviation are all overestimated for number of somatic episodes and policlinic visit. Logically, one cannot have any somatic service, like length of stay (LOS) and policlinic visits, if total amounts of these services are zero. Meaning for example, that number of somatic episodes is dependent on the total number of episodes.

To account for this relationship when using multiple imputation in Stata, one has to impute the two variables in sequence by univariate conditional imputation methods. However, Stata will first impute the total service type variable and then include the imputed values of total service as a variable when imputing somatic services. The results in Table III.11 are based on such a strategy. If $M = 1$, the somatic variables can be imputed conditioned on total service larger than zero and somatic services has to be equal to zero if total services are equal to zero. Table III.12 presents result for the latter procedure, which clearly yields better results for the somatic specific variables, without influencing the total service.

Table III.12: Test result for somatic NPR variables

	Mean	SD	Within SD	NT	N	T-bar
NPR number of episodes - only somatic[a]						
True value	1.8575	2.0572	1.6218	1,362,250	132,825	10.26
No supp.	1.8577	2.0571	1.5693			
Imputed conditioned	1.8545	2.0607	1.8264	1,358,846	143,906	9.44
NPR number of policlinic visits - only somatic[a]						
True value	1.8566	2.1690	1.6838	1,097,045	123,141	8.91
No supp.	1.8561	2.1686	1.6225			
Imputed conditioned	1.8563	2.1739	1.8955	1,094,798	139,748	7.83
NPR sum days length of stay - only somatic[a]						
True value	6.4748	6.4989	5.1547	403,610	105,657	3.82
No supp.	6.4639	6.5011	4.6571			
Imputed conditioned	6.3842	6.5001	5.2257	400,605	124,880	3.21

[a] Ordered logistic regression imputation method conditioned that total equivalent variable larger then zero

Covariates included: HBR days, age group, material status, function scores, IPLOS services, gender, education level and time dummies

Singling out the amount of imputed observations for somatic policlinic visits as an example, we can see in Table III.12 an underestimation of 2,247 observations compared with the true value. The result presented in Table III.11 show an overestimation in the region of 18,349 - 19,838 observations compared to the true value. A clear improvement, which is also seen in presented key statistics for all somatic NPR variables.

Diagnosis-related group (DRG), is a system to classify hospital cases into one out of 467 groups. Each of these groups comes with a payment weight based on the average resources used to treat patients in that DRG. The monthly sum of these DRG points per individual is one variable included the dataset. Sum DRG points is a continuous variable with a restricted range, in particular with 0.01 points as a lower limit. Using standard regression imputation will not be appropriate because it could produce negative values. Such outcomes were frequently observed in tests of the method. Two alternative likelihood based imputation techniques which in principle should solve our problen, interval or truncated regression, failed to converge in testing. This may be due to the fact that the lower limit of 0.01 features dominantly in the sample. PMM may therefore be considered as another viable option. As mentioned above, the implementation of the PMM method is extremely costly on large datasets. We therefore establish its versatility on small test datasets.

Table III.13: Test result for NPR DRG variable

	Mean	SD	Within SD	NT	N	T-bar	Min.	Max.
True value	0.5577	1.1655	1.0442	1,354,707	132,812	10.20	0.01	46.47
True value, small	0.5622	1.1571	1.0362	182,772	18,016	10.15	0.01	39.23
No supp.	0.5677	1.1846	1.0110				0.01	39.23
PMM, knn(1)	0.5614	1.1632	1.0754	181,642	19,474	9.33	0.01	39.23
PMM, knn(5)	0.5634	1.1718	1.0813	181,590	19,477	9.32	0.01	39.23
PMM, knn(5) bootstrap	0.5656	1.1748	1.0879	181,616	19,473	9.33	0.01	39.23

Covariates included: HBR days, gender, age group, marital status, function scores, IPLOS services, education level, time dummies, imputed NPR main diagnosis and NPR sector
In addition, conditioned on NPR main diagnosis>0 and NPR somatic episode>0
knn: number of nearest neighbors

The first row in Table III.13 presents the true value in the large test dataset, whereas the second row presents the true value in the reduced test dataset. There are minor differences between the true value of two test datasets. The maximum value in the complete test dataset is substantially larger compared to the small dataset. Since patients included in the smaller dataset where randomly selected, such differences can occur. This will not be an issue when imputing on the complete dataset. HBR patients are not expected to be the patient group that will need more DRG points than the maximum included. As in Section III.3.4.1, we test several versions of PMM. All of them are conditioned on the fact that NPR main diagnoses is not zero. In the DRG system, that one needs a specified diagnosis group to get DRG points. There are hardly any differences between the different PMM versions, and $knn = 5$ is chosen.

For testing imputed sum DRG points, 5 complete datasets were generated where m indicates a specific imputed dataset. The results are similar to those shown in Table III.10. The PMM imputation strategy works well on the mean, standard deviation and number of observations. However, the same mismatch on number of patients and within standard deviation is found. These findings are robust to various changes in the imputation routine as, for instance, reduction of the neighborhood, use of bootstrapping to specify that posterior estimates of model parameters are obtained, using sampling with replacement.

III.3.4.3 Minor adjustments in complete data imputation

All imputations routines that proved to be successful in the imputation experiments have been used in the final imputation exercise. The final amount of suppressed observations needed to be imputed is 150,584 since the extreme cases have been deleted. In addition to the above techniques, one minor adjustment will be applied when implementing the imputation strategy. The ordered logistic regression imputation, in some cases, produces more somatic service then given in total, which is not possible.

When imputing the final dataset, 5,429 observations had more somatic policlinic visits then total visit, and 2,432 observation had more somatic length of stay then total. In these cases, a random mechanism decided whether to use the somatic number or the total in these observations. The randomization was adjusted for the mean proportion of somatic service by age group. For example: In age group 1, for the non-suppressed, somatic policlinic visit was on average 48% of the total amount of policlinic visits in that age group. In the randomization, a probability of 0.48 was given to the somatic policlinic imputed number. This was done for all 15 different age groups. This should have minor influence on the completed result due to the small number of observations. Also, as age increases, the proportion of somatic services tends to increase, meaning that the there is an increased probability for choosing the highest value. On could therefore argue that we overestimate NPR services. This has to be kept in mind when analyzing the data.

III.3.4.4 KUHR indicator variable

As for the NPR indicator variable presented in Section III.3.4.1, one should first impute the binary KUHR indicator variable. Thereafter, all imputations for KUHR service variables have to be conditioned on the indicator variable. The same potential imputation methods considered in Section III.3.4.1 are also tested in the context of the KUHR indicator variable. We present the results in Table III.14.

Table III.14: Test result for KUHR indicator variable imputations

	All			Equal 1		
	Mean	SD	Within SD	NT	N	T-bar
True value	0.4099	0.4918	0.3705	418,996	15,465	27.09
Logistic regression	0.4104	0.4919	0.4276	419,474	19,254	21.78
PMM, knn=1	0.4100	0.4918	0.4268	419,061	19,241	21.78
PMM, knn=5	0.4096	0.4918	0.4271	418,641	19,323	21.67

Covariates included: HBR days, gender, age group, marital status, function scores, IPLOS services, education level and time dummies
knn: number of nearest neighbors

The findings exhibited in Table III.14 match those shown in Table III.10 closely. All the different techniques produce approximately the same results and have the same weaknesses. Within standard deviation and number of patients are overestimated. Mean, standard deviation and number of observations are more or less spot on compared to the true value. Imputation by logistic regression is chosen since this technique is less computer power sensitive than the PMM routine.

III.3.4.5 KUHR variables

In table III.15, the test results for all imputed KUHR service variables are presented. Also KUHR service variables have a naturally ordered structure, and their values can only be positive integers. Therefore, all variables have been imputed with ordered logistic regression conditioned on the KUHR indicator variable taking the value of one. Imputing KUHR data is more straight forward compared to data from NPR. The complication of an extra relationship between the variables as in the NPR case, does not exist in the KUHR context. For all variables M whas set to 5 and $m = 1, 3, 5$ are presented.

Our findings outlined in Section III.3.4.2 are similar to the results obtained in the KUHR context. The test results show that after imputation, means, standard deviations and number of observations are all close to the true value. Number of patients and within standard deviation for the imputed values does not match the true value very well. Especially number of patients seems to be constantly overestimated. For physiotherapy services the overestimation on number of patients receiving these service are particular severe. This has to be kept in mind when using this variable. In Table III.15, one can see that there is hardly any variation between imputed means and standard deviations between the different datasets. Also here, we argue that the extra benefit in terms of total variance of imputing $M > 1$ datasets will be marginal (cf. Rubin's combination rule (III.1). The between-imputation variance, B, will be fairly small.

Table III.15: Imputation test results for KUHR variables

	Mean	SD	Within SD	NT	N	T-bar	Min.	Max.
GP consultation								
True value	1.4158	0.8758	0.7630	1,918,345	107,448	17.85	1	25
No supp.	1.4155	0.8772	0.7482				1	25
Imputed m=1	1.4157	0.8783	0.8163	1,915,824	135,094	14.18	1	25
Imputed m=3	1.4158	0.8779	0.8164	1,914,759	135,070	14.18	1	25
Imputed m=5	1.4164	0.8784	0.8162	1,916,435	135,091	14.19	1	25
GP home visit								
True value	1.2347	0.6348	0.5023	113,310	46,646	2.66	1	21
No supp.	1.2357	0.6315	0.4608				1	15
Imputed m=1	1.2384	0.6340	0.4606	113,204	54,991	2.06	1	15
Imputed m=3	1.2383	0.6344	0.4591	114,002	55,409	2.06	1	15
Imputed m=5	1.2402	0.6391	0.4641	113,341	54,998	2.06	1	15
GP contact by phone or mail								
True value	1.7827	1.4247	1.2272	1,692,138	111,286	15.21	1	58
No supp.	1.7833	1.4254	1.1976				1	58
Imputed m=1	1.7864	1.4286	1.3142	1,692,716	137,450	12.32	1	58
Imputed m=3	1.7871	1.4264	1.3131	1,692,680	137,389	12.32	1	58
Imputed m=5	1.7872	1.4334	1.3192	1,692,844	137,466	12.31	1	58

Continued on next page

Table III.15 – *Imputation test result for KUHR variables - Continued*

	Mean	SD	Within SD	NT	N	T-bar	Min.	Max.
GP office simple contact								
True value	1.2812	1.0523	0.6902	286,641	68,489	4.19	1	27
No supp.	1.2804	1.0531	0.6542				1	27
Imputed m=1	1.2818	1.0591	0.7618	286,856	96,200	2.98	1	27
Imputed m=3	1.2805	1.0482	0.7565	286,757	95,985	2.99	1	27
Imputed m=5	1.2817	1.0579	0.7611	286,957	96,260	2.98	1	27
GP multidisciplinary cooperation meeting								
True value	1.1137	0.4117	0.3335	69,127	21,130	3.27	1	14
No supp.	1.1126	0.4067	0.3084				1	11
Imputed m=1	1.1143	0.4133	0.2925	69,196	34,708	1.99	1	11
Imputed m=3	1.1137	0.4044	0.2823	69,471	34,870	1.99	1	11
Imputed m=5	1.1114	0.3993	0.2825	69,448	35,071	1.98	1	11
GP other services								
True value	1.1412	0.4597	0.3820	278,833	77,171	3.61	1	18
No supp.	1.1421	0.4598	0.3569				1	13
Imputed m=1	1.1421	0.4579	0.3608	278,607	100,284	2.78	1	13
Imputed m=3	1.1435	0.4619	0.3664	278,380	100,295	2.78	1	13
Imputed m=5	1.1425	0.4631	0.3653	278,863	100,404	2.78	1	13
Physiotherapy consultation								
True value	1.2786	0.7502	0.5934	4,832	1,400	3.45	1	13
No supp.	1.2815	0.7812	0.5826				1	13
Imputed m=1	1.2815	0.7736	0.4621	4,838	3,109	1.56	1	13
Imputed m=3	1.2851	0.8411	0.4801	4,794	3,101	1.55	1	13
Imputed m=5	1.2754	0.7801	0.4476	4,902	3,182	1.54	1	13
Physiotherapy home visit								
True value	4.1350	2.6133	1.7821	54,803	4,911	11.16	1	26
No supp.	4.1590	2.6192	1.7291				1	21
Imputed m=1	4.1938	2.6434	1.7192	54,176	21,261	2.55	1	21
Imputed m=3	4.1990	2.6285	1.7054	54,364	21,389	2.54	1	21
Imputed m=5	4.1934	2.6515	1.7247	54,389	21,252	2.56	1	21
Physiotherapy other services								
True value	4.2725	2.9132	2.1932	534,175	35,549	15.03	1	45
No supp.	4.2762	2.9124	2.1289				1	43
Imputed m=1	4.2834	2.9199	2.3898	532,583	99,836	5.33	1	43
Imputed m=3	4.2782	2.9171	2.3913	533,872	100,243	5.33	1	43
Imputed m=5	4.2825	2.9270	2.3976	533,163	99,981	5.33	1	43

Ordered logistic regression imputation method used on all variables

Covariates included: HBR days, age group, material status, function scores, IPLOS services, gender, education level and time dummies

III.4 Data summary

Table III.16 presents summary statistics for some variables after the above imputation strategy has been implemented for HBR and non-HBR patients seperately. The number of patients and observations are the same as presented in Table III.5, but the mean number of observations, $\bar{T}$, is also presented. None of the HBR patients dies during the observation period and therefore each HBR patient is observed 60 months. The non-HBR patients also contain patients living in nursing-homes at the start of observation period. Many of those would pass away during the 5 year data period, resulting in less observations.

Table III.16: Summary statistics for some variables

	HBR		$\neg$ HBR	
	Mean	SD within	Mean	SD within
Male	0.19	-	0.40	-
Age group	9.85	-	7.20	-
East - Hospital area	0.23	0.04	0.47	0.09
North and Middle - Hospital area	0.04	0.02	0.14	0.03
West - Hospital area	0.28	0.03	0.23	0.06
South - Hospital area	0.40	0.06	0.08	0.05
Vestfold - Hospital area	0.05	0.03	0.04	0.03
Education	3.25	0.32	3.27	0.42
IPLOS hours	5.87	9.69	10.47	20.27
IPLOS institution days	0.35	2.59	2.96	4.75
NPR policlinic visit	0.43	1.22	0.48	1.15
NPR length of stay	0.43	2.63	0.58	2.82
GP consultation	0.39	0.74	0.36	0.63
GP office visits	0.06	0.33	0.05	0.28
N	1,741		153,098	
NT	104,460		7,847,222	
$\bar{T}$	60.00		51.23	

There are some clear differences between patients receiving HBR and other patients. Those differences need to be taken into account when analysing the data. For instance, there are fewer males and the average age is higher among HBR patients. The HBR lower age limit is usually set to 18 years, however the typical HBR patients are elderly people. The proportion of patients living in different hospital areas differs between the two groups. As seen in Table III.16, approximately 47% of non-HBR patients live in hospital area (HA) east. Which makes sense since this area contains three major cities in Norway, one of them being the capital, Oslo. To explain the residence-of-living proportion for HBR patients differs from the rest, we have to turn to the history of HBR implementation in Norway. HBR is not mandatory, so municipalities implemented the intervention at different rates. The municipality of Arendal, a part of HA south, was one of the first cities in Norway to fully implement HBR after running an pilot. The opposite is true for municipalities in HA east, leading to the clear differentiation between proportion of residence. IPLOS service data differs between the two groups, which is expected since all 153,098 non-HBR patients are a natural comparator to HBR. However, average monthly KUHR and NPR service are surprisingly equal. It is beyond the scope of this paper to elaborate on the latter observation, but it could tell something about ”praxis” more than about a ”need for service”.

III.5 Discussion and concluding remarks

As pointed out above, Norway has high quality register data that can be used to create unique micro-level datasets. The current paper shows how data from three national health registers can be merged to create a dataset suitable for answering important cost-effect research questions concerning HBR. However, the process of merging these registers is far from flawless, causing the need for both adaptability and patience during the process. The key dilemma lies in finding the right balance between, legal necessity to guarantee the private sphere of individuals versus the need to support the policy maker with information as precise as possible. In the case at hand, we had to accept suppressed observations to meet legal requirements with respect to anonymity of patient data, which lead to the need for an imputation strategy to achieve (restore) statistical precision.

The devised strategy proved to work remarkably well for IPLOS related data. Because IPLOS data only have minor month to month variation and few sequences, imputation via simple interpolation proved to be effective. When tested, the technique reconstructed the true value, month by month per individual, almost perfectly for certain variables and the relative frequency of perfect matches never fell below 0.88. As a consequence, the monthly mean service use and within standard use between test data and the true data hardly varied.

NPR and KUHR data, however, differ from IPLOS data and more classical imputation techniques where therefore needed. In the test results presented, all imputation techniques provide good estimates of monthly mean and standard deviation. Those are of course key measures when working with pooled regression based estimators. However, standard panel regression models like fixed-effect (FE) use within-standard deviation. Unfortunately, all imputation techniques overestimated the within standard deviation. This could bias the FE estimates and increase coefficient variance, which again can influence inference. However, in percentage, the amount of observations needed to be imputed are low and within standard deviation for the test data is still reasonable. The latter arguments should help to minimize potential bias. As a precaution, a series of robustness test should be implemented when analyzing the imputed data, so inference can be trustworthy.

All test also showed an another drawback. They failed to replicate the mean frequency of service usage. For example, as seen in Table III.11, the average number of NPR policlinic visit occurrence per patient, is 13 over the 5 year period, whereas the imputed value is 12. This is still close and impressive taking the complexity of panel data structure into account. However, this will still be a weakness if one is particularly interested in duration analysis on discrete time. The development of imputation methods which also takes this aspect into account could be a subject of future research.

As presented in the introduction, statistical estimators are typically know for the asymptotic case. For all practical purpose, meaning large sample cases. In many countries, important economic policy-maker questions might be difficult to answer due to lack of data or costly data collection. Norway is in a fairly unique situation as most health services provided is public, and mandatory health registers are in place. The infrastructure for collecting key health data for policy research therefore already exists. However, this paper substantiates the claim, that the process in by which the researcher obtains access to the data is cumbersome and close to impossible. We do understand the need for legal protection of individual information. Measures taken to achieve protection should be carefully and sensitively balanced against the need of society to learn about itself.

The key issue is that none of the application process are identical and the health registers have different ethical requirements. All applications for data and ethical requirements seems not to be designed for economic research. Policy questions that usually interest economists focus on the mean overall effect, and not at the individual level. Economist use individual historical data, but conduct analysis on the aggregated level and actually never meet the patient. Of course, one could randomly run into patients, but the research has to deliberately search in the data to find this patient. Which would in our opinion violate a researcher's ethical standard, and should be avoided regardless of official requirements. This "far-fetched" potential relationship to patients is today's standard when defining data anonymity, and some nuances need should be discussed.

These different ethical authorities are in our opining, originally not designed for applying register data. They typical focus more on classical medical studies where one needs a patient's consent. The irony in the current case, is that in all ethical applications send, one specifically emphasised that the data would be anonymous. Besides the latter argument, save storage was always questioned and therefore the final dataset is stored on a research server. Meaning, the current project ended up with approximately the same approvals and storing as identifiable clinical data, even though the dataset is anonymized.

With the above restrictions in mind, there is a chance that important research questions will stay unanswered as the time-consuming process could work as a hinderance. Even if all applications are in place, the process in which data are delivered to the client/researcher seems to lack efficiency. The fact that it takes 11 months of waiting time for a 24-hour job, shows that not all systems are ideal. A prospective researcher may benefit from our documentation in two ways: (s)he may become aware of the problem dimension and take it into account when contemplating a planning horizon for her (his) research. Hopefully, data owning institutions might consider the report as feedback to their work with a client. Norway has the potential to provide great contributions in policy-research due the many rich data registers. However, it all depends on the goals of Norwegian policy-makers. If high quality policy-research is wanted, new rules and boundaries need to be created in the near future.

III.6 References

Bai, Jushan (2009). "Panel Data Models with Interactive Fixed Effects". In: *Econometrica* 77.4, pp. 1229–1279.

Baltagi, Badi H (2013). *Econometric analysis of panel data.* eng. 5th ed. Chichester: Wiley.

Baltagi, Badi H., Chihwa Kao, and Long Liu (2008). "Asymptotic properties of estimators for the linear panel regression model with random individual effects and serially correlated errors: the case of stationary and non-stationary regressors and residuals". In: *The Econometrics Journal* 11.3, pp. 554–572.

Bersvendsen, Tore, Jochen Jungeilges, and Eirik Abildsnes (2019). "Empirical evaluation of home-based reablement: A systematic review". Manuscript submitted for publication.

Buuren, S. van, H. C. Boshuizen, and D. L. Knook (1999). "Multiple imputation of missing blood pressure covariates in survival analysis". In: *Statistics in Medicine* 18.6, pp. 681–694.

Cameron, Adrian and Pravin Trivedi (Jan. 2005). *Microeconometrics: Methods and Applications.* Cambridge University Press.

Carlin, J. B. et al. (2003). "Tools for analyzing multiple imputed datasets". In: *Stata Journal* 3.3, 226–244(19).

Chudik, Alexander and M. Hashem Pesaran (2015). "Common correlated effects estimation of heterogeneous dynamic panel data models with weakly exogenous regressors". In: *Journal of Econometrics* 188.2. Heterogeneity in Panel Data and in Nonparametric Analysis in honor of Professor Cheng Hsiao, pp. 393 –420.

Green, William H. (2003). *Econometric analysis.* 5th. Prentice Hall. Chap. 5 and Appendix E.

Kleinke, Kristian et al. (2011). "Efficient ways to impute incomplete panel data". In: *AStA Advances in Statistical Analysis* 95.4, pp. 351–373.

Little, Roderick J. A. (1988). "Missing-Data Adjustments in Large Surveys". In: *Journal of Business & Economic Statistics* 6.3, pp. 287–296.

Morris, T. P., P. Royston, and I. R. White (2014). "Tuning multiple imputation by predictive mean matching and local residual draws". eng. In: *BMC Medical Research Methodology.*

Pesaran, M. Hashem (2006). "Estimation and Inference in Large Heterogeneous Panels with a Multifactor Error Structure". In: *Econometrica* 74.4, pp. 967–1012.

Raghunathan, Trivellore et al. (2001). "A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models". In: *Survey Methodology* 27.1, pp. 85–95.

Rubin, Donald B. (1976). "Inference and Missing Data". In: *Biometrika* 63.3, pp. 581–592.

— (1986). "Statistical Matching Using File Concatenation with Adjusted Weights and Multiple Imputations". In: *Journal of Business & Economic Statistics* 4.1, pp. 87–94.

Schafer, Joseph and John Graham (2002). "Missing Data: Our View of the State of the Art". In: *Psychological Methods* 7, pp. 147–177.

Stata 15 Base Reference Manual (2017). StataCorp. College Station: Stata Press.
White, Ian R., Rhian Daniel, and Patrick Royston (2010). "Avoiding bias due to perfect prediction in multiple imputation of incomplete categorical variables". In: *Computational Statistics & Data Analysis* 54.10, pp. 2267 –2275.

Paper IV

The health cost effects of home-based reablement: Empirical evidence from Norway

The health cost effects of home-based reablement: Empirical evidence from Norway

Tore Bersvendsen[1]

[1]Department of Economics and Finance, University of Agder, tore.bersvendsen@uia.no

The coming demographical challenge has encouraged policy-makers to look into new ways of providing health care. A new somatic rehabilitation intervention, home-based reablement (HBR), aims at restoring or increasing patients' level of functioning. This study provides evidence on whether HBR reduces health service costs. A unique individual-level 5-year monthly panel for the period 2011 - 2015 is used. Utilizing a fairly new regression estimator for dynamic models, I find that HBR can be expected to reduce health service costs. The estimated effect varies between -4,875 and -6,007 NOK per month in the short run and -4,637 and -6,373 NOK per month in the long run. The strongest effect is found for specialist health services. The provider of those services differs from the stakeholder who provides HBR. All effects can be attributed to females, as results conditioned on gender show no significant effect for males.

Keywords: Rehabilitation, health costs, health intervention, dynamic panel, gender differences

JEL classification: I19, J14, C33

Acknowledgment: I would thank Jochen Jungeilges and Eirik Abildsnes for valuable comments. This work was performed on the TSD (Tjeneste for Sensitive Data) facilities, owned by the University of Oslo, operated and developed by the TSD service group at the University of Oslo, IT-Department (USIT). (tsd-drift@usit.uio.no)

Disclaimer: Data from the Norwegian Patient Register has been used in this publication. The interpretation and reporting of these data are the sole responsibility of the authors, and no endorsement by the Norwegian Patient Register is intended nor should be inferred.

IV.1 Introduction

The Western world is facing significant demographic change, and Norway is no exception. For the last decade, the Norwegian population growth rate has exceeded that of other European countries, a trend that will continue in the near future (Leknes et al., 2018). Norwegian population projections indicate that by 2060, every fifth citizen will be 70 years of age or older. As the population of older adults grows, the number of individuals facing age-related diseases, along with the number of diseases occurring simultaneously within the same person, i.e., multimorbidity, will increase (Barnett et al., 2012). In a recent paper by Eckardt et al., 2017, it is shown that health care costs tend to increase with the number of comorbidities. The estimated cost of long-term care to people over 65 years of age will double or triple by 2050 in countries in the Organisation for Economic Co-operation and Development (Oliveira Martins and Maisonneuve, 2007). As stated by Zweifel et al., 2009, p. 460, the aging population jeopardizes the financial equilibrium of social health insurance.

The increasing demand for services leads to continuous pressure on many traditional homecare providers (Ryburn et al., 2009). The coming challenges will increase the demand for all long-term services, such as nursing homes and home-based care, and traditionally, home-based care is more cost-effective. Since the expected cost of long-term care will increase dramatically, many high-income countries are actively encouraging a shift from residential care to home-based care as a potentially more effective and financially sustainable approach to the challenge (Cochrane et al., 2016). Another incentive is that older people prefer to 'age in place' (Wiles et al., 2011). The forthcoming challenges will compel healthcare services to focus more on preventive measures, early intervention, the increased use of technology, rehabilitation and healthcare services that are less manpower-intensive, and services that empower senior citizens to self-manage chronic diseases.

In this paper, I examine the individual cost effect of one new homecare service, *home-based reablement* (HBR), compared to usual care. The aim of this rehabilitation approach is to restore or increase patients' level of functioning, thereby increasing patients' self-reliance and consequently decreasing their dependence on healthcare services. Typically, the intervention is developed as a response to the forthcoming challenges mentioned above. HBR is not a standardized treatment, and the content of HBR may vary; however, all interventions have the intention of restoring or increasing an individual's level of functioning (Tuntland et al., 2014; Whitehead et al., 2014). HBR has gained significant international prominence in recent years (Cochrane et al., 2013). The intervention is of special interest because the main features, such as being time-limited, multidisciplinary, home-based, goal-oriented and person-centered, are all features many believe future health care should have.

Existing empirical evidence on the effects of HBR is typically published in the epidemiological literature and was found to be inconclusive in a recent review (Bersvendsen et al., 2019). Increased physical functioning, mainly focusing on activities of daily living (ADL), is the focus of most studies on HBR. A study by Lewin and Vandermeulen, 2010 is the first study to use functional gain as the primary outcome, and it produced some promising results. The HBR group scored significantly better on all physical measures after 3- and 12-month follow-ups. These results are consistent with earlier and more recent studies examining short-term functional effects (Tinetti et al., 2002; Parsons et al., 2013; Langeland et al., 2019). In contrast, three studies showed no statistical significance in either functional mobility or ADL (King et al., 2012; Tuntland et al., 2015; Senior et al., 2014). Moreover, the studies' results on quality of life-related measures are inconclusive, with King et al., 2012 reporting significant effects and three other studies reporting insignificant effects (Lewin et al., 2013a; Tuntland et al., 2015; Langeland et al., 2019). The most promising results of HBR are shown in potential health service reduction, such as reduced probability of residential care, emergency department visits, and hospital admissions and readmissions (Tinetti et al., 2002; Tinetti et al., 2012; Lewin et al., 2014; Senior et al., 2014). The latter findings should translate into both short-term and long-term cost reduction for HBR patients.

Three related studies estimate the cost effects of HBR (Lewin et al., 2013b; Lewin et al., 2014; Kjerstad and Tuntland, 2016). There seems to be a potential long-term cost reduction from HBR, as Lewin et al., 2013b reported cost savings but did not report on significance. From a short-term perspective, there was also a cost difference in favor of HBR, but this was not significant (Kjerstad and Tuntland, 2016), and the last study was inconclusive (Lewin et al., 2014). There are some key drawbacks of the three latter studies. Either they did not report on significance or lacked important cost elements such as hospital stay or nursing-home usage.

The present study is the first large-scale attempt to estimate the cost effects of HBR, including services from all key Norwegian healthcare providers. This is made possible by merging three national health registers in Norway at the individual level, creating a unique monthly panel. The extracted population contains data collected over 5 years, from 2011 up to and including 2015, on 155,080 individuals living in ten selected municipalities. All service data from these registers are used to estimate costs for different stakeholders based on national reported unit costs or financial reports. The data quality is high, as all registers are mandatory and are used either in a national pay-for-performance system or for comparing productivity across municipalities. The bulk of the data refers to geriatric patients above 70 years old, as the usual HBR patient is older. Estimating the cost effect of HBR based on individual data poses some interesting econometric challenges. Due to the individuality of the treatment, one should allow for heterogeneous effects in a dynamic model because current overall health costs naturally are affected by previous costs. This needs to be addressed in a setting where variables and the error term are cross-sectionally dependent. Since it is difficult to argue that the factors driving the cross-section dependence in the variables and error term are not correlated, standard estimators might contain bias. To avoid this problem, a recent large N and T estimator developed by Chudik and Pesaran, 2015 is used, and an introduction to the estimator is provided.

The results indicate that HBR patients on average have lower costs after treatment than non-HBR patients. For the main treatment variable, the estimated monthly short-run total cost savings varies between -4,875 and -6,007 NOK. This includes the first period of HBR and the entire subsequent period. On average, the mean length of the main binary treatment variable is over 15 months. The short-run effect translates into a predicted long-run cost reduction effect between -4,637 and -6,373 NOK. Moreover, the cost reduction increases as the time horizon between the first HBR period and the observation period increases. The effect is mainly found in specialist health services, which is a different area from where the intervention is implemented. Interestingly, especially for policy-makers, no effect is shown for men when regressions are conditioned on gender. Explaining the gender difference could be of interest for future research. All results pass a series of robustness tests.

This paper contributes mainly to two distinct literatures. First, the study contributes to the small but growing literature on the cost effect of new non-medical services such as vocational rehabilitation (Engström et al., 2017), cognitive-behavioral group therapy for cancer patients (Sabariego et al., 2011), tele-healthcare (Witt Udsen et al., 2017), telephone- and face-to-face-delivered counseling intervention (Berndt et al., 2016) and nurse-based case management (Seidl et al., 2015). These studies show the great variety of new ideas all aiming to improve current healthcare. They mainly tend to rely on randomized controlled trials (RCTs) or cluster trials. In many research areas, RCT is regarded as the *gold standard*, if such a standard exists (Cartwright, 2007), but RCT if not without problems, especially in social science (Deaton and Cartwright, 2018). In contrast, the current study views the introduction of HBR in Norwegian municipalities as an instance of a natural experiment and studies it with appropriate methods using data already assembled at a high cost by society. Second, the present study also sheds light on the new literature focusing on traditional primary care, for instance, increased long-term subsidization (Costa-Font et al., 2018), extended primary care practice opening hours (Bruni et al., 2016), the importance of informal care (Cecchini, 2018, Urwin et al., 2019) and long-term medical costs of Alzheimer's disease (Sopina et al., 2019).

In the next section, the Norwegian health system and implementation of HBR is explained. Section IV.3 presents the data and cost estimation, along with a description of the treatment variables and matching strategy. Section IV.4 thoroughly explains the empirical application, including the approach developed by Chudik and Pesaran, 2015. In Section IV.5, the results are presented, and Section IV.6 contains a series of robustness tests. The paper ends with a discussion of the findings and concluding remarks alongside implications for future research.

IV.2 The Norwegian health system and home-based reablement

The Norwegian national government is solely responsible for providing healthcare to the population. For instance, the share of specialist care rendered by private, for-profit hospitals is low. In 2014, such institutions provided less than 0.2% of somatic hospital stays and 6.8% of daytime stays[1]. Primary care and specialist care are organized differently. For specialist care, the Ministry of Health has a direct role through its

[1] https://helsedirektoratet.no/publikasjoner/private-aktorer-i-spesialisthelsetjenesten-omfang-og-utvikling-2010-2014

ownership of hospitals. All public hospitals are associated with one of four regional healthcare authorities (RHAs). The Ministry of Health provides each RHA with an annual budget augmented by operational directives on general goals to be achieved. RHAs are responsible for providing specialist inpatient somatic and psychiatric care alongside alcohol and substance abuse treatments. When a patient is referred to or treated at a hospital, policlinic or even a contract specialist, considerable information is digitally registered at the treatment site. A large part of this information is sent to the national health register named the *Norwegian Patient Register* (NPR), which is operated by *the Norwegian Directorate of Health* (NDH). This register plays a key role in the funding mechanism for somatic services. Hospitals combine an annual budget and a pay-for-performance system referred to as Diagnosis-related Groups (DRG), which is based on data from NPR. Other specialist services are funded by a budget, but the data are still registered in NPR.

General practitioners (GPs) and private physiotherapists are organized independently and often collaborate with peers. GPs act as gatekeepers and refer patients to more complex care services. Although GPs and private physiotherapists are self-employed, most are still part of the public system through contracts with municipalities. The GPs are financed by three different sources, capitation per patient from municipalities, patient payments and fee-for-services from the *The Norwegian Health Economics Administration* (NHEA). The fee-for-service is a pay-for-performance system where one is reimbursed according a set of detailed activities. The register for the *Control and Payment of Primary care Reimbursement Scheme* (KUHR) is the register designed to manage this system. Private physiotherapists have the same finance system, except they do not receive capitation from the municipalities. If the reader is interested in further details, Ringard et al., 2013 provides an extensive description of the Norwegian health system.

One special feature of the Norwegian health system is the municipalities' key role. They are responsible for providing a wide range of primary care services, such as rehabilitation, physiotherapy, preventive medicine, health promotion and others. Undoubtedly, the main health care services provided by the municipalities are long-term care at home (home services) and nursing homes. The *Registry for Individual-based Nursing and Care Statistics* (IPLOS) contains health information for individuals who have applied for or received primary care from the municipality. IPLOS differs from the two other registers, NPR and KUHR, as it is not used in a governmental funding system. However, it is still a legal requirement that the municipalities use IPLOS, and therefore, the quality is still regarded as high. Grants from the national government and local taxes are the main sources of financing for municipal authorities. Local municipality councils allocate these grants to all services provided by the municipality, including social care, schools, kindergartens and others. These national health registers play a key role in creating the dataset described in Section IV.3.1.

Implementing HBR is not mandatory for Norwegian municipalities, although approximately 230 out of 422 municipalities render such services, and 10 of them are included in this study. The ten included municipalities were chosen due to feasibility considerations (see Section IV.3.1) and based on a set of explicit criteria. Under (i), only municipalities that have implemented the HBR strategy alongside a conventional care strategy qualify for inclusion. Under the size criterion (ii), populous municipalities become candidates for selection. Therefore, one are aiming for high potential variation at the municipality level. Finally, to avoid capturing predominantly transient phenomena, the experience criterion (iii) favors the inclusion of municipalities that have gathered sufficient experience in administering the care strategy itself and have accumulated experience concerning the management of the respective HBR units.

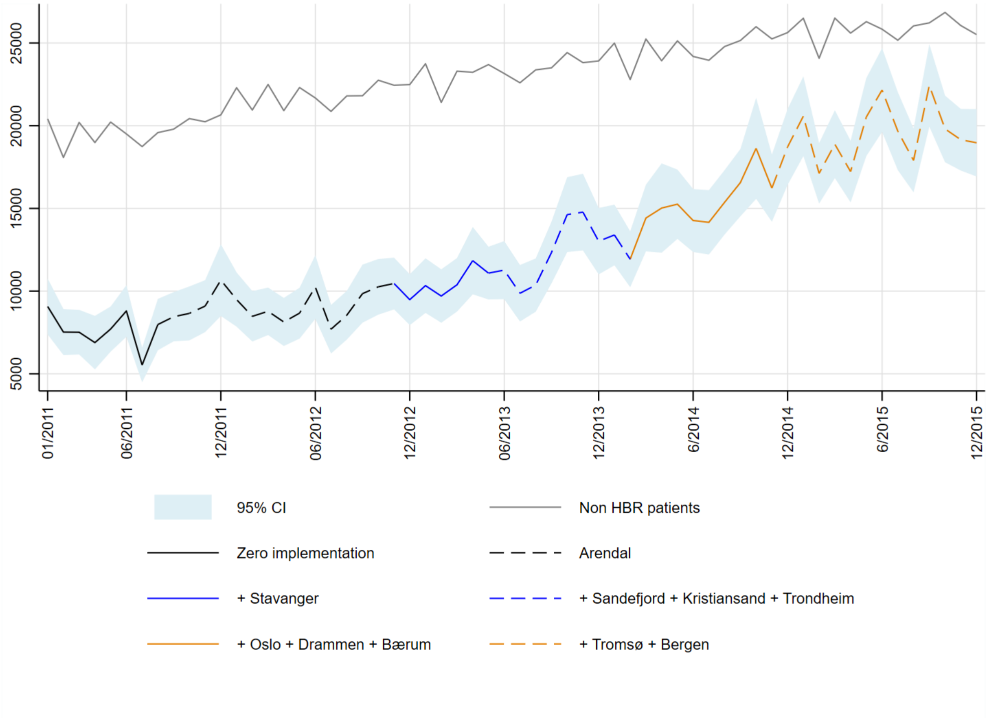

Figure IV.1: Monthly mean total cost development for HBR and non-HBR patients, 2011 - 2015, measured in NOK. The data are based on the complete sample of 155,080 individuals presented in Section IV.3.1. In addition, the figure illustrates roughly when HBR was implemented in the various municipalities considered here.

HBR shares features with other rehabilitation services and is distinguished by a reorientation of homecare away from treating diseases toward maximizing independence (Cochrane et al., 2016). Typically, a multidisciplinary team, which often comprises a nurse, physiotherapist and occupational therapist, works with the patient towards a defined goal. In HBR, significant resources are provided to focus on intensive rehabilitation measures in the patient's home for a limited period. Usual care, such as home nurse, is the typical alternative to HBR. HBR patients will often receive homecare service after HBR, but hopefully at a lower rate. Since HBR is not mandatory, the municipalities are autonomous with respect to the design of the intervention. However, for the municipalities included in our study, there is limited variation with respect to the inclusion criteria, duration of the intervention, and approach to implementation of HBR.

None of the municipalities work with explicit exclusion criteria, but they tend to administer the treatment to older, functionally impaired individuals who do not experience severe cognitive problems. Typically, an HBR phase lasts 4 weeks, with 5 treatment days per week. While in one municipality HBR lasts 6 weeks, it lasts for only 3 weeks – with 7 treatment days per week – in another. Thus, in total days, there is little variation in the length of the intervention. A patient is offered HBR as an alternative to traditional treatment. The patient then decides whether to take the offer, meaning that patients self-select when offered HBR, but few patients refuse HBR. All municipalities started with a pilot before implementing the service full time. The respective implementation years vary between 2011 and 2015, as illustrated in Figure IV.1.

The monthly mean costs of HBR and non-HBR patients are presented in Figure IV.1. The mean HBR cost is lower but increases more rapidly. The main reason that HBR has such a low mean cost is that most HBR patients are fairly new to services. This means that the observed periods before HBR treatment will have close to zero costs. This is not the case for non-HBR patients, who for instance include long-term nursing-home patients, an expensive treatment. The difference between the groups and how this is addressed are presented in Section IV.3.3.

IV.3 Data and cost estimation

IV.3.1 Data

Section IV.2 explained the functions of three Norwegian health registers. These data repositories have been assembled at high cost by Norwegian society, and data from these registers constitute the basis of this study. The three registers, IPLOS, KUHR and NPR, have been connected at an individual level, including socio-demographic data from Statistics Norway (SSB). The merged dataset contains observations collected over 5 years, from 2011 up to and including 2015, in ten selected municipalities. Merging these registers at such a large scale is, to our knowledge, not common and created substantial difficulties that need to be addressed. The root of the complications lies in the fact that IPLOS is a pseudonymous register, whereas NPR and KUHR use personal number (P.no) for identification.

As seen from the variable list presented in Table IV.14 in Appendix IV.A, IPLOS contains data on nursing home and traditional home-based services, such as homecare and home nurses. Detailed measures of other more specialized primary care services, such as elderly day care and support persons, are not included due to anonymization. Function scores, which measure ADL on 17 items, are included in the dataset. However, when a patient is not receiving services from the municipality, scores are not observed.

Data included from specialist care, obtained from NPR, include monthly aggregated service data separated into either number of episodes or sum length of stay.

The most costly diagnoses group in a given month and the individual monthly sum of DRG points are also provided. KUHR contains data from GPs and private physiotherapists. As mentioned in Section IV.2, the latter occupations are reimbursed by registering for a set of activities. The dataset contains monthly aggregate sums of these activities by different categories.

HBR is implemented in the municipalities, and therefore IPLOS is the natural starting point for the creation of the dataset. The population is all patients registered in the IPLOS at least once during the period 2011 - 2015 in the included municipalities. None of the registers included contains information concerning a patient's use of HBR, and P.no of individuals who had received HBR were collected "manually". To avoid the apparent difficulties and cost associated with collecting HBR data in this manner for all Norwegian municipalities, only observations from selected administrative entities have been included in this study. SSB merged all register data after having received them.

The end result of the merging process creates a unique panel dataset, structured such that individuals constitute the cross-section and the data are at a monthly frequency. The sample contains 155,380 individuals, and approximately 1.4% of them, i.e., 2,230 individuals, received HBR at least once. All individuals were observed for 60 months (5 years), but not all observations include service data. These empty observations are considered "healthy observations" in the sense that no health service was needed. The strategy of creating 60 observations for all individuals – irrespective of the de facto use of health services – proved to be an appropriate response to the technical problems caused by the need to merge data from several registers. Since observations following an individual's death month were deleted, exceptions to the "60 observations rule" may occur.

Obtaining access to these data is not straightforward and requires a thorough application process. The project was approved by *The Regional Committees of Medical and Health Research Ethics (REK)* and *The Norwegian Data Protection Authority (NDPA).* Necessary guidelines and verdicts from these institutions where followed. In sum, the data used in this study are costly, both in terms of time and monetary expenses. A potential risk in rich individual-specific datasets is the possibility of identifying individuals. SSB therefore needed to anonymize the data before delivering it to the researcher. SSB implements a pragmatic rule of thumb: If each cluster of identifiable variables contains at least five individuals, then the data set fulfills the anonymity requirement. The anonymization algorithm proceeds in two steps. First, it identifies clusters of fewer than 5 individuals in the space spanned by four identifiable variables: HBR group (yes/no), Gender, Hospital area and Age group. This "cluster step" is replicated for each of the 60 time periods. In the second step, the necessary deletions are carried out before the dataset is handed over. Because of anonymization difficulties, the age variable was replaced with 15 different age groups[2] and the municipality information was replaced with hospital area.

The anonymization algorithm from SSB marked 180,036 observations as suppressed, which is 2.3% of the total 7,981,709 observations. Service data on suppressed

[2]Anonymization required that age had to be aggregated into 15 different age groups: 0-39, 40-59, 60-65, 66-70, 71-72, 73-74, 75-76, 77-78, 79-80, 81-82, 83-84, 85-86, 87-88, 89-90 and 91+.

observations where deleted, and 144,979 (1.8%) of the marked observations originally contained some service data. This implies that there are 35,057 marked observations that were originally empty/healthy observations. Of the total 133,800 HBR observations, 89,747 were marked as suppressed. These are not necessarily related to the HBR period when the intervention was given but often to periods before and after the intervention. Deleted service data are equivalent to missing values, and imputation was conducted. The methods were chosen based on several test imputations on nonmissing values and the nature of the variables. A detailed description and discussion of the imputation procedure and data collection process are available (Bersvendsen and Jungeilges, 2019). In the sections below, when discussing primary care, IPLOS-related data are referenced. When discussing specialist care, data coming from NPR are used.

IV.3.2 Cost estimation

In the current setting, cost is defined as costs related to health institutions, which in Norway are run by the public sector. The costs are registered in accounts, and financial reports from 2017 are the bases for estimating costs. The necessary data are publicly available but require some institutional insight. The variable list presented in Table IV.14 also provides the unit of measure for different service variables. The cost estimation strategy is separated into three steps. First, estimate a cost per unit measured in NOK for all different service variables. The second step is to multiply observed service data by the estimated cost per unit to obtain the cost for each health service variable in monetary terms. In the final step, these service-specific costs are summarized into groups of interest, all with equal weight, that will act as dependent variables. The basic rule is that all institutional costs are included and adjusted by Norwegian VTA rules, and deviations from this rule will be noted.

Unit cost estimation for somatic specialist care services is straightforward, as NDH produces reports on institutional cost and activity usage for specialist care[3]. Somatic specialist care is financed through an activity-based system, where DRG points are the activity driver. All somatic hospital visits, policlinic visits and stays, are given DRG points. Therefore, to estimate the somatic service cost, one only needs an average cost per DRG point. This is directly obtained from published NDH reports. Psychiatric services provided by specialist care do not have an activity-based financing system. The number of psychiatric policlinic visits and days stayed are observed in the dataset. NDH provides the average cost per psychiatric visit and per day. Both of these variables are observed, meaning that all unit costs for specialist services are directly obtained from public reports. Thus, it is simply a matter of multiplying observed data by estimated average unit costs.

As mentioned in Section IV.2, the KUHR register is a system designed to accommodate reimbursement requests made to the government by health institutions and therapists. The register is owned by NDH and operated by NHEA, an external agency of NDH. GP and Physiotherapist are reimbursed based on a set of detailed activities that each have a separate code. These activities were aggregated into different

[3]www.helsedirektoratet.no/rapporter/samdata-spesialisthelsetjenesten/ Press link

categories by KUHR before being included in the dataset. However, KUHR provides a list of which activity codes are included in the different aggregated KUHR service variables. NHEA produces annual detailed reports[4] [5] presenting the total number of reimbursements and the total sum of reimbursements per activity code in NOK, which is the basis for KUHR service variable unit costs. To obtain the latter unit cost, one first has to summarize the number and total amount of reimbursement for the codes included in the specific KUHR service variable and thereafter divide these numbers. An example is illustrated below.

$$\text{Unit cost KUHR variable} = \frac{\text{Sum amount reimbursed for included codes}}{\text{Sum number of reimbursements for included codes}}$$

The main weakness in basing KUHR unit costs on reimbursement is that these reimbursements do not cover the total institutional costs of GPs and physiotherapists. To my knowledge, there is no estimated total cost per aggregated KUHR activity. Increased KUHR unit costs are tested in Section IV.6.4.

The unit cost estimation for services from NPR and KUHR are national estimates and not only based on the included municipalities. This should not be an issue, especially since all hospital areas in Norway are represented. For primary care, however, the estimated cost is based on the included municipalities. All included municipalities are larger cities, and they tend to organize their services differently from minor cities, which might influence the cost.

Table IV.1: Cost per unit for each service variable

Cost group	Variable	Per unit in NOK
Primary care cost (IPLOS)	Hours of home-nurse and homecare	594
	Days in care institution	2,627
Specialist care cost (NPR)	Total DRG points	52,442
	Policlinic contacts psychiatry	3,087
	Length of stay (days) psychiatry	12,483
GP/Physiotherapy cost (KUHR)	Other GP services	238
	Multidisciplinary cooperation meeting	404
	GP home visit	261
	GP contact by phone or mail	62
	GP office simple contact	35
	GP consultation	66
	Physiotherapy consultation	279
	Physiotherapy home visit	111
	Other Physiotherapy services	146

Estimating the unit cost for primary care based on publicly available data requires some understanding of the system and reporting structure. One needs to estimate two unit costs: (i) the cost per hour for home nurses and homecare and (ii) the cost per day for care institutions.

[4]https://www.helfo.no/om-helfo/analyse-og-kontrollrapporter-fra-helfo/ Press link
[5]https://www.helfo.no/om-helfo/analyse-og-kontrollrapporter-fra-helfo/ Press link

SSB has made data accessible on several areas, one of which is the annual financial records for municipalities. The records are reported in different three-digit codes. As an example, *253*, is called *health and care services in institutions*, and *261* is called *institution facilities*. Both of these codes cover all cost and user income related to institutions provided by municipalities. Aggregate usage of institutional days is also made public by SSB, which obtains data from the IPLOS register. In our data, the number of days in an institution covers all types of visits, short-term visits, rehabilitation and long-term visits covered by the municipality. Therefore, all registered institution days must be included when calculating unit costs. The registered cost also includes all types of institution visits, and all data are available at the municipality level[6]. One then has the following:

$$\text{Unit cost per day} = \frac{\text{Sum cost registered at 253 and 261 in 2017}}{\text{Total days provided by institutions in 2017}}.$$

The same structure can be identified for homecare and home-nursing care using the code *254*, called *health and care services to patients in private residences*. Aggregated hours of use for different IPLOS services are also made public. However, the observed amount of home-nurse and homecare hours is under code 254, but in addition, 254 also includes two types of special services that are categorized in the other care services variable. However, 86.3% of the hours are related to services included in the observed hours variable are registered under 254. Thus, the average cost per hour for 254, including the two extra services, is estimated here. This appears unlikely to affect the "true" average cost without the two additional services in 254 because they are performed by the same profession[7]. In Section IV.6.4, an adjusted price is tested. The unit cost per hour is therefore

$$\text{Unit cost per hour} = \frac{\text{Sum cost registered at 254 in 2017}}{\text{Total amount of hours provided in 2017}}.$$

Once all unit costs (see Table IV.1) are obtained, one can calculate costs for different groups that can act as dependent variables. For each patient i at time t, one estimates these overall cost groups in NOK:

$$\begin{aligned}
\text{Primary care cost}_{it} &= (Hours_{it} * 594) + (Days_{it} * 2,627) \\
\text{Specialist care cost}_{it} &= (\text{DRG points}_{it} * 52,442) + (\text{Days psychiatry}_{it} * 12,483) \\
&\quad +(\text{Policlinic contacts psychiatry}_{it} * 3,087) \\
\text{KUHR cost}_{it} &= \sum_{j=1}^{9} p_j * \text{KHUR variable}_{jit}, \\
&\quad \text{where } p_j = \text{unit cost for the } j\text{th KUHR variable} \\
\text{Total cost}_{it} &= \text{Primary care cost}_{it} + \text{Specialist care cost}_{it} \\
&\quad +\text{KUHR cost}_{it}
\end{aligned}$$

[6]For further details regarding the public data used for estimating average cost per day, see SSB tables 11644 and 12367.

[7]For further details regarding the public data used for estimating average cost per hour, see SSB tables 11643 and 12367.

The results will be presented when using *Total cost*, *Primary care cost* and *Specialist care cost* as dependent variables. The results for *KUHR cost* will be discussed in Section IV.6.4.

One potential weakness is related to the data quality of the health registers, which again depends on the fact that all agents register equally. As mentioned above, NPR and KUHR are both used in a national activity-based financing system, which has been used for several years. Clear routines for registration in these registers exist, and the data quality is regarded as high. It is mandatory for municipalities to report in the national IPLOS register. However, there are less clear routines for how to report IPLOS data. This is particularly relevant since HBR is a new service and not mandatory for the municipalities to offer. This potential drawback should be minor in the present study. Eight of the ten included municipalities participate in a collaboration network that meets regularly each year. During these meetings at the administration level, operational aspects of data registration are regularly discussed. The other two municipalities have close relationships with either one or several of the other eight. The activities of this network should minimize the potential bias introduced by different interpretations of data registration in IPLOS and financial reports sent to SSB.

IV.3.3 Matching strategy

For two reasons, I opt for a matching strategy to create a smaller and more balanced dataset that is well suited to estimate the cost effects of HBR. First, the complete dataset explained in Section IV.3.1 contains all patients who received some sort of primary care during the period 2011 - 2015. The complete dataset therefore contains patients receiving services that are not a natural logical comparison for HBR. For example, there are patients in the complete data set who never receive any amount of homecare or home-nurse hours, which is the target group for HBR. There could be patients who live in long-term institutions for the entire observed period. Additionally, the complete dataset includes young patients with serious needs, who would not be eligible for HBR. Comparing the means of *age group* and *gender* reveals a clear difference between the 2,230 HBR patients and the others. For instance, the mean of *age group* is 9.4 for the HBR group and 7.2 for non-HBR. Additionally, the proportion of men differs between the groups by 10% in absolute terms, where HBR had less. Figure IV.1 also illustrates the clear difference between HBR and all other services. The second reason for creating a relevant sub-sample is due to feasibility issues related to estimation. For the reasons presented in Section IV.4.4, it is difficult to estimate the regression model on the complete dataset because of computer power and Stata matsize issues.

Of the 2,230 HBR patients in the original dataset, 489 are dropped because at least 93% of their observations are suppressed. Which such a large number of suppressed observations, it seems meaningless to impute missing values. Fifty-two patients of non-HBR users are also dropped for the same reason. The complete dataset begins with 1,741 HBR patients and 153,098 non-HBR patients, for a total of 7,951,682 observations, before the matching strategy was implemented.

Since non-HBR patients who have not received any homecare or home-nurse hours are not a natural comparison group to HBR patients, they are dropped. The present study is interested in estimating the cost effects of HBR, including the treatment period. There are patients who are still receiving/receiving HBR for the first time in December 2015, meaning that no periods are observed after HBR. These 117 patients will not provide any information on the research question and are therefore dropped. HBR is initially designed as a home-based service and can be received either at first entry to home services or at a later stage. Ideally, one would match patients on a clearly defined cutoff for both groups. Thus, the cutoff is set as the first entry into home services, where first entry is defined as a patient with zero home-service hours 12 months before the first hours are registered. Since only a few patients received HBR in 2011, little HBR data is lost. Having such a clear cutoff will increase matching precision because it is possible to match 12 months of pre-cutoff data.

I then calculate the mean cost and standard deviation (SD) for the 12-months prior to first entry on each of the three cost groups and total cost. At the defined point, first entry has a cross-section data structure. At this point, *nearest-neighbor matching* is applied. This method pairs observations with the closest m matches between the two groups using a multi-dimensional set of variables. The number of closest matches m is set to 1, and the specific Stata program used is `nnmatch` by Abadie et al., 2004. In this program, one has to set up a dependent variable and treatment variable along with a variable list used in the matching. Total cost is set as the dependent variable and HBR group, a binary variable, is the treatment variable. The 12-month mean costs and standard deviation prior to first entry are used as matching variables. Additionally, traditional covariates such as *Age group*, *Education*, *Gender*, *Hospital areas*, *Sum observations*, *NPR main diagnoses*, *Marital status* and *Time* are used. The sum of observations is included in the variable list to ensure that HBR patients are not matched with many patients who die during the period. None of the HBR patients died during the sample period, and therefore, the sum of observations is always 60 (5 years of monthly data) for these patients. The `nnmatch` command allows for an *exact* option that allows one to specify a set of variables, where the aim is to obtain an exact matching (or one as exact as possible). The variable list for the exact option cannot overlap with the complete list of matching variables, so mean *total cost* and *time* are dropped when using the exact option. In addition to the 1,088 included HBR patients, the matching procedure found 927 non-HBR patients. This results in a matched dataset with 2,015 patients and 120,808 observations.

IV.3.4 Descriptive statistics

The matching strategy seems to be satisfactory, as all socio-demographic are fairly similar between the two groups presented in Table IV.2. There is no difference between the proportion of males in the two groups, the no-HBR group only has a 1% larger share. Martial status is also essentially identical in terms of both mean and within SD between the two groups. This is a categorical variable with 5 different categories, where 2 equals *married or in a partnership* and 3 indicates *widow or widower*. The two groups have an essentially equal distribution in percentage terms between the different categories, where 40% of the observations are registered as *widow or widower*. The two groups differ slightly in their proportion of residence. For non-HBR users, the mean proportion of patients living in hospital area East is 27% compared to 23% for the HBR group. The largest differences are found for hospital area South, where there is a 6% larger proportion of HBR patients. There is a 2% difference between the groups for hospital areas North and Middle. Regarding the SD within values for HA, they are all low and fairly equal. This indicates that some patients do move, but this is rare for both groups. More interesting, SD within is almost identical between the two groups for all HA variables. With such little movement, the possible effect of the mean difference should be captured by most panel data estimators.

As mentioned above, we do not observe age but 15 different age groups. The closest integer for the mean age group for both groups was 9, meaning that the patient was between 79 and 80 years old in 2015. The average mean age group is slightly higher for the HBR group, but both groups still have 28% of the patients in age groups 8 - 11. Education level, a 9-level ordinal variable, is identical between the groups. The closest integer for both of them is 3, which indicates high school, meaning a maximum of 12 years of school. Education category 4 also indicates high school, but with one additional year.

Table IV.2: Summary statistics

	HBR		$\neg$ HBR	
	Mean	Within SD	Mean	Within SD
Primary care cost	2,180	7,810	4,437	12,154
Specialist care cost	7,794	35,851	6,194	32,224
GP and Physiotherapy cost	123	261	130	227
Total cost	10,096	37,032	10,761	35,032
Male	0.22	-	0.23	-
Marital status	2.60	0.27	2.59	0.26
East - Hospital area	0.23	0.04	0.27	0.07
North and Middle - Hospital area	0.04	0.02	0.06	0.03
West - Hospital area	0.28	0.04	0.26	0.04
South - Hospital area	0.39	0.07	0.33	0.07
Vestfold - Hospital area	0.05	0.03	0.06	0.03
Age group	9.34	-	8.95	-
Education	3.27	0.34	3.27	0.42
FG 1 - Social functioning	0.41	0.48	0.53	0.65
FG 2 - Cognitive failure	0.38	0.46	0.50	0.63
FG 3 - Look after health	0.67	0.80	0.87	1.00
FG 4 - Housekeeping	0.72	0.88	0.78	0.94
FG 5 - Self care	0.54	0.63	0.57	0.67
n	1,088		927	
N	65,280		55,528	
$\bar{T}$	60.00		59.96	

The monthly mean and within SD for primary care costs are both lower for the

HBR group than for the matched non-HBR patients. The opposite is observed for specialist care. For GP and physiotherapy costs, the mean and SD within have minor differences. The monthly mean total cost over the 5-year period differs little. However, the within SD variation is 2,000 NOK higher for the HBR group.

Here, I use function groups (FG) instead of 17 different function score variables. Operating with FG is the norm for SSB when they publish reports based on function scores. The function groups are created in the same manner as done by the SSB, structuring 15 of the 17 different function scores[8] into 5 different groups. *Hearing* and *eyesight* are excluded because they lacked significance in a factor analysis conducted by SSB.

Mean FGs 4 and 5, *housekeeping* and *self-care*, are similar between the two groups, which also holds for SE within. Clear differences can be found for FGs 2 and 3, *cognitive failure* and *look after own health*, where the means are both lower for the HBR group. This is reasonable. One of the inclusion criteria for HBR is that patients should have some degree of cognitive presence. The reason for this is that patients need to be able to set their own goals. Looking after one's own health is also expected because HBR patients undergo a thorough program that includes several training exercises. One intention of HBR is that patients should change their mindset to be more self-reliant. There is also a marginal difference in FG 1, *social functioning*.

Table IV.3: Summary cost statistics by gender

	HBR		¬ HBR	
	Male	Female	Male	Female
Primary care cost	1,413	2,394	4,723	4,350
	(5,724)	(8,301)	(12,401)	(12,078)
Specialist care cost	8,633	7,558	7,455	5,811
	(38,522)	(35,067)	(38,869)	(29,916)
GP and physiotherapy cost	110	127	118	134
	(246)	(265)	(228)	(227)
Total cost	10,157	10,079	12,295	10,295
	(38,974)	(36,470)	(41,436)	(32,841)
Age group	8.32	9.63	8.17	9.18
Education	3.61	3.18	3.65	3.16
n	238	850	216	711
N	14,280	51,000	12,947	42,635
$\bar{T}$	60.00	60.00	59.94	59.96

Within standard deviation in parentheses

Table IV.3 presents the monthly mean costs separated by gender and the respective treatment group. Interestingly, the mean primary costs are higher for females in the HBR group and lower for specialist care compared to males. The difference in both cost groups is approximately 1,110 NOK. The total cost differences between males

[8] For further details, see Appendix A in `https://www.ssb.no/helse/artikler-og-publikasjoner/_attachment/318105?_ts=15dcac9dff0`

and females for HBR patients are minor. In the non-HBR group, there is a minor difference between genders and mean primary costs. For specialist care, the mean cost for males is also higher in the non-HBR group, which also results in a difference in mean total costs between genders. There appear to be gender differences in service usage measured by costs. Assessing whether the potential HBR effect differs between genders is therefore of interest.

IV.3.5 Treatment variables

To identify the effect of HBR on different cost outcome measures, I create binary treatment indicators. The effect on pre-treatment cost is not of interest because potential HBR patients should not differ substantial from traditional homecare patients. However, one of the goals of HBR is for the treatment to reduce the need for health services, which again should translate into reduced costs. During the HBR period, costs should actually be higher than those of traditional care because of the greater intensity of care, meaning more service hours. A potential reduction in cost post-HBR therefore comes at an additional cost. Since my interest here is in the total effect, the service amount used during HBR also needs to be captured. The main treatment indicator, T_{it}, therefore equals 1 beginning in the first month the patient received HBR for the first time and all periods thereafter. This means that the number of periods in which T_{it} equals 1 will vary because patients receive HBR at different points in time. The mean length of time from the first period of HBR to the last observed period is 15 months. To capture the development over time, three other indicators are defined with the same definition but with a different maximum number of periods for which the treatment indicator can be equal 1. These time horizons are set at 6, 8 and 12 months, and the corresponding treatment variables are defined as T_{it}^{6}, T_{it}^{8} and T_{it}^{12}. All start at the first period when a patient receives HBR for the first time. Figure IV.2 illustrates these four different indicators. I run separate regressions for each of the different treatment variables, and present results for all of them below.

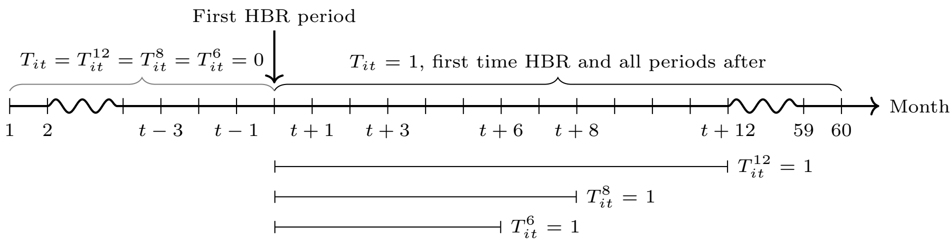

Figure IV.2: Illustrates the four different treatment indicators of interest. All start at the same point, but their time horizons differ. T_{it} does not have a limited time horizon, while T_{it}^{12}, T_{it}^{8} and T_{it}^{6} and are limited at 12, 8 and 6 months, respectively.

IV.4 Empirical strategy

IV.4.1 Regression setup

The relationship between health costs and our variables of interest is assumed to be linear. I specify the following *random trend model*

$$y_{it} = \rho y_{it-1} + \delta T_{it} + \mathbf{s}_{it}\boldsymbol{\theta} + g_i t + a_i + \eta_t + u_{it},\ i = 1, 2, ..., 2015;\ t = 1, 2, ..., 60 \quad \text{(IV.1)}$$

where each individual i is allowed to have its own time trend, $g_i t$ (Wooldridge, 2002, p. 315). In equation (IV.1), a_i captures possible unobserved area heterogeneity, and in addition to the standard random trend model, η_t is added, which accounts for unobserved time effects. The model also differs from the classical random trend model in that it is dynamic. T_{it} is the variable of interest, $\mathbf{s}_{it}$ is a $1 \times k$ row vector and contains observed covariates varying across i and t, and u_{it} denotes the error term. The unobserved coefficients ρ and δ are scalars, and $\boldsymbol{\theta}$ is a $k \times 1$ vector. Given the strict exogeneity assumption, one approach to estimate equation (IV.1) is to difference away a_i:

$$\Delta y_{it} = \rho \Delta y_{it-1} + \delta \Delta T_{it} + \Delta \mathbf{s}_{it}\boldsymbol{\theta} + g_i + \Delta \eta_t + \Delta u_{it}, \quad \text{(IV.2)}$$

where Δ denotes the *first difference*, and exploit that $g_i t - g_i(t-1) = g_i$. Equation (IV.2) now resembles a standard dynamic panel model and can be estimated with standard dynamic models. Traditionally, there are concerns with using fixed effect (FE) estimators on dynamic panel models. Such estimators are inconsistent due to the Nickell bias discovered by Nickell, 1981. In the current case, Nickell bias will not be an issue because of the large T, and thus the FE estimator would still be consistent (Judson and Owen, 1999,Baltagi, 2013, p. 155).

Table IV.4: FE estimation regression on Total cost – error term diagnostics – p-values presented

	Sub-sample					Matched
	1	2	3	4	5	data
Pesaran CD test[a]	0.000	0.000	0.000	0.000	0.000	0.000
Modified Wald test[b]	0.000	0.000	0.000	0.000	0.000	0.000
Wooldrige auto correlation test[c]	0.000	0.000	0.000	0.000	0.000	0.000
n	1,121	1,122	1,117	1,118	1,116	2,015

[a] Pesaran cross-sectional dependence test from 2004 under the null hypothesis of independence

[b] Modified Wald test under the null hypothesis of homoscedasticity

[c] Wooldridge autocorrelation test under the null hypothesis of zero correlation

Covariates included: T_{it}, lagged Total cost, five hospital areas, marital status 2 - 5, Education and time dummies

Table IV.4 presents the regression diagnostic for the model (IV.2) using the T_{it} treatment variable. Due to feasibility issues, one cannot directly perform these tests on the initial dataset. Therefore, 5 sub-samples were randomly selected, each containing 1,124 patients. Since working with first-differenced variables, and the Pesaran cross-sectional dependence test does not work well with fewer than 5 time observations (Pesaran, 2004), deviation in number of patients will occur. The latter is because different diagnostic tests is performed only on patients with 6 or more time observations, causing the n in Table IV.4 to be less than 1,124 for the sub-samples. As seen, I clearly cannot reject that the errors are autocorrelated and heteroscedastic (HAC). This holds for all sub-samples and the matched data used for the analysis. The test results presented in Table IV.4 are only for one regression setup, but they do not change when introducing different combinations of covariates. The results presented are for the random trend model, but the test result also holds for a classical two-way FE model.

IV.4.2 Cross-sectional dependence

In addition to having to address HAC, the information in Table IV.4 also implies cross-sectional dependence (CSD) in the error term,

$$cov(u_{it}, u_{jt}) \neq 0 \text{ for some } t \text{ and some } i \neq j. \tag{IV.3}$$

That is, at each time t, the errors may be correlated across patients. There are currently two main strands in the literature for dealing with CSD in panels with large N: *spatial models* and the *residual multifactor approach*, also known as the *factor structure approach* (Pesaran and Tosetti, 2011; Sarafidis and Wansbeek, 2012). Spatial models use the concept of a distance metric, assuming that the structure of CSD is related to distance and locations of units. Such a pre-defined metric is often difficult to find, but there are studies that use this concept with an *economic distance* (Sarafidis and Wansbeek, 2012). The alternative approach, the *factor structure approach*, assumes that the error term contains a finite number of unobserved common factors such that,

$$u_{it} = \boldsymbol{\lambda}_i' \mathbf{f}_t + \varepsilon_{it}, \tag{IV.4}$$

where $\mathbf{f}_t$ is an $m \times 1$ vector of unknown common factors, $\boldsymbol{\lambda}_i$ is an $m \times 1$ vector of unknown factor loadings, and ε_{it} denotes the idiosyncratic error that is assumed to be independently distributed among the regressors. This allows the unknown common factors, economy-wide shocks, to influence all units with different intensities. Such a factor structure approach is well suited to the current research question for reasons elaborated in the discussion in section (IV.7). Spatial CSD can be viewed as a special factor structure approach (Sarafidis and Wansbeek, 2012). Bai, 2009 illustrated that the model presented in equation (IV.1) is actually a special case of the factor structure error structure. Let $m = 2$ and the common factors and loadings be such for all i and t,

$$\mathbf{f_t} = \begin{bmatrix} 1 \\ \eta_t \end{bmatrix} \text{ and } \boldsymbol{\gamma}_i = \begin{bmatrix} a_i \\ 1 \end{bmatrix}, \tag{IV.5}$$

then

$$\boldsymbol{\gamma}_i' \mathbf{f}_t = a_i + \eta_t. \tag{IV.6}$$

The appropriate estimation technique for models with CSD largely depends on the sources and nature of these CSDs. Especially for non-dynamic models, the key question is whether the unobserved common factors, $\mathbf{f}_t$ in equation (IV.4), are correlated with the regressors.

Table IV.5: Variable-specific Pesaran CD test for cross-sectional dependence

	Sub-sample										Matched data	
	1		2		3		4		5			
	CD[a]	p[b]	CD[a]	p[b]	CD[a]	p[b]	CD[a]	p[b]	CD[a]	p[b]	CD[a]	p[b]
ΔTotal cost	12.8	0.00	15.9	0.00	22.7	0.00	22.7	0.00	21.3	0.00	9.5	0.00
ΔPrimary cost	124.7	0.00	117.3	0.00	132.1	0.00	127.6	0.00	115.8	0.00	79.9	0.00
ΔSpecialist cost	2.8	0.00	4.3	0.00	8.3	0.00	9.1	0.00	9.0	0.00	8.4	0.00
ΔGP/Phys cost	47.3	0.00	38.7	0.00	45.5	0.00	39.1	0.00	28.4	0.00	72.9	0.00
ΔHBR Days	0.0	0.94	0.0	1.00	0.0	0.99	0.0	0.99	0.0	1.00	12.6	0.00
ΔEducation	0.1	0.97	0.5	0.61	0.7	0.50	0.8	0.43	0.3	0.73	2.5	0.01
ΔFG 1	4.3	0.00	2.9	0.00	2.3	0.02	5.5	0.00	4.3	0.00	20.4	0.00
ΔFG 2	4.0	0.00	2.5	0.01	3.1	0.02	4.0	0.00	3.5	0.00	19.0	0.00
ΔFG 3	3.6	0.00	2.4	0.02	2.2	0.03	3.4	0.00	1.9	0.06	18.8	0.00
ΔFG 4	4.1	0.00	1.9	0.05	1.9	0.08	3.6	0.00	2.3	0.02	23.8	0.00
ΔFG 5	5.1	0.00	4.7	0.00	5.6	0.00	6.5	0.00	4.1	0.00	22.5	0.00

[a] The CD-statistic is distributed $\mathcal{N}(0,1)$, and extreme values indicate strong correlation

[b] *p-values*, and H_0 : cross-sectional independence

The results in Table IV.5 clearly suggest that for the matched data, all variables that potentially could be included in the regression and are CSD. This is also confirmed by studying the five random sub-samples, except *HBR days* and *Education.* The latter evidence, in addition to the results from Table IV.4, clearly indicates that there is CSD in the regressors and error terms. Assuming that the common factors causing CSD in the error term and regressors are independent of one another, conventional panel data estimators will still be consistent but naturally affect inference (Chudik and Pesaran, 2013). In this case, one could obtain valid inference by, for instance, using a nonparametric robust variance-covariance matrix discussed by Driscoll and Kraay, 1998.

That the regressors are uncorrelated with unobserved common factors is the sufficient condition for consistency, but if this is invalid, then conventional estimators may be inconsistent (Andrews, 2005). However, in dynamic panel models, the latter argument may not hold. Phillips and Sul, 2007 show that the dynamic panel data model with CSD estimated with FE will be inconsistent even with independent common factors. These findings are in line with recent results from Everaert and De Groote, 2016, who also warn against using FE on a dynamic model even with large T. In the current research setting, these considerations need to be taken into account when estimating the model, not only because of the dynamic model but also because the common factors are most likely not independent of the regressors. The logic behind this assumption is presented in the discussion in section (IV.7).

IV.4.3 Heterogenous coefficients

As in most treatment effect studies, the mean treatment effect δ in (IV.2) is of interest. Traditionally, this translates into estimating a regression model with homogenous coefficients, such as equation (IV.1). For HBR-specific reasons, it seems logical that the regressors have heterogeneous effects on health costs, especially because of the micro-level data. Patients live in different surroundings, there homes differ, and they have different relationships, previous health status and so forth. Some patients might need more home-nurse hours because their homes are impractical, even though their function scores might not differ. Patients will also react differently to HBR since it is a personalized treatment with patient-specific goals.

The homogenous coefficient assumption seems implausible for many applications and can actually cause conventional models to be inconsistent where the bias can be substantial (Pesaran and Smith, 1995). As shown by Pesaran and Smith, 1995, pooled and aggregated estimators are not consistent in dynamic models, even with large N and T. They show that when regressors are serially correlated, incorrectly ignoring coefficient heterogeneity creates serial correlation in the disturbance. Even with $T \to \infty$, this causes inconsistent estimates in models with lagged dependent variables. Their *mean group* (MG) estimator, running N regressions and averaging the coefficients, is consistent for large N and T.

Table IV.6: $\tilde{\Delta}$ test for slope homogeneity[9]

Matched data	
$\tilde{\Delta}$ Statistic	p-value
-19.007	0.000***

Model (2) in Table IV.7 with T_{it} is used, $\tau = 3$
$\tilde{\Delta}$ distributed $\mathcal{N}(0,1)$, H_0: homogenous slopes
* p<0.10, ** p<0.05, *** p<0.01

Pesaran and Yamagata, 2008 proposed a standardized version of Swamy's test of slope homogeneity for panel data under the null of homogeneity, where N could be relatively large compared to T. They proposed a test statistic denoted $\tilde{\Delta}$, which is shown to have a standard normal distribution as $(N,T) \xrightarrow{j} \infty$ such that $\sqrt{N}/T^2 \to 0$ or $N/T \to \kappa$ for stationary dynamic models. Table IV.6 presents the $\tilde{\Delta}$ and the respective p-value test results, and one reject the null hypothesis of homogenous slopes at the 1% significance level[9]. The results in Table IV.6 clearly provide indications that homogenous coefficients seem unrealistic in the current setting. The result also holds for different models, including different sets of covariates.

For illustration and not inference, I estimate model (IV.2) with FE and MG using total cost as the dependent variable and obtain the following estimates, $\hat{\delta}^{FE}$, $\hat{\delta}_i^{MG}$, $\hat{\rho}^{FE}$ and $\hat{\rho}_i^{MG}$. Based on these results, I estimate the individual difference between FE estimates and individual-specific MG estimated coefficients, with the following definitions $\delta_i^{Diff} = \hat{\delta}^{FE} - \hat{\delta}_i^{MG}$ and $\rho_i^{Diff} = \hat{\rho}^{FE} - \hat{\rho}_i^{MG}$. The latter individual differences, δ_i^{Diff} and ρ_i^{Diff}, are plotted and presented in Figure IV.3 to demonstrate coefficient variability.

One would expect that if homogenous slopes where true, there would be some minor variations around the zero mark. However, as seen in Figure IV.3, there is a wide spread for δ_i^{Diff} and ρ_i^{Diff} . In particular since, the plots in Figure IV.3 do not contain the upper or lower 5% percentiles to avoid extreme values for robust inspections. The blue dots indicate patients for which the individual MG coefficients $\hat{\delta}_i^{MG}$ are zero, indicating the non-HBR group. These blue dots do not occur in ρ_i^{Diff} because all patients have a non-zero MG coefficient. Interestingly, when the mass is centered close to zero, as for ρ_i^{Diff}, the FE and mean MG coefficients do not differ greatly.

[9]The `xthst` command by *Tore Berscendsen and Jan Ditzen* is used. Originally, the test by Pesaran and Yamagata, 2008 is proposed for strictly exogenous regressors or autoregressive models. However, when developing `xthst` simulation results show that by adding cross sectional averages to the model, the test works for situations with cross-section dependence in the error term and variables. The latter is therefore used when performing in the current setting. Presented results do also hold for non-dynamic model without cross sectional averages and standard AR model.

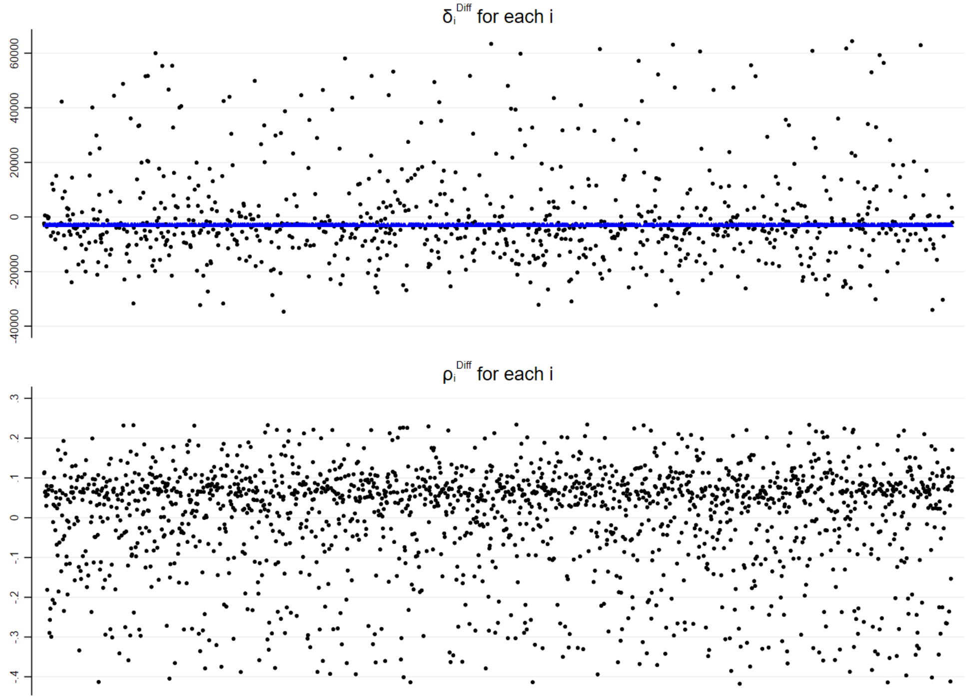

Figure IV.3: Difference between FE coefficients and individual-specific MG coefficients, δ_i^{Diff} and ρ_i^{Diff}, for model (IV.2). The covariates included in the estimation are Education, five hospital areas, marital status 2 - 5 and quarterly time dummies

Based on the $\tilde{\Delta}$ test results, Figure IV.3 and reasoning, I select a model with heterogeneous coefficients with a common factor error structure.

$$\begin{aligned} \Delta y_{it} &= \rho_i \Delta y_{it-1} + \delta_i \Delta T_{it} + \Delta \mathbf{s}_{it} \boldsymbol{\theta}_i + g_i + \boldsymbol{\gamma}_i' \mathbf{f}_t + \varepsilon_{it}, \\ & \quad \text{where } i = 1, 2, ..., 2015; \; t = 2, 3, ..., 60. \end{aligned} \tag{IV.7}$$

Note that equation (IV.7) does not contain η_t unobserved time effects, as in (IV.2, since these will be captured by $\mathbf{f}_t$.

IV.4.4 Estimation

Pesaran, 2006 proposed an approach to estimate panel data models with a common factor error structure called *Common Correlated Effects* (CCE). The CCE estimator of heterogeneous coefficients is consistent as $(N, T) \xrightarrow{j} \infty$, provided that certain rank conditions regarding factor loadings are satisfied. The CCE estimator was originally not proposed to cover dynamic models, and Chudik and Pesaran, 2015 extended the CCE approach to allow lagged dependent variables as regressors. They demonstrate that the cross-sectional averages of observed variables and their lags are able to sufficiently approximate the unobserved factors $\mathbf{f}_t$. This result holds under certain conditions, the key assumptions of which are presented below.

In application, this implies that one runs N least squares regressions of equation (IV.7), including the cross-sectional averages and the their lags of all observed variables as regressors. The number of lags must be chosen. Thereafter, one estimates the average coefficient based on all estimated coefficients. A brief summary of the estimator presented by Chudik and Pesaran, 2015 follows below.

Let us slightly rewrite equation (IV.7),

$$\Delta y_{it} = \rho_i \Delta y_{it-1} + \Delta \mathbf{x}_{it} \boldsymbol{\beta}_i + g_i + u_{it}, \tag{IV.8}$$

where u_{it} follows the same common factor error structure as in equation (IV.4),

$$u_{it} = \boldsymbol{\gamma}_i' \mathbf{f}_t + \varepsilon_{it}.$$

In (IV.8), $\mathbf{x}_{it}$ contains both $\mathbf{s}_{it}$ and T_{it} from equation (IV.7) and therefore has dimension $1 \times (k+1)$. Since the unobserved factors $\mathbf{f}_t$ could be correlated with the regressors, a general model for individual-specific regressors is adopted (Pesaran, 2006):

$$\Delta \mathbf{x}_{it}' = \mathbf{r}_i + \boldsymbol{\Gamma}_i' \mathbf{f}_t + \boldsymbol{\upsilon}_{it}, \tag{IV.9}$$

where $\boldsymbol{\Gamma}_i$ is an $m \times (k+1)$ matrix. The regressors are also determined by some unobserved individual fixed effect $\mathbf{r}_i$, and $\boldsymbol{\upsilon}_{it}$ are idiosyncratic errors. Define $\mathbf{C}_i = (\boldsymbol{\gamma}_i, \boldsymbol{\Gamma}_i)'$, $\boldsymbol{\pi}_i = (\rho_i, \boldsymbol{\beta}_i)'$ and $\mathbf{z}_{it} = (\Delta y_{it}, \Delta \mathbf{x}_{it})'$, and then the following six dynamic CCE assumptions are presented:

Assumption 1: The individual error terms ε_{it} and $\boldsymbol{\upsilon}_{it}$ are distributed independently for all i and t. The $\boldsymbol{\varepsilon}_t$ vector is weakly cross-sectionally dependent.

Assumption 2: The unobserved common factors, $\mathbf{f}_t$, are covariance stationary with absolute summable autocovariances and are independently distributed from the individual errors presented above.

Assumption 3: Factor loadings $\boldsymbol{\gamma}_i'$ and $\boldsymbol{\Gamma}_i'$ are independently and identically distributed across i and $\mathbf{f}_t$, for all i and t, with mean $\boldsymbol{\gamma}$ and $\boldsymbol{\Gamma}$. They are generated from the following random model:

$$\boldsymbol{\gamma}_i = \boldsymbol{\gamma} + \eta_{\gamma i},\ \eta_{\gamma i} \sim IID(\mathbf{0}, \boldsymbol{\Omega}_\gamma),\ \text{for } i = 1, 2, ..., N,$$

and

$$vec(\boldsymbol{\Gamma}_i) = vec(\boldsymbol{\Gamma}) + \eta_{\Gamma i},\ \eta_{\Gamma i} \sim IID(\mathbf{0}, \boldsymbol{\Omega}_\Gamma),\ \text{for } i = 1, 2, ..., N.$$

Assumption 4: The vector of coefficients, $\boldsymbol{\pi}_i$ follows the random coefficient model

$$\boldsymbol{\pi}_i = \boldsymbol{\pi} + \eta_{\pi i},\ \eta_{\pi i} \sim IID(\mathbf{0}, \boldsymbol{\Omega}_\pi),\ \text{for } i = 1, 2, ..., N.$$

Assumption 5: Regressors and covariates are either strictly exogenous or weakly exogenous. For the latter, certain assumptions must be fulfilled. See Chudik and Pesaran, 2015 page 396 for further details.

In the current case, strict exogeneity is assumed, with $E(\mathbf{x}_{is}\varepsilon_{it}) = 0$ for all s and t.

Assumption 6: Define $\mathbf{C} = E(\mathbf{C}_i) = (\boldsymbol{\gamma}, \boldsymbol{\Gamma})'$. The $(k+1+1) \times m$ matrix $\mathbf{C}$ has full column rank.

This is a key assumption, with some estimation implications discussed in Section IV.4.5. This also means that the number of unobserved factors is no larger than the number of observed variables, $m \leq (k+1+1)$.

Before expressing the estimated coefficients denoted $\hat{\boldsymbol{\pi}}_i = (\hat{\rho}_i, \hat{\boldsymbol{\beta}}_i)'$, one first has to define the following matrices:

$$\boldsymbol{\Xi}_i = \begin{pmatrix} y_{i,\tau} & \mathbf{x}_{i,\tau+1} \\ y_{i,\tau+1} & \mathbf{x}_{i,\tau+2} \\ \vdots & \vdots \\ y_{i,T-1} & \mathbf{x}_{i,T} \end{pmatrix}, \ \bar{\mathbf{Q}}_\omega = \begin{pmatrix} 1 & \bar{\mathbf{z}}'_{\omega,\tau+1} & \bar{\mathbf{z}}'_{\omega,\tau} & \cdots & \bar{\mathbf{z}}'_{\omega,1} \\ 1 & \bar{\mathbf{z}}'_{\omega,\tau+2} & \bar{\mathbf{z}}'_{\omega,\tau+1} & \cdots & \bar{\mathbf{z}}'_{\omega,2} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & \bar{\mathbf{z}}'_{\omega,T} & \bar{\mathbf{z}}'_{\omega,T-1} & \cdots & \bar{\mathbf{z}}'_{\omega,T-\tau} \end{pmatrix}, \tag{IV.10}$$

where $\bar{\mathbf{z}}_{\omega t} = (\Delta\bar{y}_{\omega t}, \Delta\bar{\mathbf{x}}_{\omega t})' = \sum_{i=1}^{N} \omega_i \mathbf{z}_{it}$. This is an $k+1+1$ vector of weighted cross-sectional averages of the *observed* variables in (IV.7) or (IV.8), and τ is the number of chosen lags. The initial recommended setting for how many lagged cross-sectional averages to include in the regression is $\tau = T^{1/3}$. However, additional simulations by Chudik and Pesaran, 2015 show that the choice of lag order depends on the object of interest, β or ρ. In the former case, $0.75\tau = T^{1/3}$ seems to be preferable, and this setting is therefore used in our estimation. The weights, ω_i, need to satisfy some conditions. Let $\mathbf{w} = (\omega_1, \omega_2, ..., \omega_N)$ be an $N \times 1$ vector, and the following conditions need to be fulfilled:

$$\|\mathbf{w}\| = O(N^{-\frac{1}{2}}), \ \frac{\omega_i}{\|\mathbf{w}\|} = O(N^{-\frac{1}{2}}) \text{ and } \sum_{i=1}^{N} \omega_i = 1.$$

The projection matrix $\bar{\mathbf{M}}_q = \mathbf{I}_{T-\tau} - \bar{\mathbf{Q}}_\omega(\bar{\mathbf{Q}}'_\omega\bar{\mathbf{Q}}_\omega)^+\bar{\mathbf{Q}}'_\omega$ where $\mathbf{I}_{T-\tau}$ is the identity matrix with dimensions $(T-\tau)\times(T-\tau)$ and $^+$ denotes the Moore-Penrose generalized inverse. Given that assumptions 1 - 6 hold, the coefficient can be expressed as

$$\hat{\boldsymbol{\pi}}_i = (\boldsymbol{\Xi}'_i\bar{\mathbf{M}}_q\boldsymbol{\Xi}_i)^{-1}\boldsymbol{\Xi}'_i\bar{\mathbf{M}}_q\mathbf{y}_i, \tag{IV.11}$$

where $\mathbf{y}_i = (y_{i,\tau+1}, y_{i,\tau+1}, ..., y_{i,T})'$. The mean group estimator, like the coefficients presented in Section IV.5, is then given by

$$\hat{\boldsymbol{\pi}} = \frac{1}{N}\sum_{i=1}^{N} \hat{\boldsymbol{\pi}}_i. \tag{IV.12}$$

The mean group estimator in (IV.12) is the *Dynamic CCE* (DCCE) estimator presented by Chudik and Pesaran, 2015. This estimator is applied separately for each of the four different treatment variables explained in Section IV.3.5.

Since least squares is performed on all N patients, the standard least squares assumptions must be satisfied at the individual level. This is especially important in the current case since several variables are categorical. One covariate in the DCCE setup adds $1+\tau$ extra variables to the regression, which will make it difficult to fulfill the full rank assumption. The regressions can therefore not include all covariates, and several setups need to be tested. Computing separate regressions for more than 150,000 individuals is not ideal and is time consuming. One would encounter computational limitations, which is another argument for creating a matched dataset.

IV.4.5 Potential endogenous regressors and modeling technicalities

Including the function group variables in the regression will be an issue due to the strict exogeneity assumption. The HBR treatment would affect function scores and therefore make the function group variable endogenous. Function scores could actually work as independent variables. The relationship could also run in the opposite direction; function scores would decide whether a patient can receive HBR. This could potentially make the treatment variables endogenous. Either way, an endogenous variable would violate the strict exogeneity assumption and potentially bias the results. Therefore, the function group variables are not included in the regression models. The regression model controls for unobserved time-invariant individuality and trend. Some heterogeneity across patients is captured, and allowing for a heterogeneous slope would, to some degree, compensate for the loss of precision by not being able to control for function scores.

One way to estimate a DCCE regression in statistical software such as Stata is to first run N regressions including the $\tau + 1$ additional cross-sectional mean variables. The N estimates are thereafter averaged to obtain the mean estimated coefficients. This procedure is time-consuming, especially since each N regression will contain several variables. A faster approach is to utilize the *Frisch-Waugh-Lovell* theorem (Frisch and Waugh, 1933; Lovell, 1963) and partial out the cross-sectional averages. Next, run N regressions on the partitioned variables. This is the procedure used by the Stata `xtdcce2` command developed by Ditzen, 2018.

Assumption 6, full column rank of $\mathbf{C}$, is required for estimating individual coefficients. However, it is not always necessary when focusing on the overall mean estimates, $\hat{\boldsymbol{\pi}}$. When the common factors, $\mathbf{f}_t$, are serially uncorrelated, the full column rank is not needed for consistency. To assess whether the common factors in the current study are serially correlated is difficult because they are unobserved, which also means that $\mathbf{C}$ is unobserved.

The cross-sectional averages and their lags are used for approximating $\mathbf{C}$. Thus, a matrix $\mathbf{Z}$, which includes all $\tau + 1$ cross-sectional averages of $\mathbf{z}_{it}$, is used for this approximation. Collinearity checks are necessary to ensure that $\mathbf{Z}'\mathbf{Z}$ is invertible and thus that assumption 6 is not validated. Including hospital area and marital status in $\mathbf{z}_{it}$ would cause $\mathbf{Z}'\mathbf{Z}$ to violate assumption 6 because of collinearity issues. The reason for this conclusion requires further insight into IPLOS registration. All data updates capturing changes in education level, marital status and hospital area are performed once each year. Thus, changes in data will always occur in January of each year, so at 4 possible points in time during the sample period. Martial status and area are categorical variables, which are captured by dummy variables in the regression setup. There is limited movement between the different martial status categories, and all changes occur simultaneously. The cross-sectional averages of these dummies will nearly perfectly predict one another, causing collinearity issues because a patient can only be in one category. For example, if a patient moves for marital status 2 to 5, the cross-sectional mean marital status dummy variable 2 would decrease by the same amount as the increase in marital status dummy 5. These two dummies would therefore perfectly predict one another and, by definition, make the variables collinear. The latter argument holds for hospital areas even though few patients move to non-included municipalities.

Education level is an ordered variable and can be included in $\mathbf{z}_{it}$. However, as mentioned above, changes in education also occur only in January. Education level hardly changes among the included patients, which makes sense because they are mostly elderly. Since there are few changes, the cross-sectional averages for the change in education level in January can actually predict changes in martial status and hospital areas. This would violate the collinearity assumption. To avoid such issues, cross-sectional averages of changes in education level cannot be combined with martial status and hospital ares in the regression setup. Because of the above data idiosyncrasies, some changes need to be made in (IV.7).

Covariates that cannot be included in $\mathbf{z}_{it}$, meaning hospital area or martial status, are captured by $\mathbf{c}_{it}$. Because of matsize issues, neither marital status nor hospital area dummies can be included in the same model. When adding $\mathbf{c}_{it}$, $\mathbf{s}_{it}$ will be reduced to covariates that can be included in the cross-sectional averages, HBR days and education level. After these adjustments and adding cross-sectional averages, the regression model can be written as follows:

$$\Delta y_{it} = \rho_i \Delta y_{it-1} + \delta_i \Delta T_{it} + \Delta \mathbf{s}_{it}\boldsymbol{\theta}_i + \sum_{\tau=0}^{\tau} \boldsymbol{\mu}_{i,\tau}\bar{\mathbf{z}}_{t-\tau} + \Delta \mathbf{c}_{it}\boldsymbol{\phi}_i + \varepsilon_{it}, \quad \text{(IV.13)}$$
$$\text{where } i = 1, 2, ..., 2015;\ t = 2, 3, ..., 60.$$

where $\boldsymbol{\mu}_{i,\tau}$ represents the $\tau + 1$ coefficients for the added cross-sectional averages. The dimension of $\mathbf{s}_{it}$ has been reduced to $1 \times (k - l)$, and $\mathbf{c}_{it}$ has dimension $1 \times l$. Equation (IV.13) is estimated by MG to obtain DCCE estimates presented below.

IV.5 Results

The DCCE regression results are presented for three different dependent variables: (i) total cost, (ii) primary care cost and (iii) specialist care cost, as described in Section IV.3.2. All tables presented follow the same structure, where the results of the different treatment indicators are presented alongside the lagged dependent variable. For primary care cost, the results for the HBR days variable are also presented. The results are presented for models with difference covariate settings. The long-run treatment effect, $\hat{\delta}_{LR}$, is estimated as follows:

$$\hat{\delta}_{LR} = \frac{1}{N}\left(\sum_{i=1}^{N}\frac{\hat{\delta}_i}{1-\hat{\rho}_i}\right),$$

where $\hat{\delta}_i$ and $\hat{\rho}_i$ are individual estimates obtained from (IV.13).

IV.5.1 Total costs

As seen in Table IV.7, the DCCE estimated coefficient for T_{it}, $\hat{\delta}$, is clearly significant in all specifications. The respective p-value runs from 0.001 - 0.002, and all $\hat{\delta}$ coefficients are negative, which indicates that HBR significantly reduces total health costs. The estimated cost reduction from the first month of HBR varies from -4,875 to -6,007 NOK.

The above pattern repeats when studying the estimated coefficient, $\hat{\delta}$, for the two treatment variables, T^6 and T^8. All estimated coefficients, $\hat{\delta}$, are significant at the 1% level. The $\hat{\delta}$ coefficients vary between -2,784 and -3,630, all indicating that HBR already has a cost reduction effect 6 months after the first HBR period. The cost reduction effect is on average 2,225 NOK less for T^6 than for the unrestricted T_{it} treatment variable. However, for T^{12}, the estimated cost reduction is lower and significant at the 5% level, except in (4), where $\hat{\delta}$ is only significant at the 10% level. This is surprising, and there could be a health-increasing cost event for HBR patients between 8 and 12 months after the first HBR period compared with usual care. The cause of this increase is unknown and should be investigated in future research.

Table IV.7: Total cost regression results – Dynamic CCE, $\tau = 3$

	(1) Δ Total cost	(2) Δ Total cost	(3) Δ Total cost	(4) Δ Total cost
$\Delta\bar{T}_{it}$	-4875.3***	-5860.8***	-5872.8***	-6007.0***
	(0.001)	(0.002)	(0.002)	(0.001)
Δ Lagged Total cost	-0.405***	-0.402***	-0.401***	-0.401***
	(0.000)	(0.000)	(0.000)	(0.000)
R^2	0.76	0.76	0.73	0.74
$\Delta\bar{T}^{12}$	-2442.3**	-2980.1**	-2770.6**	-2810.1*
	(0.037)	(0.049)	(0.064)	(0.061)
Δ Lagged Total cost	-0.406***	-0.412***	-0.410***	-0.410***
	(0.000)	(0.000)	(0.000)	(0.000)
R^2	0.77	0.73	0.73	0.74
$\Delta\bar{T}^{8}$	-3124.1***	-3576.6***	-3559.4***	-3530.8***
	(0.005)	(0.008)	(0.008)	(0.009)
Δ Lagged Total cost	-0.409***	-0.420***	-0.416***	-0.417***
	(0.000)	(0.000)	(0.000)	(0.000)
R^2	0.77	0.73	0.73	0.74
$\Delta\bar{T}^{6}$	-2783.7***	-3574.9***	-3621.4***	-3629.9***
	(0.003)	(0.002)	(0.002)	(0.002)
Δ Lagged Total cost	-0.411***	-0.423***	-0.420***	-0.420***
	(0.000)	(0.000)	(0.000)	(0.000)
R^2	0.76	0.72	0.72	0.73
Included in, $\mathbf{s}_{it}$, cross-sectional averages added:				
Δ HBR days		Yes	Yes	Yes
Δ Education level		Yes		
Included in, $\mathbf{c}_{it}$, cross-sectional averages not added:				
Δ Marital status 2-5			Yes	
Δ 5 hospital areas				Yes

p-values in parentheses
* p<0.10, ** p<0.05, *** p<0.01

The estimated dynamic coefficient, $\hat{\rho}$, is clearly significant for all models independent of treatment variables. There is little variation in the coefficients, as all are just above -0.4. At first glance, the negative sign seems illogical. This indicates that the cost in the previous period reduces that in the current period. However, there is a logical explanation for the negative sign. The dependent variable, total cost, contains the cost of three different stakeholders of health providers, where primary care and specialist care, is by a large margin the largest contributor. This is clear in the descriptive statistics in Table IV.2. Specialist health services, often provided in a hospital, are usually costly and intensive, and all are performed within a short time span. Primary care costs are lower, and the services provided each month vary little. Much of the limited variation observed between months in primary care is actually explained by the number of days in each month. The variation in total cost will therefore mostly be due to specialist health services. Therefore, if one patient ended an intensive hospital stay in the previous period, $t-1$, the cost in the current period, t, would definitely be less than in $t-1$, which would result in a negative coefficient. Following this analogy, one would not expect such a clear negative coefficient when only studying primary care costs. All other included variables are insignificant at the 5% level except for martial status dummy number 2, indicating *married or in partnership*, which exhibits a higher cost compared to un-married patients.

Because this is a dynamic model, one can estimate the potential long-run effect. As noted above, the estimated lagged dependent coefficient, $\hat{\rho}$, is essentially equal across the different treatment variables. This would mean that the long-run estimates would follow the same development as described above, all depending on which treatment variable is the basis for the estimates. Based on T_{it}, the estimated long-run effect, $\hat{\rho}_{LR}$, will vary between -4,637 and -6,373 NOK, depending on the model considered.

IV.5.2 Primary care costs

Following the above argument regarding the lagged dependent variable coefficient, ρ, the results in Table IV.8 do not indicate a clear significant negative coefficient. For models (2) - (4), all estimated lagged primary care cost coefficients, $\hat{\rho}$, are positive, which holds for all treatment variables. Because of the slow development of primary care costs, no significance is observe for the dynamic coefficient $\hat{\rho}$ at the 5% level for models (2) - (4). As an example of the slow-moving dynamics, the mean primary care cost in December 2015 for the treatment group is only 207 NOK higher than that in January 2015.

Studying model (1) in Table IV.8 reveals the opposite results on the lagged dependent coefficient, ρ. Model (1) results in a negative coefficient and is significant across all treatment variables. The same difference in coefficients between model (1) and the rest can be found when studying the treatment variables. The key difference between models (1) and (2) - (4) is the introduction of HBR days provided. This variable captures the intensity, treatment length, and the weight of HBR provided between different months if the treatment runs over several periods. This variable is equal to zero whenever HBR is not provided. Controlling for HBR days seems logical, especially when focusing on primary care, and the HBR days coefficient is clearly significant in all models and across treatment variables. The HBR coefficient size varies between 132 and 155, meaning that having more HBR days will increase costs. This is expected because HBR is an intensive treatment.

Studying the estimated $\hat{\delta}$ related to the different treatment variables for models (2) - (4) reveals an interesting pattern. The monthly cost savings from HBR increases over time after the first HBR treatment. For variable T_{it}, the estimated monthly cost savings, $\hat{\delta}$, varies between -1,227 and -1,257 NOK across models (2) - (4). This value is almost double that of T^6, which varies between -724 and -749 NOK in monthly cost savings. When estimating $\hat{\delta}$ for the indicators T^8 and T^{12}, cost savings increases gradually between the two. All estimated coefficients $\hat{\delta}$ are significant at the 5% level. Model (1) actually shows the complete opposite result from models (2) - (4). However, as the former does not control for HBR days, (1) is therefore less able to capture the individuality of the HBR treatment. The HBR intervention is person-centered and goal-oriented. One patient can need fewer days to complete the agreed goal than another patient. The length of the intervention will therefore vary across patients, and longer intervention times would entail higher primary care costs. This variation would not be captured in a running treatment variable but is captured when including HBR days. The latter variable could to some degree explain the HBR patients' function level during HBR treatment because one would expect a less functional patient to use more HBR days. Logically, capturing the individuality of the treatment seems important.

The long-run estimates, $\hat{\rho}_{LR}$, based on models (2) - (4) using T_{it} as the treatment variable vary between -1,300 and -1,848 NOK. For model (1), the estimated long-run effect is 661.

Table IV.8: Primary care cost regression results – Dynamic CCE, $\tau = 3$

	(1) Δ Primary care cost	(2) Δ Primary care cost	(3) Δ Primary care cost	(4) Δ Primary care cost
$\Delta\bar{T}_{it}$	974.8***	-1227.1***	-1255.9***	-1257.1***
	(0.000)	(0.001)	(0.001)	(0.001)
Δ Lagged Primary care cost	-0.0558***	0.0571*	0.0486	0.0510*
	(0.000)	(0.057)	(0.103)	(0.086)
Δ HBR days		153.4***	154.7***	154.7***
		(0.000)	(0.000)	(0.000)
R^2	0.80	0.74	0.75	0.75
$\Delta\bar{T}^{12}$	577.0**	-1014.4***	-1047.3***	-1038.7***
	(0.015)	(0.004)	(0.003)	(0.004)
Δ Lagged Primary care cost	-0.0549***	0.0631*	0.0551	0.0578*
	(0.000)	(0.065)	(0.105)	(0.088)
Δ HBR days		132.3***	133.3***	133.6***
		(0.000)	(0.000)	(0.000)
R^2	0.82	0.75	0.75	0.75
$\Delta\bar{T}^{8}$	623.8***	-893.6**	-941.6***	-934.2**
	(0.005)	(0.013)	(0.009)	(0.009)
Δ Lagged Primary care cost	-0.0831***	0.0216	0.0196	0.0192
	(0.000)	(0.479)	(0.516)	(0.525)
Δ HBR days		135.5***	137.0***	137.1***
		(0.000)	(0.000)	(0.000)
R^2	0.82	0.75	0.75	0.76
$\Delta\bar{T}^{6}$	435.9**	-724.1**	-745.1**	-748.5**
	(0.045)	(0.027)	(0.022)	(0.021)
Δ Lagged Primary care cost	-0.0699***	0.0109	0.00791	0.00751
	(0.000)	(0.624)	(0.719)	(0.733)
Δ HBR days		139.0***	139.8***	140.0***
		(0.000)	(0.000)	(0.000)
R^2	0.83	0.74	0.75	0.75
Included in, $\mathbf{s}_{it}$, cross-sectional averages added:				
Δ Education level		Yes		
Included in, $\mathbf{c}_{it}$, cross-sectional averages not added:				
Δ Marital status 2-5			Yes	
Δ 5 hospital areas				Yes

p-values in parentheses
* p<0.10, ** p<0.05, *** p<0.01

IV.5.3 Specialist care costs

The results when specialist care cost is the dependent variable are presented in Table IV.9. In short, the results in Table IV.9 are essentially identical to the findings for total costs. However, the $\hat{\delta}$ coefficients in model (1) are in the same region as for models (2) - (4), whereas for the corresponding total cost coefficients, there was a small jump between (1) and (2) - (4). Controlling for HBR days caused this change[10]. This effect is not observed for specialist health costs.

For the treatment variables, T_{it}, T^6 and T^8, all estimated coefficients, $\hat{\delta}$, are significant at the 5% level. For T_{it}, estimated cost savings varies between -5,663 and -5,923 NOK. Regarding the lagged dependent variable, ρ, all coefficients are significant and approximately 0.03 smaller than those estimated for total cost. The interpretation presented above for total cost also holds for specialist care costs.

The estimated lagged dependent coefficient, $\hat{\rho}$, is fairly stable in all settings. Therefore, as seen for total costs, the long-run estimates would follow the same pattern as the short-run findings in Table IV.9, depending on which indicator is the basis for the estimates. Based on T_{it}, the estimated long-run effect $\hat{\rho}_{LR}$ for specialist care costs will vary between -5,317 and -5,650 NOK, depending on the model considered.

[10]Test regressions were estimated without introducing HBR days and Education only and vice versa. The effect was clearly caused by HBR days

Table IV.9: Specialist care cost regression results – Dynamic CCE, $\tau = 3$

	(1) Δ Specialist health cost	(2) Δ Specialist health cost	(3) Δ Specialist health cost	(4) Δ Specialist health cost
$\Delta\bar{T}_{it}$	-5922.5***	-5662.6***	-5710.6***	-5838.3***
	(0.000)	(0.002)	(0.002)	(0.002)
Δ Lagged Specialist care cost	-0.432***	-0.432***	-0.431***	-0.431***
	(0.000)	(0.000)	(0.000)	(0.000)
R^2	0.75	0.73	0.73	0.73
$\Delta\bar{T}^{12}$	-3174.4***	-3020.7**	-2755.3*	-2799.8**
	(0.005)	(0.044)	(0.063)	(0.059)
Δ Lagged Specialist care cost	-0.433***	-0.440***	-0.439***	-0.439***
	(0.000)	(0.000)	(0.000)	(0.000)
R^2	0.76	0.73	0.73	0.73
$\Delta\bar{T}^{8}$	-3651.4***	-3271.2**	-3231.3**	-3204.4**
	(0.001)	(0.015)	(0.015)	(0.016)
Δ Lagged Specialist care cost	-0.435***	-0.448***	-0.445***	-0.445***
	(0.000)	(0.000)	(0.000)	(0.000)
R^2	0.76	0.72	0.73	0.73
$\Delta\bar{T}^{6}$	-3126.0***	-3232.8***	-3239.1***	-3252.0***
	(0.001)	(0.004)	(0.004)	(0.004)
Δ Lagged Specialist care cost	-0.439***	-0.453***	-0.451***	-0.450***
	(0.000)	(0.000)	(0.000)	(0.000)
R^2	0.76	0.71	0.72	0.72
Included in, $\mathbf{s}_{it}$, cross-sectional averages added:				
Δ HBR days		Yes	Yes	Yes
Δ Education level		Yes		
Included in, $\mathbf{c}_{it}$, cross-sectional averages not added:				
Δ Marital status 2-5			Yes	
Δ 5 hospital areas				Yes

p-values in parentheses
* p<0.10, ** p<0.05, *** p<0.01

IV.5.4 Gender-specific regressions

To estimate whether there is a potential treatment effect difference between genders, I conducted DCCE estimation conditional on the binary gender variable. Interestingly, this reveals some clear differences. Table IV.10 presents the conditional regression results for males only on total cost. None of the different $\hat{\delta}$ coefficients are significant at any level, but most of the coefficients are negative. This is in stark contrast to the earlier results presented in Table IV.7, where all coefficients were significant at some level. This finding indicates that all previous significant treatment effects are driven by females, as males have no significant effect. In Appendix IV.C, conditional regression results for all dependent variables are presented by gender. Regardless of whether total cost, primary care or specialist care cost are considered, all of the results have the same pattern. Females have a significant cost reduction effect from HBR, while males do not.

The lagged dependent variables are essentially identical between genders and are clearly significant for total and specialist health costs. HBR days are also only significant when primary care is the dependent variable for both genders. The coefficients are, however, different between the two. One additional HBR day for males, ceteris paribus, increases primary costs by approximately 320 NOK, which is more than 3 times higher than that for females. Interestingly, the mean hours one month prior to the first HBR period are actually 2 hours lower for men. Mean hours in the first month of HBR are, however, equal for both genders. Therefore, there was a larger increase in service hours between the two time periods for men.

Table IV.10: Males only – Total cost regression results – Dynamic CCE, $\tau = 3$

	(1) Δ Total cost	(2) Δ Total cost	(3) Δ Total cost	(4) Δ Total cost
$\Delta\bar{T}_{it}$	356.4	-1067.6	-1013.7	-1248.9
	(0.890)	(0.771)	(0.782)	(0.732)
Δ Lagged Total cost	-0.400***	-0.405***	-0.396***	-0.398***
	(0.000)	(0.000)	(0.000)	(0.000)
R^2	0.78	0.76	0.76	0.77
$\Delta\bar{T}^{12}$	1810.3	947.7	922.0	662.0
	(0.414)	(0.780)	(0.779)	(0.841)
Δ Lagged Total cost	-0.395***	-0.404***	-0.395***	-0.397***
	(0.000)	(0.000)	(0.000)	(0.000)
R^2	0.78	0.76	0.75	0.76
$\Delta\bar{T}^{8}$	942.7	-893.4	-383.7	-953.8
	(0.654)	(0.739)	(0.885)	(0.721)
Δ Lagged Total cost	-0.402***	-0.411***	-0.404***	-0.405***
	(0.000)	(0.000)	(0.000)	(0.000)
R^2	0.78	0.76	0.77	0.77
$\Delta\bar{T}^{6}$	913.5	-1546.5	-1386.6	-1520.7
	(0.633)	(0.438)	(0.487)	(0.445)
Δ Lagged Total cost	-0.398***	-0.414***	-0.406***	-0.407***
	(0.000)	(0.000)	(0.000)	(0.000)
R^2	0.75	0.75	0.75	0.76
Included in, $\mathbf{s}_{it}$, cross-sectional averages added:				
Δ HBR days		Yes	Yes	Yes
Δ Education level		Yes		
Included in, $\mathbf{c}_{it}$, cross-sectional averages not added:				
Δ Marital status 2-5			Yes	
Δ 5 hospital areas				Yes

p-values in parentheses
* p<0.10, ** p<0.05, *** p<0.01

IV.6 Robustness tests

To my knowledge, there is no clear alternative estimator to DCCE that could play a constructive role in a robustness test. The simulation results presented by Chudik and Pesaran, 2015 show that not all potential bias in the estimated coefficients is removed with DCCE. They suggest some alternative bias corrections that could work as robustness tests. These results are presented below. Since there could be uncertainty in the cost per hour of home-nurse care and KUHR unit costs might be low, results with adjusted unit costs are presented below.

IV.6.1 Different lags, τ, selections

As shown in Section IV.4.4, the DCCE estimator depends on the number of added lagged cross-sectional average variables chosen, τ. Chudik and Pesaran, 2015 note that choosing lag order is difficult in practice because it depends on several unknown aspects. They suggest that different lag orders could be tested and investigated. This would be especially important for the current case because the lag chosen is smaller than the recommended setting. This was discussed in Section IV.4.4, and therefore the recommended setting, $\tau = 4$, should be studied in detail. The results for $\tau = 1, 2, 4$ are presented in Appendix IV.D.

For total cost, setting the lag order to 4 hardly changes the different estimated treatment effects. The estimated cost reduction from HBR is well within the same range as presented previously, which holds for all different treatment variables. Regarding inference, all $\hat{\delta}$ for T_{it} and T^6 and are significant at the 1% level, which is equal to the previously presented results. Regarding the estimated coefficient, $\hat{\delta}$, for the T^8 variable, the significance level changes to 5% for models (2) - (4), where the original results were significant at the 1% level. For estimated $\hat{\delta}$ reflecting the T^{12} variable, the significance level goes from 5% to 10% for models (2) and (3). The latter is also the least convincing of the previously presented results. Regarding the estimated lagged dependent variable coefficients, $\hat{\rho}$, they are essentially equal and all clearly significant. The results for lag settings 1 and 2 do not change the inference from the original results. All $\hat{\delta}$ coefficients for all the different treatment variables are close to the original results, with one exception. For the main treatment variable, T_{it}, the estimated cost reduction is approximately 1,200 higher with only one lag for models (2) - (4). As seen in the presented results, there is a shift between model (1) and the other models for all lag settings, but the difference is less pronounced for $\tau = 4$.

Changing lag settings for DCCE estimation using primary care cost as the dependent variable reveals the same pattern as presented above. For all lag settings, the signs of the estimated coefficients change after controlling for HBR days. The estimated treatment effect $\hat{\delta}$ in models (2) - (4) increases as the time horizon for the treatment variables increases. All $\hat{\delta}$ coefficients for models (2) - (4), regarding the lag setting, are in the same region as presented above, and the minimum significance level is 5%. HBR days are also significant in all settings, and the coefficient is almost the identical to that in the $\tau = 3$ setting. The estimated lagged dependent coefficient

$\hat{\rho}$ is still not significant.

Regarding specialist health costs, all estimated $\hat{\delta}$ coefficients are in the same general range as the previous results. The exception to this rule is for $\tau = 1$. Here, the estimated cost reduction of HBR is higher overall, especially for the main treatment variable, T_{it}. This is the same finding as discussed earlier for total costs. The estimated coefficient on the lagged dependent variable, $\hat{\rho}$, is essentially identical for all considered values of τ, which also holds for the significance level. In conclusion, the presented results seem to hold for all dependent variables for different settings of τ.

IV.6.2 Jackknife bias-correction

Simulations provided by Chudik and Pesaran, 2015 show that the DCCE technique outperforms regular estimators for δ and the lagged dependent coefficient, ρ. However, not all potential biases might be removed through DCCE estimation. They show that *jackknife* bias correction works well but tends to slightly over-correct, and their supplementary material shows that this holds for different simulation settings. The *root mean square error* (RMSE) for the ρ parameter is also reduced with the adjustment, which holds for all combinations of N and T. However, the simulation results are different for regressors, such as δ in equation (IV.13). For the regressors, the simulation results showed that the DCCE estimator almost removes all bias in the estimated coefficient and that the jackknife adjustment only slightly improves the results. This improvement comes at a cost, as the related RMSE actually increases in the jackknife result. This would affect the p-values and, naturally, inference. Therefore, the jackknife result is used as a robustness check for the estimated lagged dependent coefficient $\hat{\rho}$ and related inference. Since RMSE increases with the jackknife adjustment for δ, this adjustment only acts as a robustness test for estimated coefficients $\hat{\delta}$. As a robustness test regarding inference for $\hat{\delta}$, I employ bootstrap-corrected standard errors. These results are presented in Section IV.6.3.

Table IV.11 presents the results for total cost estimated with the jackknife-adjusted DCCE with a lag order of 3. The results for the two other dependent variables and for $\tau = 4$ are presented in Appendix IV.E. As seen in Table IV.11, for all 4 models and across all different treatment variables, the estimated lagged dependent coefficient $\hat{\rho}$ is clearly significant. All coefficients remain negative but are, however, slightly lower than the results presented earlier in Table IV.7. The results presented in the latter table show that the $\hat{\rho}$ coefficient varied from approximately -0.40 to -0.42 depending on which treatment variable was under scrutiny. With jackknife-adjusted estimates, these coefficients vary between -0.46 and -0.51. This would lead to a reduced potential long-run effect. Based on the simulation results obtained by Chudik and Pesaran, 2015, the latter results are over-adjusted but still closer to the *true value* than the results without the adjustment. Tables IV.E.1 and IV.E.2 present the jackknife-adjusted estimates for primary care and specialist health care costs. In the latter case, there are few changes in the lagged coefficients relative to the previously presented results, and they are all clearly significant.

Table IV.11: Total cost regression results – Dynamic CCE, $\tau = 3$, Jackknife adjusted

	(1) Δ Total cost	(2) Δ Total cost	(3) Δ Total cost	(4) Δ Total cost
$\Delta\bar{T}_{it}$	-6305.7***	-5351.0	-7036.2*	-7211.7***
	(0.004)	(0.102)	(0.011)	(0.009)
Δ Lagged Total cost	-0.487***	-0.464***	-0.480***	-0.481***
	(0.000)	(0.000)	(0.000)	(0.000)
$\Delta\bar{T}^{12}$	-2166.1	-3545.2	-2285.2	-2392.1
	(0.200)	(0.139)	(0.303)	(0.281)
Δ Lagged Total cost	-0.489***	-0.486***	-0.491***	-0.492***
	(0.000)	(0.000)	(0.000)	(0.000)
$\Delta\bar{T}^{8}$	-3173.2**	-5080.4**	-3750.3*	-3736.8*
	(0.046)	(0.017)	(0.053)	(0.054)
Δ Lagged Total cost	-0.490***	-0.501***	-0.496***	-0.497***
	(0.000)	(0.000)	(0.000)	(0.000)
$\Delta\bar{T}^{6}$	-3169.4**	-3290.7*	-4089.1*	-4121.7**
	(0.023)	(0.074)	(0.019)	(0.018)
Δ Lagged Total cost	-0.484***	-0.510***	-0.493***	-0.493***
	(0.000)	(0.000)	(0.000)	(0.000)
Included in, $\mathbf{s}_{it}$, cross-sectional averages added:				
Δ HBR days		Yes	Yes	Yes
Δ Education level		Yes		
Included in, $\mathbf{c}_{it}$, cross-sectional averages not added:				
Δ Marital status 2-5			Yes	
Δ 5 hospital areas				Yes

p-values in parentheses
* p<0.10, ** p<0.05, *** p<0.01

Studying the different treatment coefficients, as presented in Table IV.11, reveals greater differences between the 4 models compared to the earlier results. All coefficients are similar to the previously presented results in Table IV.7, with two exceptions. For T^8 in model (2) and T_{it} in model (1), the jackknife-adjusted coefficient estimates are 1,500 and 1,430 NOK lower, indicating a higher cost of HBR compared to the initial results. However, this does not change the fact presented earlier that HBR seems to reduce total costs. For primary care costs only, the jackknife results, shown in Table IV.E.1, remain similar to the original results. This also holds for specialist health care costs presented in Table IV.E.2.

IV.6.3 Wild bootstrap standard error correction

As a robustness check for treatment effect inference, I compare the initial p-values with p-values based on wild bootstrap-adjusted standard errors. Because of feasibility issues, wild bootstrap-adjusted standard errors are only considered using the treatment variable T_{it}, with $\tau = 3$. Applying the bootstrap technique to the DCCE model is time-consuming. For example, running the 4 models for one dependent variable and only one treatment indicator requires more than 14 hours for 50 repetitions. These estimates are calculated using the Stata 15 multicore version; 50 repetitions may be sufficient to capture the trend.

Table IV.12: Comparing initial and wild bootstrap-adjusted p-values for the main treatment indicator, T_{it}

			(1)	(2)	(3)	(4)
Total cost	Initial results	p-value	0.001	0.002	0.002	0.001
	Wild bootstrap	p-value	0.000	0.003	0.096	0.039
Primary care	Initial results	p-value	0.000	0.001	0.001	0.001
	Wild bootstrap	p-value	0.000	0.000	0.000	0.003
Specialist care	Initial results	p-value	0.000	0.002	0.002	0.002
	Wild bootstrap	p-value	0.000	0.025	0.027	0.152

Models 1 - 5 are identical to earlier models. Wild bootstrap is conducted with 50 repetitions. Lag order, τ, is set to 3

Table IV.12 provides a comparison with the p-values of the initial model, with the same values reflecting the wild bootstrap-adjusted standard errors. All initial results had a p-value lower than 0.01, hence an significance level of 1%, for all models and dependent variables. The wild bootstrap results show the same significance level with 5 exceptions. In 3 of these cases, models (2) and (3) with specialist care costs as the dependent variable and model (4) with total costs as the dependent variable, adjusted based p-values, still indicate a significance level of 5%. The largest difference is found for model (4), where specialist care cost is the dependent variable. If one were to base inference on the bootstrap-adjusted p-value for the latter case, the conclusion would have been the opposite of the to initial results. However, the latter case exhibits an abnormal trend and therefore argues that the initial results hold and thus pass the robustness test.

IV.6.4 Sensitivity analysis

As mentioned in Section IV.3.2, costs reported in the publicly accessible data for homecare and home nurse contain two additional types of services. They constitute 13.7% of all home hours registered in public databases. The results presented above are based on estimated unit costs when these two additional services, *care salary for private persons* and *assistance outside institution*, are included. One could argue that these services are performed by less-experienced employees with lower wages and that they therefore lower the overall estimated unit cost per hour used. As a robustness test to address this potential error, I estimate the DCCE regression on total and primary care costs with increased unit costs per hour. The increased cost is estimated by including the cost of the two additional services but not including the hours. The same cost is therefore divided by 13.7% fewer hours, hence the increased unit cost. This change yields in a unit cost per hour of 688 NOK compared to 594 NOK initial value. Appendix IV.F presents the results with the increased unit cost, and the presented results do not change.

Presented in Section IV.3.2, the estimated unit cost of KUHR services, GP/ physiotherapy, is based on reimbursement rates. This income is only one of several income sources for GPs and physiotherapists and would therefore not cover all costs associated with these services. Unofficial estimates by KUHR approximate that these reimbursements cover 30% - 40% of total GP/physiotherapy costs. As a test, KHUR cost is increased by 65%, and thereafter, DCCE estimation is conducted on adjusted total cost, which includes this increase. The results from the latter setup are presented in Table IV.13. The results with this adjustment are essentially identical to the earlier results in Table IV.7. The results are not sensitive to changes in the unit cost of KUHR services because there are no significant differences between the groups in GP/physiotherapy costs.

Table IV.13: Total cost with 65% increase in KUHR unit cost, $\tau = 3$

	(1) Δ Total cost adj.	(2) Δ Total cost adj.	(3) Δ Total cost adj.	(4) Δ Total cost adj.
$\Delta\bar{T}_{it}$	-4866.8***	-5837.5***	-5849.0***	-5983.6***
	(0.001)	(0.002)	(0.002)	(0.001)
Δ Lagged Total cost adj.	-0.406***	-0.403***	-0.402***	-0.402***
	(0.000)	(0.000)	(0.000)	(0.000)
R^2	0.76	0.73	0.73	0.74
$\Delta\bar{T}^{12}$	-2439.0**	-2971.4**	-2762.5*	-2802.0*
	(0.037)	(0.049)	(0.064)	(0.061)
Δ Lagged Total cost adj.	-0.406***	-0.413***	-0.411***	-0.411***
	(0.000)	(0.000)	(0.000)	(0.000)
R^2	0.77	0.73	0.73	0.74
$\Delta\bar{T}^{8}$	-3120.4***	-3565.2***	-3547.2***	-3518.8***
	(0.005)	(0.009)	(0.008)	(0.009)
Δ Lagged Total cost adj.	-0.410***	-0.420***	-0.417***	-0.417***
	(0.000)	(0.000)	(0.000)	(0.000)
R^2	0.77	0.73	0.73	0.74
$\Delta\bar{T}^{6}$	-2774.8***	-3564.6***	-3609.5***	-3618.4***
	(0.003)	(0.002)	(0.002)	(0.002)
Δ Lagged Total cost adj.	-0.411***	-0.424***	-0.421***	-0.421***
	(0.000)	(0.000)	(0.000)	(0.000)
R^2	0.76	0.72	0.72	0.73
Included in, $\mathbf{s}_{it}$, cross-sectional averages added:				
Δ HBR days		Yes	Yes	Yes
Δ Education level		Yes		
Included in, $\mathbf{c}_{it}$, cross-sectional averages not added:				
Δ Marital status 2-5			Yes	
Δ 5 hospital areas				Yes

p-values in parentheses
* p<0.10, ** p<0.05, *** p<0.01

IV.7 Discussion and conclusion

The analysis above estimated institutional service cost differences between patients receiving HBR and those receiving traditional services. The choice of using the DCCE estimator instead of standard panel data estimators mainly relies on one key argument relating to the common factor in equation (IV.4), $\mathbf{f}_t$. This (these) unobserved common factor(s) is (are) arguably closely related to the potential reason that the regressors are cross-sectionally dependent. The argument is best explained with a concrete example of one potential unobserved common factor. In January 2012, namely at a single point in time, the Norwegian government implemented a national health reform called *The Coordination Reform*. This reform affected basically all health stakeholders in Norway and, therefore, the dependent variables and again the regressors since I use a dynamic model. Traditionally, one would run a two-way FE model that would include $t-1$ time dummies and argue that the effect of the reform would be captured by these dummies.

As mentioned in equations (IV.5) - (IV.6), the two-way model is a special version of the common factor error structure. The problem with the latter solution is that one assumes that such a reform would affect every individual equally. This is probably far from the truth, as the immediate effect would differ between municipalities and hospitals and therefore also lead to individual differences. Another unobserved common factor that most likely has considerable influence is how municipalities have organized the decision-making regarding service allocation. All operate with administrative offices, which make the decisions before primary care services are executed. All of the employees in these offices have similar backgrounds and eduction, and they will follow certain norms. These norms, however, are likely to differ slightly across municipalities and also across individuals.

As shown in Table IV.4, the test results for the classical two-way model clearly indicated the presence of cross-sectional dependence. The p-values for the CD-statistic are 0.000. Considering the same statistic for the DCCE regression results on T_{it} reveals a low CD-statistic with a p-value varying between $|0.43|$ and $|0.96|$. The latter results clearly indicate that the DCCE estimator worked as intended. However, these results do not hold for primary care costs. The CD statistic remains high in this case, resulting in a p-value of 0.000. This finding indicates that the primary care results might have some limitations. Future research should test whether primary care remains CSD. There might be a deeper econometric issue related to the slow movement in the dependent variable.

The DCCE estimator used in this paper is basically a version of the MG estimator. Thus, one estimate N different coefficients and use the mean as a consistent estimator of the *true value.* Whether the mean is relevant for the current case is debatable. As seen in Figure IV.4, there is substantial variation in the estimated individual treatment effect related to T_{it} in model (1) on total cost. As shown in Table IV.7, the coefficient is -4,875, which includes all non-HBR patients with zero effect. The mean coefficient for HBR patients is only -9,029, and the individual estimated coefficients, $\hat{\delta}_i$, for HBR patients are the basis for Figure IV.4. As seen in Figure IV.4, the median effect is actually positive at 812 NOK, which would indicate that HBR actually increases total costs. The mean HBR only $\hat{\delta}_i$ coefficient without the upper and lower 5% percentiles is -3,769. The estimated overall mean treatment effect based on the latter is still negative at -1,842 but indicates less cost reduction than originally presented. This shows that the estimated treatment effect is influenced by extreme cases. It is not the case that all HBR patients experience a cost reduction from the treatment. Several HBR patients actually exhibit a cost increase. In total, there are 1,088 HBR patients in the sample; 559 have a positive δ coefficient, and 529 have a negative coefficient. Therefore, 30 less patients have cost reductions than experience increased costs according to T_{it}. Policy-makers should be aware that not all patients are affected by HBR and that there is considerable variability in the potential effects. The consistency of the DCCE estimator is based on the mean. Further research could investigate the consistency of this estimator under other alternatives, such as the median.

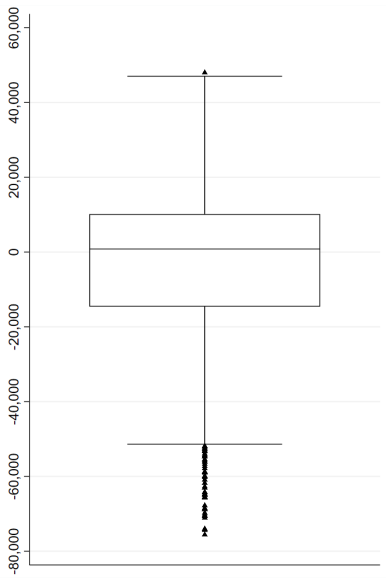

Figure IV.4: Box-plot of individual $\hat{\delta}_i$ coefficients from model (1) in Table IV.7 for the HBR group. Upper and lower 5% percentiles excluded

The results in Table IV.7 clearly indicate that HBR reduces the monthly short-run total cost. Interestingly, the effect grows at the time horizon expends. Setting the time horizon at 6 or 8 months after the first intervention results in approximately, depending on the model, 1,800 - 2,500 NOK less in cost reduction relative to the case when there is no limit on the time horizon. The latter finding is definitely the desired pattern. However, setting the time horizon at a maximum of 12 months after the first HBR period yields results than contradict the previous pattern. Here, the estimated short-run coefficients are at the lowest, and the significance level has dropped to 5% for models (1) - (3) and to 10 % for model (4). All the other results based on total costs are significant at the 1% level. Future research could explain this drop between 8 and 12 months after intervention. This finding could explain why Kjerstad and Tuntland, 2016 did not find any significant results after 9 months. Studying primary care costs alone reveals the desired pattern with increased cost reduction over time after first HBR treatment. The estimated monthly cost reduction is, however, smaller when studying primary care only and varies between 720 and 1,260 NOK. There is a larger effect when studying only specialist health costs than when considering only primary care. The findings on specialist health costs are closely related to total costs and follow an identical pattern as total costs. Independent of treatment variable, model, and dependent variable, the R^2 is always above 0.71, which would indicate that models have good fit.

The findings related to primary care costs in the current study are in line with previous results. Lewin et al., 2013b estimated the 5-year cost savings from HBR on homecare service costs in Australia. They concluded that HBR patients used AU $12,600 less then those using conventional care. In terms of the monthly cost reduction, this translates to AU $210, which is equivalent to 1,319 NOK[11], which are close to findings in Table IV.8 for the main treatment variable. Similar findings are obtained in Norway by Kjerstad and Tuntland, 2016, where their estimated 9-month cost reduction was 953 NOK, which again is close to the findings presented for the T^8 variable. In the latter study, the findings were insignificant, and the Australian study did not provide inference. The study by Lewin et al., 2014 did include some hospital costs alongside homecare costs, and their estimated monthly cost savings were AU $181, equivalent to 1,137 NOK[10]. The key difference between the current study and previous efforts is that all specialist care costs and nursing home costs are also included here. This benefits the current study because the main cost reduction is found for specialist care, which previous studies were not able to capture.

The latter finding is especially interesting for policy-makers because HBR is promoted and implemented by municipalities, hence for primary care. Thus, one agent makes another agent better off, ceteris paribus. In Norway, these two agents collaborate but report to different agencies and are not linked financially. If specialist health care costs decline due to external decisions, these providers will fill that free capacity with other patients. These providers are paid via pay-for-performance and have incentives to fill their capacity in the short run with other services. By making one stakeholder better off, society would actually generate more health services, meaning increased health costs, ceteris paribus. If the overall economic goal is to reduce total health costs, such an effect would actually work against that goal. The same argument will hold for municipalities to some degree. If HBR is not treated as a substitute but represents an additional service, it would actually generate more primary care services, ceteris paribus. However, the municipalities are the only provider of primary care and also control HBR, and adjustments should therefore be easier. Regarding total cost, converging towards a long-run *steady-state* equilibrium would reduce costs by between -4,637 and -6,373 NOK per month based on the main treatment variable.

For municipal policy-makers, the gender-specific results pose some interesting questions. All previously discussed results are driven by females, as conditional regression on males revealed no HBR effect on costs. Designing gender-specific interventions would generate a series of ethical questions. When speaking to practitioners, two main

[11] Based on the exchange rate on November 8, AU $1 = 6.2803 NOK.

explanations were proposed for this finding. First, providers might not be effective enough in setting goals for men. As most HBR workers are women, there could be communication difficulties in motivating and discussing goals. The second explanation is closely related to the first but reflect that HBR focuses on typical tasks. The rehabilitation focuses mainly on ADLs, such as cooking, cleaning, and doing laundry. The typical HBR patient is elderly, usually older than 70 years, and traditionally males of that generation are not responsible for these classical household activities. Attempting to motivate them to "re-learn" skills that might not be intuitive for men could therefore be difficult, and HBR would not make a difference. Future research could investigate whether there are aspects of HBR one could adjust such that men will find it more useful.

The present study is not without limitations, and in particular, two relate to estimated costs. First, during the HBR intervention period, the cost per hour is potentially higher than that used when estimating costs here. During the intervention period, both physiotherapists and occupational therapists are usually involved in providing the registered service hours. In Norway, these occupations traditionally tend to have a higher salary than nurses and auxiliary nurses. The price per hour used here reflects cost of the latter two occupations, as these are the dominant ones in municipal homecare services. A slightly increased price during the HBR period would also increase the cost of HBR. However, the cost increase would be minor because the change would only occur for a few observations. This would probably not be sufficient to change the fact that HBR seems to reduce costs, but the cost effect would be smaller. As mentioned in Section IV.3.2, the estimated KUHR unit cost is based on reimbursements and not the total cost relating to these services. As shown in Section IV.6.4, DCCE regressions including a 65% increase in GP and physiotherapy costs did not reveal any differences from the previously presented results. The potential issues in using reimbursement as a basis for unit cost appear limited.

Because of legal rules, some service data were deleted because of anonymization and were later imputed. To minimize the potential downside of these missing values, an adequate imputation strategy was chosen. The strategy relied on findings from Kleinke et al., 2011, which show that traditional imputation techniques were far from being outperformed by newer techniques specifically designed for panel data. The missing values follow a uniform pattern and are *missing at random* (MAR). The latter feature is especially desirable because most imputation techniques assume MAR. Because of the uniform missing pattern, it was possible to impute variables separately. The techniques adopted here were therefore based on variable-specific characteristics. Testing the applied techniques on real data revealed a clear pattern. The imputed results were nearly identical to the true mean and standard deviation. However, all imputations overestimated the within standard deviation, which is the variation used in DCCE estimation. This could lead to biased coefficients and increased standard errors, which would affect inference. Dropping missing data and avoiding imputation is also not a viable option because it would eliminate too much HBR data. All results pass a series of robustness tests, indicating that potential bias introduced because of imputed data might not affect the results. However, regardless of how good state-of-the-art imputation techniques are, they introduce uncertainty because they are not the "true" observations.

As discussed in Section IV.4.5, it is not possible to use the function group variables in the regression because of endogeneity issues. Both groups have equal scores the period before the first entry into the service. However, the function scores prior to first entry are based on only the few patients who received some other service, such as a *support person*. Most persons in the sample actually have a zero function score prior to the first entry. The dataset therefore does not contain any good measure of the function level prior to first entry. Lagged costs are potentially the best measure available in the dataset. Costs should reflect service usage, and service usage should to some degree reflect function or sickness level. The model used controls for all unobserved time-invariant factors and potential trends, so a large degree of uncertainty could be captured. However, one cannot completely rule out the possibility that the effects observed above are solely due to HBR itself; they might also capture some unobserved pre-treatment functional differences or other unobserved time-varying changes. Such a change could be motivated by HBR patients changing their motivation for living more healthfully with increased knowledge after treatment. The effect could therefore be caused by a change in the mental state and not due to HBR training per se. Either way, HBR had a positive effect, what causes the effect would just differ.

In conclusion, HBR shows some promising results in terms of cost reduction. From a positivistic perspective, the largest estimated potential monthly cost savings of HBR is 6,000 NOK, as shown in Table IV.7 model (4). On an annual basis, this would be 72,000 NOK per HBR patient, which is just above 10% of the cost of one person-year for nurses[12]. Technically, this means that after 10 HBR patients, one could eliminate one full-time worker the first year after the patients received HBR, ceteris paribus. Again, this is a positivistic view but illustrates the potential benefits of HBR. However, realizing the potential benefits is more difficult because the main effects is found at a different stakeholder. However, the cost reduction is most likely insufficient to compensate for the coming demographical challenges. To do so, more structural changes would need to occur, but HBR and its ideas are a part of the solution. Currently, HBR has few exclusion criteria, and thus there is substantial large variation in the individual treatment effects. Future research should investigate what distinguishes patients with and without effects such that the potential cost-reducing effect of HBR could be maximized.

[12]Based on a rough estimate of nurses work day/evening shifts, this would be 700,000 NOK for one year.

IV.8 References

Abadie, Alberto et al. (2004). “Implementing Matching Estimators for Average Treatment Effects in Stata”. In: *The Stata Journal* 4.3, pp. 290–311.

Andrews, Donald W. K. (2005). “Cross-Section Regression with Common Shocks”. In: *Econometrica* 73.5, pp. 1551–1585.

Bai, Jushan (2009). “Panel Data Models with Interactive Fixed Effects”. In: *Econometrica* 77.4, pp. 1229–1279.

Baltagi, Badi H (2013). *Econometric analysis of panel data.* eng. 5th ed. Chichester: Wiley.

Barnett, Karen et al. (2012). “Epidemiology of multimorbidity and implications for health care, research, and medical education: a cross-sectional study”. In: *The Lancet* 380.9836, pp. 37 –43.

Berndt, Nadine et al. (2016). “Economic evaluation of a telephone- and face-to-face-delivered counseling intervention for smoking cessation in patients with coronary heart disease”. In: *The European Journal of Health Economics* 17.3, pp. 269–285.

Bersvendsen, Tore and Jochen Jungeilges (2019). “Creating a unique panel from Norwegian health register data: Technicalities and difficulties”. Manuscript included in Ph.d dissertation.

Bersvendsen, Tore, Jochen Jungeilges, and Eirik Abildsnes (2019). “Empirical evaluation of home-based reablement: A systematic review”. Manuscript submitted for publication.

Bruni, Matteo Lippi, Irene Mammi, and Cristina Ugolini (2016). “Does the extension of primary care practice opening hours reduce the use of emergency services?” In: *Journal of Health Economics* 50, pp. 144 –155.

Cartwright, Nancy (2007). “Are RCTs the Gold Standard?” In: *BioSocieties* 2.1, pp. 11–20.

Cecchini, Michele (2018). “The hidden economics of informal elder-care in the United States”. In: *The Journal of the Economics of Ageing* 12, pp. 218 –224.

Chudik, Alexander and M Pesaran (2013). *Large panel data models with cross-sectional dependence: a survey.* Globalization and Monetary Policy Institute Working Paper 153. Federal Reserve Bank of Dallas.

Chudik, Alexander and M. Hashem Pesaran (2015). “Common correlated effects estimation of heterogeneous dynamic panel data models with weakly exogenous regressors”. In: *Journal of Econometrics* 188.2. Heterogeneity in Panel Data and in Nonparametric Analysis in honor of Professor Cheng Hsiao, pp. 393 –420.

Cochrane, A. et al. (2013). “Home-care 're-ablement' services for maintaining and improving older adults' functional independence”. In: *Cochrane Database of Systematic Reviews* 11.

— (2016). “Time-limited home-care reablement services for maintaining and improving the functional independence of older adults”. In: *Cochrane Database of Systematic Reviews* 10.

Costa-Font, Joan, Sergi Jimenez-Martin, and Cristina Vilaplana (2018). “Does long-term care subsidization reduce hospital admissions and utilization?” In: *Journal of Health Economics* 58, pp. 43 –66.

Deaton, Angus and Nancy Cartwright (2018). "Understanding and misunderstanding randomized controlled trials". In: *Social Science & Medicine* 210. Randomized Controlled Trials and Evidence-based Policy: A Multidisciplinary Dialogue, pp. 2 –21.

Ditzen, J. (2018). "Estimating dynamic common-correlated effects in Stata". In: *Stata Journal* 18.3, 585–617(33).

Driscoll, John C. and Aart C. Kraay (1998). "Consistent Covariance Matrix Estimation with Spatially Dependent Panel Data". In: *The Review of Economics and Statistics* 80.4, pp. 549–560.

Eckardt, Matthias et al. (2017). "Analysis of Health Care Costs in Elderly Patients with Multiple Chronic Conditions Using a Finite Mixture of Generalized Linear Models". In: *Health Economics* 26.5, pp. 582–599.

Engström, Per, Pathric Hägglund, and Per Johansson (2017). "Early Interventions and Disability Insurance: Experience from a Field Experiment". In: *The Economic Journal* 127.600, pp. 363–392.

Everaert, Gerdie and Tom De Groote (2016). "Common Correlated Effects Estimation of Dynamic Panels with Cross-Sectional Dependence." In: *Econometric Reviews* 35.3, pp. 428–463.

Frisch, Ragnar and Frederick V. Waugh (1933). "Partial Time Regressions as Compared with Individual Trends". In: *Econometrica* 1.4, pp. 387–401.

Harris, Richard D.F. and Elias Tzavalis (1999). "Inference for unit roots in dynamic panels where the time dimension is fixed". In: *Journal of Econometrics* 91.2, pp. 201 –226.

Im, Kyung So, M.Hashem Pesaran, and Yongcheol Shin (2003). "Testing for unit roots in heterogeneous panels". In: *Journal of Econometrics* 115.1, pp. 53 –74.

Judson, Ruth A and Ann L Owen (1999). "Estimating dynamic panel data models: a guide for macroeconomists". In: *Economics Letters* 65.1, pp. 9 –15.

Kao, Chihwa (1999). "Spurious regression and residual-based tests for cointegration in panel data". In: *Journal of Econometrics* 90.1, pp. 1 –44.

King, Anna I. I. et al. (2012). "Assessing the impact of a restorative home care service in New Zealand: a cluster randomised controlled trial." In: *Health & Social Care in the Community* 20.4, pp. 365 –374.

Kjerstad, Egil and Hanne Tuntland (2016). "Reablement in community-dwelling older adults: a cost-effectiveness analysis alongside a randomized controlled trial". In: *Health Economics Review* 6.1, pp. 1–10.

Kleinke, Kristian et al. (2011). "Efficient ways to impute incomplete panel data". In: *AStA Advances in Statistical Analysis* 95.4, pp. 351–373.

Langeland, E. et al. (2019). "A multicenter investigation of reablement in Norway: a clinical controlled trial". In: *BMC Geriatrics* 19.1, p. 29.

Leknes, Stefan et al. (2018). *Befolkningsframskrivingene 2018. Modeller, forutsetninger og resultater*. Tech. rep. Rapporter 2018/21. Statistics Norway.

Lewin, Gill and Suzanne Vandermeulen (2010). "A non-randomised controlled trial of the Home Independence Program (HIP): an Australian restorative programme for older home-care clients". In: *Health & Social Care in the Community* 18.1, pp. 91–99.

Lewin, Gill et al. (2013a). “A randomised controlled trial of the Home Independence Program, an Australian restorative home-care programme for older adults”. In: *Health & Social Care in the Community* 21.1, pp. 69–78.

Lewin, Gill et al. (2014). “A comparison of the home-care and healthcare service use and costs of older Australians randomised to receive a restorative or a conventional home-care service”. In: *Health & Social Care in the Community* 22.3, pp. 328–336.

Lewin, Gill F, Helman S Alfonso, and Janine J Alan (2013b). “Evidence for the long term cost effectiveness of home care reablement programs”. eng. In: *Clinical interventions in aging* 8, p. 1273.

Lovell, Michael C. (1963). “Seasonal Adjustment of Economic Time Series and Multiple Regression Analysis”. In: *Journal of the American Statistical Association* 58.304, pp. 993–1010.

Nickell, Stephen (1981). “Biases in Dynamic Models with Fixed Effects”. In: *Econometrica* 49.6, pp. 1417–1426.

Oliveira Martins, Joaquim and Christine Maisonneuve (2007). “The drivers of public expenditure on health and long-term care”. In: *OECD Economic Studies* 2006, pp. 115–154.

Parsons, J.G. et al. (2013). “A randomized controlled trial to determine the effect of a model of restorative home care on physical function and social support among older people”. In: *Archives of Physical Medicine and Rehabilitation* 94 (6).

Pesaran, M. Hashem (2004). *‘General Diagnostic Tests for Cross Section Dependence in Panels’*. Cambridge Working Papers in Economics 0435. Faculty of Economics, University of Cambridge.

— (2006). “Estimation and Inference in Large Heterogeneous Panels with a Multifactor Error Structure”. In: *Econometrica* 74.4, pp. 967–1012.

Pesaran, M. Hashem and Elisa Tosetti (2011). “Large panels with common factors and spatial correlation”. In: *Journal of Econometrics* 161.2, pp. 182–202.

Pesaran, M. Hashem and Takashi Yamagata (2008). “Testing slope homogeneity in large panels”. In: *Journal of Econometrics* 142.1, pp. 50 –93.

Pesaran, M.Hashem and Ron Smith (1995). “Estimating long-run relationships from dynamic heterogeneous panels”. In: *Journal of Econometrics* 68.1, pp. 79–113.

Phillips, Peter C.B. and Donggyu Sul (2007). “Bias in dynamic panel estimation with fixed effects, incidental trends and cross section dependence”. In: *Journal of Econometrics* 137.1, pp. 162–188.

Ringard, A. et al. (2013). “Norway: health system review”. In: *Health systems in transition* 15 (8), pp. 1–162.

Ryburn, Bridget, Yvonne Wells, and Peter Foreman (2009). “Enabling independence: restorative approaches to home care provision for frail older adults”. In: *Health & Social Care in the Community* 17.3, pp. 225–234.

Sabariego, C. et al. (2011). “Cost-effectiveness of cognitive-behavioral group therapy for dysfunctional fear of progression in cancer patients”. In: *The European Journal of Health Economics* 12.5, pp. 489–497.

Sarafidis, Vasilis and Tom Wansbeek (2012). “Cross-Sectional Dependence in Panel Data Analysis.” In: *Econometric Reviews* 31.5, pp. 483–531.

Seidl, Hildegard et al. (2015). "Cost-effectiveness of nurse-based case management versus usual care for elderly patients with myocardial infarction: results from the KORINNA study". In: *The European Journal of Health Economics* 16.6, pp. 671–681.

Senior, Hugh E. J. et al. (2014). "Promoting independence in frail older people: a randomised controlled trial of a restorative care service in New Zealand". In: *Age and Ageing* 43.3, pp. 418–424.

Sopina, Elizaveta et al. (2019). "Long-term medical costs of Alzheimer's disease: matched cohort analysis". In: *The European Journal of Health Economics* 20.3, pp. 333–342.

Tinetti, Mary E. et al. (2002). "Evaluation of Restorative Care vs Usual Care for Older Adults Receiving an Acute Episode of Home Care". In: *JAMA* 287.16, pp. 2098–2105.

Tinetti, Mary E. et al. (2012). "Effect of a Restorative Model of Posthospital Home Care on Hospital Readmissions". In: *Journal of the American Geriatrics Society* 60.8, pp. 1521–1526.

Tuntland, Hanne et al. (2014). "Reablement in community-dwelling adults: study protocol for a randomised controlled trial". In: *BMC Geriatrics* 14.1, p. 139.

Tuntland, Hanne et al. (2015). "Reablement in community-dwelling older adults: a randomised controlled trial". In: *BMC Geriatrics* 15.1, p. 145.

Urwin, Sean, Yiu-Shing Lau, and Thomas Mason (2019). "Investigating the relationship between formal and informal care: An application using panel data for people living together". In: *Health Economics* 28.8, pp. 984–997.

Westerlund, Joakim (2005). "New Simple Tests for Panel Cointegration." In: *Econometric Reviews* 24.3, pp. 297 –316.

Whitehead, Phillip J. et al. (2014). "Occupational Therapy in HomEcare Re-ablement Services (OTHERS): study protocol for a randomized controlled trial". In: *Trials* 15.1, p. 447.

Wiles, Janine L. et al. (2011). "The Meaning of "Aging in Place" to Older People". In: *The Gerontologist* 52.3, pp. 357–366.

Witt Udsen, Flemming et al. (2017). "Subgroup analysis of telehealthcare for patients with chronic obstructive pulmonary disease: the cluster-randomized Danish Telecare North Trial". eng. In: *ClinicoEconomics and outcomes research : CEOR* 9, p. 391.

Wooldridge, Jeffry M. (2002). *Econometric analysis of cross section and panel data.* Cambridge, Mass: MIT press. Chap. 20.

Zweifel, Peter, Friedrich Breyer, and Mathias Kifmann (2009). *Health Economics.* 2nd ed. Springer-Verlag Berling Heidelberg.

IV.A Appendix A: Variable list

Table IV.14: Variable list included in the dataset

	Variable	Comments
Demographical	Gender	Binary variable
	Marital status	Divided into 5 groups
	Education level	Nine levels, including a category ”no or unknown”
	Age group	15 different age groups: 0-39, 40-59, 60-65, 66-70, 71-72, 73-74, 75-76, 77-78, 79-80, 81-82, 83-84, 85-86, 87-88, 89-90 and 91+
	Year/month of death	Year and month the patient died
	Suppressed	Binary variable marking the 180,036 observations that have been suppressed
HBR	HBR total observations	Total number of months in which HBR has been provided during the 5 year period
	HBR total days	Total number of days when HBR has been provided during the 5 year period
	HBR days	HBR days during a specific month
IPLOS	Hours	Hours of home-nurse and homecare
	Days in care institution	Includes long-term and short-term institution
	Other care services	Binary variable indicating whether the patient has used care services other than those given above. A support person is such service
	Function score, 1-17	ADL score 1-5 on 17 different items
KUHR	Other GP services Multidisciplinary cooperation meeting GP home visit GP contact by phone or mail GP office simple contact GP consultation Physiotherapy consultation Physiotherapy home visit Other physiotherapy services	GP and physiotherapist are reimbursed following a set of detailed activities. The dataset contains the aggregated sum of these activities by different categories.

Continued on next page

Table IV.14 – *Variable list - Continued*

	Variable	**Comments**
NPR	Hospital episodes	Number of hospital episodes, both somatic and psychiatry
	Hospital episodes somatic	Number of hospital episodes, somatic only
	Policlinic contacts	Number of policlinic contacts, both somatic and psychiatry
	Policlinic contacts somatic	Number of policlinic contacts, somatic only
	Length of stay hospital	Days of hospital stay, both somatic and psychiatry
	Length of stay hospital somatic	Days of hospital stay, somatic only
	Sector	Five different groups of hospital stay
	Total DRG points	Total DRG points, can be calculated as a cost according to the national unit cost per DRG point
	Most costly diagnosis group	27 different diagnosis groups

IV.B Appendix B:Unit root and co-integration tests

Unit root tests

Table IV.B.1 present p-values for two panel unit root tests, Harris and Tzavalis, 1999 and Im et al., 2003. The null hypothesis is that all panels contain a unit root. As can clearly be seen in Table IV.B.1, one rejected the null and conclude that the variables contain a unit root. This is also as expected because we are working with first-differenced variables.

Table IV.B.1: Unit root tests, p-values presented

	Harris-Tzavalis test		Im-Pesaran-Shin test[a]	
	demean[b]	demean + trend[c]	demean[b]	demean + trend[c]
ΔTotal cost	0.00	0.00	0.00	0.00
ΔPrimary cost	0.00	0.00	0.00	0.00
ΔSpecialist cost	0.00	0.00	0.00	0.00
ΔGP/Phys cost	0.00	0.00	0.00	0.00
ΔHBR Days	0.00	0.00	0.00	0.00
ΔEducation	0.00	0.00	0.00	0.00
ΔFG 1	0.00	0.00	0.00	0.00
ΔFG 2	0.00	0.00	0.00	0.00
ΔFG 3	0.00	0.00	0.00	0.00
ΔFG 4	0.00	0.00	0.00	0.00
ΔFG 5	0.00	0.00	0.00	0.00

[a] *BIC*, Bayesian Information Criterion is used for lag selection
[b] *demean* first subtract the cross-sectional average from the series.
Is used to mitigate the impact of CSD
[c] *demean + trend*, same as above but also includes a linear time trend

Co-integration tests

Several co-integration tests are performed. The results from two tests, Kao, 1999 and Westerlund, 2005, are presented in Table IV.B.2. All tests have the same common null hypothesis of no co-integration. The p-values presented in Table IV.B.2 clearly reject the null hypothesis, and it seems that one are working with a co-integrated panel.

Table IV.B.2: Co-integration test, p-values presented

	Kao test			Westerlund test			
		Lag[a]	lag + demean[b]		demean[c]	trend[d]	demean + trend[e]
p-values	0.00	0.00	0.00	0.00	0.00	0.00	0.00

[a] *BIC*, Bayesian Information Criterion is used for lag selection

[b] the above plus *demean*, first subtract the cross-sectional average from the series. Is used to mitigate the impact of CSD

[c] *demean*, same as above but not including lag adjustment

[d] *trend*, panel-specific linear time trends in the model for dependent variable on the covariates

[e] *demean + trend*, same as above but also subtract the cross-sectional average from the series

IV.C Appendix C: Gender-specific regression results

Male

Table IV.C.1: Males only – Primary care cost regression results – Dynamic CCE, $\tau = 3$

	(1) Δ Primary care cost	(2) Δ Primary care cost	(3) Δ Primary care cost	(4) Δ Primary care cost
$\Delta\bar{T}_{it}$	2938.6***	-1914.9	-1938.7	-1934.1
	(0.000)	(0.126)	(0.122)	(0.123)
Δ Lagged Primary care cost	-0.105***	0.0745	0.0700	0.0738
	(0.000)	(0.342)	(0.372)	(0.346)
Δ HBR days		325.1***	325.7***	325.6***
		(0.000)	(0.000)	(0.000)
$\Delta\bar{T}^{12}$	2215.2***	-1707.6	-1717.0	-1719.8
	(0.000)	(0.157)	(0.155)	(0.154)
Δ Lagged Primary care cost	-0.0905***	0.0972	0.0941	0.0974
	(0.000)	(0.216)	(0.231)	(0.215)
Δ HBR days		307.2***	307.4***	307.6***
		(0.000)	(0.000)	(0.000)
$\Delta\bar{T}^{8}$	1991.8***	-2136.6*	-2171.8*	-2161.0*
	(0.000)	(0.100)	(0.095)	(0.097)
Δ Lagged Primary care cost	-0.112***	0.0807	0.0772	0.0802
	(0.000)	(0.306)	(0.328)	(0.310)
Δ HBR days		322.8***	323.7***	323.5***
		(0.000)	(0.000)	(0.000)
$\Delta\bar{T}^{6}$	1705.8***	-1654.4	-1660.4	-1661.5
	(0.000)	(0.168)	(0.166)	(0.166)
Δ Lagged Primary care cost	-0.0933***	0.0782	0.0744	0.0773
	(0.000)	(0.318)	(0.342)	(0.323)
Δ HBR days		299.9***	300.2***	300.2***
		(0.000)	(0.000)	(0.000)
Included in, s_{it}, cross-sectional averages added:				
Δ Education level		Yes		
Included in, c_{it}, cross-sectional averages not added:				
Δ Marital status 2-5			Yes	
Δ 5 hospital areas				Yes

p-values in parentheses
* p<0.10, ** p<0.05, *** p<0.01

Table IV.C.2: Males only – Specialist health cost regression results – Dynamic CCE, $\tau = 3$

	(1) Δ Specialist health cost	(2) Δ Specialist health cost	(3) Δ Specialist health cost	(4) Δ Specialist health cost
$\Delta \bar{T}_{it}$	-2379.1	-1841.0	-1682.5	-1882.0
	(0.359)	(0.614)	(0.646)	(0.605)
Δ Lagged Specialist care cost	-0.424***	-0.428***	-0.418***	-0.420***
	(0.000)	(0.000)	(0.000)	(0.000)
$\Delta \bar{T}^{12}$	-186.3	695.1	632.9	403.8
	(0.932)	(0.836)	(0.847)	(0.902)
Δ Lagged Specialist care cost	-0.418***	-0.429***	-0.419***	-0.421***
	(0.000)	(0.000)	(0.000)	(0.000)
$\Delta \bar{T}^{8}$	-866.4	-1330.9	-776.6	-1335.7
	(0.681)	(0.616)	(0.769)	(0.614)
Δ Lagged Specialist care cost	-0.425***	-0.433***	-0.426***	-0.427***
	(0.000)	(0.000)	(0.000)	(0.000)
$\Delta \bar{T}^{6}$	-656.2	-1659.6	-1466.2	-1588.9
	(0.728)	(0.399)	(0.456)	(0.419)
Δ Lagged Specialist care cost	-0.421***	-0.439***	-0.431***	-0.431***
	(0.000)	(0.000)	(0.000)	(0.000)
Included in, $\mathbf{s}_{it}$, cross-sectional averages added:				
Δ HBR days		Yes	Yes	Yes
Δ Education level		Yes		
Included in, $\mathbf{c}_{it}$, cross-sectional averages not added:				
Δ Marital status 2-5			Yes	
Δ 5 hospital areas				Yes

p-values in parentheses
* p<0.10, ** p<0.05, *** p<0.01

Female

Table IV.C.3: Females only – Total cost regression results – Dynamic CCE, $\tau = 3$

	(1)	(2)	(3)	(4)
	Δ Total cost	Δ Total cost	Δ Total cost	Δ Total cost
$\Delta \bar{T}_{it}$	-6127.0***	-6314.4***	-6531.4***	-6529.0***
	(0.001)	(0.004)	(0.003)	(0.003)
Δ Lagged Total cost	-0.406***	-0.405***	-0.405***	-0.405***
	(0.000)	(0.000)	(0.000)	(0.000)
R^2	0.75	0.71	0.72	0.72
$\Delta \bar{T}^{12}$	-3341.9**	-3560.6**	-2757.8	-2745.2
	(0.013)	(0.045)	(0.118)	(0.120)
Δ Lagged Total cost	-0.406***	-0.420***	-0.421***	-0.421***
	(0.000)	(0.000)	(0.000)	(0.000)
R^2	0.76	0.71	0.72	0.72
$\Delta \bar{T}^{8}$	-3981.1***	-4214.3***	-4202.1***	-4120.0***
	(0.001)	(0.004)	(0.004)	(0.005)
Δ Lagged Total cost	-0.409***	-0.425***	-0.423***	-0.423***
	(0.000)	(0.000)	(0.000)	(0.000)
R^2	0.76	0.71	0.72	0.72
$\Delta \bar{T}^{6}$	-3308.5***	-3512.4***	-3749.0***	-3681.7***
	(0.002)	(0.007)	(0.004)	(0.004)
Δ Lagged Total cost	-0.413***	-0.424***	-0.423***	-0.423***
	(0.000)	(0.000)	(0.000)	(0.000)
R^2	0.76	0.71	0.72	0.72
Included in, $\mathbf{s}_{it}$, cross-sectional averages added:				
Δ HBR days		Yes	Yes	Yes
Δ Education level		Yes		
Included in, $\mathbf{c}_{it}$, cross-sectional averages not added:				
Δ Marital status 2-5			Yes	
Δ 5 hospital areas				Yes

p-values in parentheses
* p<0.10, ** p<0.05, *** p<0.01

Table IV.C.4: Females only – Primary care cost regression results – Dynamic CCE, $\tau = 3$

	(1) Δ Primary care cost	(2) Δ Primary care cost	(3) Δ Primary care cost	(4) Δ Primary care cost
$\Delta\bar{T}_{it}$	389.9	-1035.1***	-1099.4***	-1099.1***
	(0.223)	(0.001)	(0.001)	(0.001)
Δ Lagged Primary care cost	-0.0514***	0.0387	0.0261	0.0283
	(0.001)	(0.146)	(0.334)	(0.291)
Δ HBR days		106.4***	109.5***	109.5***
		(0.000)	(0.000)	(0.000)
$\Delta\bar{T}^{12}$	94.24	-936.3***	-982.6***	-971.6***
	(0.732)	(0.002)	(0.001)	(0.001)
Δ Lagged Primary care cost	-0.0486***	0.0524	0.0377	0.0401
	(0.001)	(0.174)	(0.309)	(0.278)
Δ HBR days		80.20***	82.19***	82.55***
		(0.009)	(0.006)	(0.006)
$\Delta\bar{T}^{8}$	218.6	-536.1**	-597.2**	-587.7**
	(0.411)	(0.028)	(0.011)	(0.012)
Δ Lagged Primary care cost	-0.0790***	0.000426	-0.00392	-0.00535
	(0.000)	(0.989)	(0.894)	(0.856)
Δ HBR days		79.33***	82.66***	82.94***
		(0.004)	(0.002)	(0.002)
$\Delta\bar{T}^{6}$	20.33	-506.8**	-545.7**	-544.6**
	(0.938)	(0.040)	(0.019)	(0.019)
Δ Lagged Primary care cost	-0.0656***	-0.00676	-0.0106	-0.0116
	(0.000)	(0.703)	(0.537)	(0.499)
Δ HBR days		92.14***	93.93***	94.27***
		(0.000)	(0.000)	(0.000)
Included in, $\mathbf{s}_{it}$, cross-sectional averages added:				
Δ Education level		Yes		
Included in, $\mathbf{c}_{it}$, cross-sectional averages not added:				
Δ Marital status 2-5			Yes	
Δ 5 hospital areas				Yes

p-values in parentheses

* p<0.10, ** p<0.05, *** p<0.01

Table IV.C.5: Females only – Specialist health cost regression results – Dynamic CCE, $\tau = 3$

	(1) Δ Specialist health cost	(2) Δ Specialist health cost	(3) Δ Specialist health cost	(4) Δ Specialist health cost
$\Delta\bar{T}_{it}$	-6776.4***	-6164.9***	-6465.5***	-6449.3***
	(0.000)	(0.004)	(0.003)	(0.003)
Δ Lagged Specialist care cost	-0.434***	-0.437***	-0.437***	-0.436***
	(0.000)	(0.000)	(0.000)	(0.000)
$\Delta\bar{T}^{12}$	-3792.6***	-3647.9**	-2770.2	-2779.1
	(0.003)	(0.035)	(0.107)	(0.106)
Δ Lagged Specialist care cost	-0.434***	-0.449***	-0.450***	-0.450***
	(0.000)	(0.000)	(0.000)	(0.000)
$\Delta\bar{T}^{8}$	-4334.4***	-4055.8***	-3960.9***	-3890.8***
	(0.000)	(0.005)	(0.005)	(0.006)
Δ Lagged Specialist care cost	-0.437***	-0.455***	-0.454***	-0.454***
	(0.000)	(0.000)	(0.000)	(0.000)
$\Delta\bar{T}^{6}$	-3412.4***	-3268.7**	-3402.7**	-3340.7***
	(0.001)	(0.010)	(0.007)	(0.008)
Δ Lagged Specialist care cost	-0.443***	-0.458***	-0.456***	-0.456***
	(0.000)	(0.000)	(0.000)	(0.000)
Included in, $\mathbf{s}_{it}$, cross-sectional averages added:				
Δ HBR days		Yes	Yes	Yes
Δ Education level		Yes		
Included in, $\mathbf{c}_{it}$, cross-sectional averages not added:				
Δ Marital status 2-5			Yes	
Δ 5 hospital areas				Yes

p-values in parentheses
* p<0.10, ** p<0.05, *** p<0.01

IV.D Appendix D: Result tables for different τ values

Total cost

Table IV.D.1: Total cost regression results – Dynamic CCE, $\tau = 4$

	(1)	(2)	(3)	(4)
	Δ Total cost	Δ Total cost	Δ Total cost	Δ Total cost
$\Delta\bar{T}_{it}$	-5227.2***	-5737.4***	-5540.8***	-5752.1***
	(0.001)	(0.005)	(0.005)	(0.004)
Δ Lagged Total cost	-0.402***	-0.409***	-0.409***	-0.409***
	(0.000)	(0.000)	(0.000)	(0.000)
$\Delta\bar{T}^{12}$	-2665.0**	-3135.1*	-2820.5*	-2850.1*
	(0.035)	(0.068)	(0.088)	(0.086)
Δ Lagged Total cost	-0.397***	-0.405***	-0.406***	-0.407***
	(0.000)	(0.000)	(0.000)	(0.000)
$\Delta\bar{T}^{8}$	-3184.4***	-3167.7**	-3221.7**	-3252.1**
	(0.005)	(0.039)	(0.030)	(0.029)
Δ Lagged Total cost	-0.410***	-0.430***	-0.427***	-0.427***
	(0.000)	(0.000)	(0.000)	(0.000)
$\Delta\bar{T}^{6}$	-3189.0***	-3826.5***	-3865.0***	-3880.4***
	(0.001)	(0.003)	(0.002)	(0.002)
Δ Lagged Total cost	-0.408***	-0.423***	-0.420***	-0.420***
	(0.000)	(0.000)	(0.000)	(0.000)
Included in, $\mathbf{s}_{it}$, cross-sectional averages added:				
Δ HBR days		Yes	Yes	Yes
Δ Education level		Yes		
Included in, $\mathbf{c}_{it}$, cross-sectional averages not added:				
Δ Marital status 2-5			Yes	
Δ 5 hospital areas				Yes

p-values in parentheses

* p<0.10, ** p<0.05, *** p<0.01

Table IV.D.2: Total cost regression results – Dynamic CCE, $\tau = 2$

	(1)	(2)	(3)	(4)
	Δ Total cost	Δ Total cost	Δ Total cost	Δ Total cost
$\Delta \bar{T}_{it}$	-4814.8***	-6085.3***	-6123.2***	-6204.3***
	(0.001)	(0.001)	(0.001)	(0.001)
Δ Lagged Total cost	-0.409***	-0.406***	-0.404***	-0.404***
	(0.000)	(0.000)	(0.000)	(0.000)
$\Delta \bar{T}^{12}$	-2321.0**	-3380.7**	-3192.9**	-3201.5**
	(0.040)	(0.016)	(0.023)	(0.023)
Δ Lagged Total cost	-0.406***	-0.409***	-0.408***	-0.409***
	(0.000)	(0.000)	(0.000)	(0.000)
$\Delta \bar{T}^{8}$	-2383.6**	-2872.9**	-2827.9**	-2812.9**
	(0.022)	(0.021)	(0.023)	(0.023)
Δ Lagged Total cost	-0.409***	-0.415***	-0.413***	-0.414***
	(0.000)	(0.000)	(0.000)	(0.000)
$\Delta \bar{T}^{6}$	-2310.9***	-3100.3***	-3119.0***	-3107.3***
	(0.010)	(0.003)	(0.003)	(0.003)
Δ Lagged Total cost	-0.412***	-0.423***	-0.421***	-0.421***
	(0.000)	(0.000)	(0.000)	(0.000)
Included in, $\mathbf{s}_{it}$, cross-sectional averages added:				
Δ HBR days		Yes	Yes	Yes
Δ Education level		Yes		
Included in, $\mathbf{c}_{it}$, cross-sectional averages not added:				
Δ Marital status 2-5			Yes	
Δ 5 hospital areas				Yes

p-values in parentheses
* p<0.10, ** p<0.05, *** p<0.01

Table IV.D.3: Total cost regression results – Dynamic CCE, $\tau = 1$

	(1)	(2)	(3)	(4)
	Δ Total cost	Δ Total cost	Δ Total cost	Δ Total cost
$\Delta\bar{T}_{it}$	-4900.2***	-7124.8***	-7122.6***	-7196.1***
	(0.001)	(0.000)	(0.000)	(0.000)
Δ Lagged Total cost	-0.406***	-0.409***	-0.408***	-0.408***
	(0.000)	(0.000)	(0.000)	(0.000)
$\Delta\bar{T}^{12}$	-2437.4**	-3870.2***	-3793.2***	-3799.5***
	(0.024)	(0.004)	(0.005)	(0.004)
Δ Lagged Total cost	-0.405***	-0.413***	-0.412***	-0.412***
	(0.000)	(0.000)	(0.000)	(0.000)
$\Delta\bar{T}^{8}$	-2618.6***	-3391.9***	-3324.5***	-3308.6***
	(0.009)	(0.005)	(0.006)	(0.006)
Δ Lagged Total cost	-0.410***	-0.417***	-0.416***	-0.416***
	(0.000)	(0.000)	(0.000)	(0.000)
$\Delta\bar{T}^{6}$	-2476.8***	-3334.9***	-3299.9***	-3281.3***
	(0.004)	(0.001)	(0.001)	(0.001)
Δ Lagged Total cost	-0.411***	-0.420***	-0.419***	-0.419***
	(0.000)	(0.000)	(0.000)	(0.000)
Included in, $\mathbf{s}_{it}$, cross-sectional averages added:				
Δ HBR days		Yes	Yes	Yes
Δ Education level		Yes		
Included in, $\mathbf{c}_{it}$, cross-sectional averages not added:				
Δ Marital status 2-5			Yes	
Δ 5 hospital areas				Yes

p-values in parentheses
* p<0.10, ** p<0.05, *** p<0.01

Primary care cost

Table IV.D.4: Primary care cost regression results – Dynamic CCE, $\tau = 4$

	(1) Δ Primary care cost	(2) Δ Primary care cost	(3) Δ Primary care cost	(4) Δ Primary care cost
$\Delta\bar{T}_{it}$	1008.7***	-1026.7***	-1107.9***	-1106.4***
	(0.000)	(0.005)	(0.005)	(0.003)
Δ Lagged Primary care cost	-0.0458***	0.0525*	0.0417	0.0446
	(0.001)	(0.072)	(0.152)	(0.124)
Δ HBR days		153.8***	156.3***	156.3***
		(0.000)	(0.000)	(0.000)
$\Delta\bar{T}^{12}$	684.9***	-867.7**	-961.8***	-957.1***
	(0.006)	(0.015)	(0.006)	(0.006)
Δ Lagged Primary care cost	-0.0434***	0.0624*	0.0565*	0.0595*
	(0.002)	(0.064)	(0.092)	(0.074)
Δ HBR days		127.6***	130.1***	130.5***
		(0.000)	(0.000)	(0.000)
$\Delta\bar{T}^{8}$	611.3***	-840.6**	-907.5**	-904.2**
	(0.007)	(0.019)	(0.011)	(0.011)
Δ Lagged Primary care cost	-0.0777***	0.0367	0.0318	0.0313
	(0.000)	(0.206)	(0.261)	(0.268)
Δ HBR days		135.7***	136.9***	137.2***
		(0.000)	(0.000)	(0.000)
$\Delta\bar{T}^{6}$	423.3*	-659.6**	-684.6**	-685.5**
	(0.057)	(0.043)	(0.034)	(0.033)
Δ Lagged Primary care cost	-0.0598***	0.0366	0.0303	0.0294
	(0.000)	(0.108)	(0.178)	(0.190)
Δ HBR days		138.3***	140.1***	140.2***
		(0.000)	(0.000)	(0.000)
Included in, $\mathbf{s}_{it}$, cross-sectional averages added:				
Δ Education level		Yes		
Included in, $\mathbf{c}_{it}$, cross-sectional averages not added:				
Δ Marital status 2-5			Yes	
Δ 5 hospital areas				Yes

p-values in parentheses
* p<0.10, ** p<0.05, *** p<0.01

Table IV.D.5: Primary care cost regression results – Dynamic CCE, $\tau = 2$

	(1) Δ Primary care cost	(2) Δ Primary care cost	(3) Δ Primary care cost	(4) Δ Primary care cost
$\Delta\bar{T}_{it}$	990.5***	-1187.9***	-1215.8***	-1216.3***
	(0.000)	(0.002)	(0.001)	(0.001)
Δ Lagged Primary care cost	-0.0724***	0.0439	0.0382	0.0405
	(0.000)	(0.131)	(0.184)	(0.157)
Δ HBR days		149.6***	150.5***	150.6***
		(0.000)	(0.000)	(0.000)
$\Delta\bar{T}^{12}$	604.5**	-1037.9***	-1073.1***	-1065.4***
	(0.012)	(0.003)	(0.002)	(0.002)
Δ Lagged Primary care cost	-0.0651***	0.0450	0.0391	0.0424
	(0.000)	(0.168)	(0.225)	(0.186)
Δ HBR days		130.3***	131.4***	131.8***
		(0.000)	(0.000)	(0.000)
$\Delta\bar{T}^{8}$	574.7***	-955.4***	-984.6***	-977.6***
	(0.009)	(0.007)	(0.006)	(0.006)
Δ Lagged Primary care cost	-0.0838***	0.0176	0.0166	0.0162
	(0.000)	(0.538)	(0.557)	(0.568)
Δ HBR days		132.7***	133.3***	133.6***
		(0.000)	(0.000)	(0.000)
$\Delta\bar{T}^{6}$	460.6**	-761.1**	-770.5**	-769.4**
	(0.035)	(0.019)	(0.018)	(0.018)
Δ Lagged Primary care cost	-0.0791***	-0.00372	-0.00510	-0.00567
	(0.000)	(0.864)	(0.814)	(0.793)
Δ HBR days		137.3***	138.2***	138.5***
		(0.000)	(0.000)	(0.000)
Included in, $\mathbf{s}_{it}$, cross-sectional averages added:				
Δ Education level		Yes		
Included in, $\mathbf{c}_{it}$, cross-sectional averages not added:				
Δ Marital status 2-5			Yes	
Δ 5 hospital areas				Yes

p-values in parentheses
* p<0.10, ** p<0.05, *** p<0.01

Table IV.D.6: Primary care cost regression results – Dynamic CCE, $\tau = 1$

	(1) Δ Primary care cost	(2) Δ Primary care cost	(3) Δ Primary care cost	(4) Δ Primary care cost
$\Delta\bar{T}_{it}$	881.1***	-1351.8***	-1357.5***	-1357.9***
	(0.001)	(0.000)	(0.000)	(0.000)
Δ Lagged Primary care cost	-0.0742***	0.0199	0.0190	0.0184
	(0.000)	(0.458)	(0.477)	(0.491)
Δ HBR days		150.5***	150.6***	150.6***
		(0.000)	(0.000)	(0.000)
$\Delta\bar{T}^{12}$	594.7***	-1150.1***	-1171.5***	-1164.9***
	(0.009)	(0.001)	(0.000)	(0.000)
Δ Lagged Primary care cost	-0.0749***	0.0120	0.0112	0.0110
	(0.000)	(0.670)	(0.690)	(0.695)
Δ HBR days		136.4***	137.3***	137.8***
		(0.000)	(0.000)	(0.000)
$\Delta\bar{T}^{8}$	518.7**	-985.3***	-1000.3***	-994.8***
	(0.016)	(0.005)	(0.004)	(0.005)
Δ Lagged Primary care cost	-0.0894***	-0.00310	-0.00244	-0.00364
	(0.000)	(0.905)	(0.925)	(0.889)
Δ HBR days		133.4***	133.8***	134.1***
		(0.000)	(0.000)	(0.000)
$\Delta\bar{T}^{6}$	386.4*	-816.1***	-821.0***	-817.4***
	(0.067)	(0.009)	(0.008)	(0.008)
Δ Lagged Primary care cost	-0.0843***	-0.0158	-0.0154	-0.0163
	(0.000)	(0.455)	(0.467)	(0.441)
Δ HBR days		136.3***	136.5***	136.8***
		(0.000)	(0.000)	(0.000)
Included in, $\mathbf{s}_{it}$, cross-sectional averages added:				
Δ Education level		Yes		
Included in, $\mathbf{c}_{it}$, cross-sectional averages not added:				
Δ Marital status 2-5			Yes	
Δ 5 hospital areas				Yes

p-values in parentheses
* p<0.10, ** p<0.05, *** p<0.01

Specialist health cost

Table IV.D.7: Specialist health cost regression results – Dynamic CCE, $\tau = 4$

	(1)	(2)	(3)	(4)
	Δ Specialist health cost	Δ Specialist health cost	Δ Specialist health cost	Δ Specialist health cost
$\Delta\bar{T}_{it}$	-6242.8***	-5468.9***	-5376.6***	-5567.4**
	(0.000)	(0.006)	(0.006)	(0.005)
Δ Lagged Specialist care cost	-0.428***	-0.440***	-0.440***	-0.440***
	(0.000)	(0.000)	(0.000)	(0.000)
$\Delta\bar{T}^{12}$	-3311.9***	-2993.3*	-2701.9*	-2699.6*
	(0.007)	(0.078)	(0.088)	(0.099)
Δ Lagged Specialist care cost	-0.425***	-0.435***	-0.435***	-0.435***
	(0.000)	(0.000)	(0.000)	(0.000)
$\Delta\bar{T}^{8}$	-3683.4***	-3102.8**	-3121.4**	-3170.2**
	(0.001)	(0.039)	(0.031)	(0.028)
Δ Lagged Specialist care cost	-0.435***	-0.457***	-0.455***	-0.455***
	(0.000)	(0.000)	(0.000)	(0.000)
$\Delta\bar{T}^{6}$	-3455.1***	-3392.7***	-3427.4***	-3459.8***
	(0.000)	(0.007)	(0.006)	(0.005)
Δ Lagged Specialist care cost	-0.436***	-0.451***	-0.449***	-0.449***
	(0.000)	(0.000)	(0.000)	(0.000)
Included in, $\mathbf{s}_{it}$, cross-sectional averages added:				
Δ HBR days		Yes	Yes	Yes
Δ Education level		Yes		
Included in, $\mathbf{c}_{it}$, cross-sectional averages not added:				
Δ Marital status 2-5			Yes	
Δ 5 hospital areas				Yes

p-values in parentheses
* p<0.10, ** p<0.05, *** p<0.01

Table IV.D.8: Specialist health cost regression results – Dynamic CCE, $\tau = 2$

	(1) Δ Specialist health cost	(2) Δ Specialist health cost	(3) Δ Specialist health cost	(4) Δ Specialist health cost
$\Delta\bar{T}_{it}$	-5881.9***	-6123.8***	-6179.9***	-6267.8***
	(0.000)	(0.001)	(0.001)	(0.001)
Δ Lagged Specialist	-0.435***	-0.434***	-0.433***	-0.433***
care cost	(0.000)	(0.000)	(0.000)	(0.000)
$\Delta\bar{T}^{12}$	-3087.4***	-3345.8**	-3125.9**	-3148.5**
	(0.005)	(0.017)	(0.025)	(0.025)
Δ Lagged Specialist	-0.433***	-0.437***	-0.436***	-0.436***
care cost	(0.000)	(0.000)	(0.000)	(0.000)
$\Delta\bar{T}^{8}$	-3009.5***	-2832.7**	-2772.7**	-2775.1**
	(0.003)	(0.022)	(0.025)	(0.022)
Δ Lagged Specialist	-0.436***	-0.443***	-0.441***	-0.441***
care cost	(0.000)	(0.000)	(0.000)	(0.000)
$\Delta\bar{T}^{6}$	-2730.5***	-2908.7***	-2921.2***	-2919.4***
	(0.001)	(0.005)	(0.005)	(0.005)
Δ Lagged Specialist	-0.440***	-0.451***	-0.449***	-0.449***
care cost	(0.000)	(0.000)	(0.000)	(0.000)
Included in, $\mathbf{s}_{it}$, cross-sectional averages added:				
Δ HBR days		Yes	Yes	Yes
Δ Education level		Yes		
Included in, $\mathbf{c}_{it}$, cross-sectional averages not added:				
Δ Marital status 2-5			Yes	
Δ 5 hospital areas				Yes

p-values in parentheses
* $p<0.10$, ** $p<0.05$, *** $p<0.01$

Table IV.D.9: Specialist health cost regression results – Dynamic CCE, $\tau = 1$

	(1) Δ Specialist health cost	(2) Δ Specialist health cost	(3) Δ Specialist health cost	(4) Δ Specialist health cost
$\Delta\bar{T}_{it}$	-5974.4***	-7164.2***	-7137.1***	-7223.5***
	(0.000)	(0.000)	(0.000)	(0.000)
Δ Lagged Specialist	-0.431***	-0.436***	-0.435***	-0.435***
care cost	(0.000)	(0.000)	(0.000)	(0.000)
$\Delta\bar{T}^{12}$	-3189.2***	-3775.1***	-3654.5***	-3679.5***
	(0.002)	(0.004)	(0.006)	(0.005)
Δ Lagged Specialist	-0.431***	-0.439***	-0.438***	-0.438***
care cost	(0.000)	(0.000)	(0.000)	(0.000)
$\Delta\bar{T}^{8}$	-3203.1***	-3255.6***	-3158.8***	-3165.2***
	(0.001)	(0.006)	(0.008)	(0.008)
Δ Lagged Specialist	-0.437***	-0.444***	-0.443***	-0.443***
care cost	(0.000)	(0.000)	(0.000)	(0.000)
$\Delta\bar{T}^{6}$	-2905.2***	-3214.9***	-3165.2***	-3168.2***
	(0.000)	(0.001)	(0.001)	(0.001)
Δ Lagged Specialist	-0.438***	-0.447***	-0.447***	-0.447***
care cost	(0.000)	(0.000)	(0.000)	(0.000)
Included in, $\mathbf{s}_{it}$, cross-sectional averages added:				
Δ HBR days		Yes	Yes	Yes
Δ Education level		Yes		
Included in, $\mathbf{c}_{it}$, cross-sectional averages not added:				
Δ Marital status 2-5			Yes	
Δ 5 hospital areas				Yes

p-values in parentheses
* p<0.10, ** p<0.05, *** p<0.01

IV.E Appendix E: Jackknife bias adjustment

Table IV.E.1: Primary care cost regression results – Dynamic CCE, $\tau = 3$, Jackknife adjusted

	(1) Δ Primary care cost	(2) Δ Primary care cost	(3) Δ Primary care cost	(4) Δ Primary care cost
$\Delta\bar{T}_{it}$	1430.4***	-1084.4	-1382.7*	-1382.7*
	(0.000)	(0.178)	(0.077)	(0.077)
Δ Lagged Primary care cost	-0.0830***	-0.0929	-0.0918	-0.0902
	(0.000)	(0.584)	(0.587)	(0.594)
Δ HBR days		179.2***	184.2***	184.2***
		(0.005)	(0.003)	(0.003)
$\Delta\bar{T}^{12}$	776.0**	-1212.4	-1083.0	-1073.6
	(0.027)	(0.123)	(0.154)	(0.157)
Δ Lagged Primary care cost	-0.0766***	-0.0581	-0.0696	-0.0680
	(0.000)	(0.734)	(0.684)	(0.691)
Δ HBR days		148.3**	152.2**	152.7**
		(0.021)	(0.017)	(0.017)
$\Delta\bar{T}^{8}$	842.0**	-583.2	-814.1	-807.0
	(0.010)	(0.436)	(0.278)	(0.282)
Δ Lagged Primary care cost	-0.118***	-0.134	-0.143	-0.144
	(0.000)	(0.429)	(0.399)	(0.396)
Δ HBR days		149.3**	153.4**	153.6**
		(0.018)	(0.015)	(0.015)
$\Delta\bar{T}^{6}$	598.6*	-714.0	-621.2	-617.2
	(0.062)	(0.332)	(0.394)	(0.397)
Δ Lagged Primary care cost	-0.0973***	-0.128	-0.137	-0.138
	(0.000)	(0.442)	(0.412)	(0.409)
Δ HBR days		169.6***	164.3***	164.6***
		(0.005)	(0.006)	(0.006)
Included in, $\mathbf{s}_{it}$, cross-sectional averages added:				
Δ Education level		Yes		
Included in, $\mathbf{c}_{it}$, cross-sectional averages not added:				
Δ Marital status 2-5			Yes	
Δ 5 hospital areas				Yes

p-values in parentheses
* p<0.10, ** p<0.05, *** p<0.01

Table IV.E.2: Specialist health cost regression results – Dynamic CCE, $\tau = 3$, Jackknife adjusted

	(1) Δ Specialist health cost	(2) Δ Specialist health cost	(3) Δ Specialist health cost	(4) Δ Specialist health cost
$\Delta\bar{T}_{it}$	-7909.1***	-5101.3	-6873.5**	-7041.0***
	(0.000)	(0.105)	(0.012)	(0.010)
Δ Lagged Specialist care cost	-0.443***	-0.421***	-0.438***	-0.439***
	(0.000)	(0.000)	(0.000)	(0.000)
$\Delta\bar{T}^{12}$	-3233.8*	-3944.7*	-2312.8	-2444.2
	(0.050)	(0.091)	(0.291)	(0.265)
Δ Lagged Specialist care cost	-0.450***	-0.452***	-0.460***	-0.460***
	(0.000)	(0.000)	(0.000)	(0.000)
$\Delta\bar{T}^{8}$	-3896.7**	-4414.2**	-3213.6*	-3239.4*
	(0.011)	(0.031)	(0.089)	(0.087)
Δ Lagged Specialist care cost	-0.447***	-0.454***	-0.452***	-0.453***
	(0.000)	(0.000)	(0.000)	(0.000)
$\Delta\bar{T}^{6}$	-3679.6***	-2854.0	-3719.3**	-3758.5**
	(0.006)	(0.114)	(0.027)	(0.025)
Δ Lagged Specialist care cost	-0.443***	-0.456***	-0.452***	-0.452***
	(0.000)	(0.000)	(0.000)	(0.000)
Included in, $\mathbf{s}_{it}$, cross-sectional averages added:				
Δ HBR days		Yes	Yes	Yes
Δ Education level		Yes		
Included in, $\mathbf{c}_{it}$, cross-sectional averages not added:				
Δ Marital status 2-5			Yes	
Δ 5 hospital areas				Yes

p-values in parentheses
* p<0.10, ** p<0.05, *** p<0.01

Table IV.E.3: Total cost regression results – Dynamic CCE, $\tau = 4$, Jackknife adjusted

	(1) Δ Total cost	(2) Δ Total cost	(3) Δ Total cost	(4) Δ Total cost
$\Delta \bar{T}_{it}$	-6972.0***	-6147.9	-6727.3**	-7169.2**
	(0.002)	(0.142)	(0.023)	(0.017)
Δ Lagged Total cost	-0.437***	-0.410***	-0.443***	-0.445***
	(0.000)	(0.000)	(0.000)	(0.000)
$\Delta \bar{T}^{12}$	-2602.7	-4672.8	-2220.4	-2327.6
	(0.155)	(0.136)	(0.378)	(0.357)
Δ Lagged Total cost	-0.434***	-0.409***	-0.455***	-0.456***
	(0.000)	(0.000)	(0.000)	(0.000)
$\Delta \bar{T}^{8}$	-3291.2**	-1812.2	-3347.2	-3378.9
	(0.044)	(0.558)	(0.116)	(0.112)
Δ Lagged Total cost	-0.454***	-0.416***	-0.434***	-0.435***
	(0.000)	(0.000)	(0.000)	(0.000)
$\Delta \bar{T}^{6}$	-3562.7**	-3746.1	-4852.0**	-4815.2**
	(0.014)	(0.159)	(0.010)	(0.011)
Δ Lagged Total cost	-0.431***	-0.481***	-0.443***	-0.443***
	(0.000)	(0.000)	(0.000)	(0.000)
Included in, $\mathbf{s}_{it}$, cross-sectional averages added:				
Δ HBR days		Yes	Yes	Yes
Δ Education level		Yes		
Included in, $\mathbf{c}_{it}$, cross-sectional averages not added:				
Δ Marital status 2-5			Yes	
Δ 5 hospital areas				Yes

p-values in parentheses

* p<0.10, ** p<0.05, *** p<0.01

Table IV.E.4: Primary care cost regression results – Dynamic CCE, $\tau = 4$, Jackknife adjusted

	(1) Δ Primary care cost	(2) Δ Primary care cost	(3) Δ Primary care cost	(4) Δ Primary care cost
$\Delta\bar{T}_{it}$	1486.6***	-1103.9	-1194.6	-1195.3
	(0.000)	(0.293)	(0.112)	(0.121)
Δ Lagged Primary care cost	-0.0675***	-0.0755	-0.0885	-0.0855
	(0.006)	(0.663)	(0.605)	(0.618)
Δ HBR days		172.5***	187.0***	186.9***
		(0.007)	(0.002)	(0.002)
$\Delta\bar{T}^{12}$	933.5***	-761.1	-838.4	-832.5
	(0.009)	(0.361)	(0.266)	(0.269)
Δ Lagged Primary care cost	-0.0729***	-0.0455	-0.0535	-0.0519
	(0.001)	(0.792)	(0.756)	(0.763)
Δ HBR days		125.3*	142.6**	142.9**
		(0.059)	(0.028)	(0.027)
$\Delta\bar{T}^{8}$	818.8**	-345.3	-539.5	-536.1
	(0.014)	(0.655)	(0.461)	(0.464)
Δ Lagged Primary care cost	-0.117***	-0.0806	-0.0887	-0.0896
	(0.000)	(0.641)	(0.604)	(0.601)
Δ HBR days		153.1**	142.8**	143.2**
		(0.016)	(0.022)	(0.021)
$\Delta\bar{T}^{6}$	594.7*	-617.0	-609.6	-601.5
	(0.065)	(0.406)	(0.402)	(0.408)
Δ Lagged Primary care cost	-0.0757***	-0.0808	-0.108	-0.110
	(0.002)	(0.638)	(0.521)	(0.513)
Δ HBR days		167.3***	164.0***	164.2***
		(0.006)	(0.006)	(0.006)
Included in, $\mathbf{s}_{it}$, cross-sectional averages added:				
Δ Education level		Yes		
Included in, $\mathbf{c}_{it}$, cross-sectional averages not added:				
Δ Marital status 2-5			Yes	
Δ 5 hospital areas				Yes

p-values in parentheses
* p<0.10, ** p<0.05, *** p<0.01

Table IV.E.5: Specialist care cost regression results – Dynamic CCE, $\tau = 4$, Jackknife adjusted

	(1) Δ Specialist health cost	(2) Δ Specialist health cost	(3) Δ Specialist health cost	(4) Δ Specialist health cost
$\Delta\bar{T}_{it}$	-8512.2***	-7658.9*	-7027.7**	-7495.4**
	(0.000)	(0.059)	(0.016)	(0.011)
Δ Lagged Specialist care cost	-0.464***	-0.434***	-0.483***	-0.484***
	(0.000)	(0.000)	(0.000)	(0.000)
$\Delta\bar{T}^{12}$	-3571.7**	-4180.5	-2342.3	-2435.2
	(0.046)	(0.165)	(0.347)	(0.330)
Δ Lagged Specialist care cost	-0.464***	-0.429***	-0.480***	-0.481***
	(0.000)	(0.000)	(0.000)	(0.000)
$\Delta\bar{T}^{8}$	-3994.5**	-3645.4	-3251.2	-3359.7
	(0.012)	(0.189)	(0.118)	(0.105)
Δ Lagged Specialist care cost	-0.466***	-0.426***	-0.457***	-0.457***
	(0.000)	(0.000)	(0.000)	(0.000)
$\Delta\bar{T}^{6}$	-3912.9***	-2437.3	-4392.6**	-4397.4**
	(0.005)	(0.329)	(0.017)	(0.017)
Δ Lagged Specialist care cost	-0.457***	-0.451***	-0.471***	-0.471***
	(0.000)	(0.000)	(0.000)	(0.000)
Included in, $\mathbf{s}_{it}$, cross-sectional averages added:				
Δ HBR days		Yes	Yes	Yes
Δ Education level		Yes		
Included in, $\mathbf{c}_{it}$, cross-sectional averages not added:				
Δ Marital status 2-5			Yes	
Δ 5 hospital areas				Yes

p-values in parentheses
* p<0.10, ** p<0.05, *** p<0.01

IV.F Appendix F: Increased price per hour

Table IV.F.1: Total cost regression results – Dynamic CCE – Increased home-nurse price, $\tau = 3$

	(1)	(2)	(3)	(4)
	Δ Total cost	Δ Total cost	Δ Total cost	Δ Total cost
$\Delta\bar{T}_{it}$	-4510.7***	-5772.2***	-5781.8***	-5916.4***
	(0.003)	(0.002)	(0.002)	(0.002)
Δ Lagged Total cost	-0.404***	-0.401***	-0.400***	-0.400***
	(0.000)	(0.000)	(0.000)	(0.000)
$\Delta\bar{T}^{12}$	-2163.4*	-2904.2*	-2701.0*	-2739.0*
	(0.065)	(0.055)	(0.071)	(0.067)
Δ Lagged Total cost	-0.404***	-0.410***	-0.408***	-0.409***
	(0.000)	(0.000)	(0.000)	(0.000)
$\Delta\bar{T}^{8}$	-2887.2***	-3539.0***	-3522.0***	-3492.7***
	(0.009)	(0.009)	(0.009)	(0.010)
Δ Lagged Total cost	-0.408***	-0.418***	-0.415***	-0.415***
	(0.000)	(0.000)	(0.000)	(0.000)
$\Delta\bar{T}^{6}$	-2573.1***	-3538.1***	-3585.2***	-3592.6***
	(0.006)	(0.003)	(0.002)	(0.002)
Δ Lagged Total cost	-0.409***	-0.421***	-0.418***	-0.419***
	(0.000)	(0.000)	(0.000)	(0.000)
Included in, $\mathbf{s}_{it}$, cross-sectional averages added:				
Δ HBR days		Yes	Yes	Yes
Δ Education level		Yes		
Included in, $\mathbf{c}_{it}$, cross-sectional averages not added:				
Δ Marital status 2-5			Yes	
Δ 5 hospital areas				Yes

p-values in parentheses
* p<0.10, ** p<0.05, *** p<0.01

Table IV.F.2: Primary care cost regression results – Dynamic CCE – Increased home-nurse price, $\tau = 3$

	(1) Δ Primary care cost	(2) Δ Primary care cost	(3) Δ Primary care cost	(4) Δ Primary care cost
$\Delta \bar{T}_{it}$	1341.4***	-1329.8***	-1362.3***	-1363.9***
	(0.000)	(0.002)	(0.001)	(0.001)
Δ Lagged Primary care cost	-0.0557***	0.0577*	0.0495*	0.0518*
	(0.000)	(0.055)	(0.097)	(0.082)
Δ HBR days		185.5***	186.7***	186.8***
		(0.000)	(0.000)	(0.000)
$\Delta \bar{T}^{12}$	862.0***	-1104.7***	-1145.1***	-1134.6***
	(0.001)	(0.006)	(0.004)	(0.005)
Δ Lagged Primary care cost	-0.0542***	0.0641*	0.0563*	0.0590*
	(0.000)	(0.061)	(0.099)	(0.083)
Δ HBR days		162.2***	163.4***	163.8***
		(0.000)	(0.000)	(0.000)
$\Delta \bar{T}^{8}$	889.1***	-981.4**	-1035.8**	-1027.1**
	(0.000)	(0.016)	(0.011)	(0.012)
Δ Lagged Primary care cost	-0.0817***	0.0235	0.0216	0.0211
	(0.000)	(0.444)	(0.476)	(0.486)
Δ HBR days		166.5***	168.1***	168.3***
		(0.000)	(0.000)	(0.000)
$\Delta \bar{T}^{6}$	654.5***	-782.5**	-806.4**	-808.7**
	(0.005)	(0.036)	(0.029)	(0.028)
Δ Lagged Primary care cost	-0.0685***	0.0130	0.00992	0.00946
	(0.000)	(0.561)	(0.654)	(0.668)
Δ HBR days		170.5***	171.3***	171.5***
		(0.000)	(0.000)	(0.000)
Included in, s_{it}, cross-sectional averages added:				
Δ Education level		Yes		
Included in, c_{it}, cross-sectional averages not added:				
Δ Marital status 2-5			Yes	
Δ 5 hospital areas				Yes

p-values in parentheses

* p<0.10, ** p<0.05, *** p<0.01