Pimperl, A; Schulte, T; Mhlbacher, A; Rosenmller, M; Busse, R; Groene, O; Rodriguez, HP; Hildebrandt, H (2016) Evaluating the Impact of an Accountable Care Organization on Population Health: the Quasi-Experimental Design of the German Gesundes Kinzigtal. Population health management. ISSN 1942-7891 DOI: https://doi.org/10.1089/pop.2016.0036

Evaluating the Impact of an Accountable Care Organization on Population Health – the Quasi-Experimental Design of the German "Gesundes Kinzigtal"

Alexander Pimperl, Dr. phil.[1]*, Timo Schulte, MBA, Dipl.-Kfm.[2,3], Axel Mühlbacher, Dr. rer. oec., Dipl.-Kfm.[4], Magdalena Rosenmöller, Ph.D, M.D., MBA[5], Reinhard Busse[6], Dr. med, MPH, FFPH, Oliver Groene, PhD, MSc, MA[3,7], Hector P. Rodriguez, PhD, MPH[1], Helmut Hildebrandt[3,8]

* Corresponding author: Alexander Pimperl; apimperl@berkeley.edu

[1] University of California, Berkeley, School of Public Health - Health Policy and Management, Berkeley, CA, United States

[2] Department of Health, University of Witten/Herdecke, Witten, Germany

[3] OptiMedis AG, Hamburg, Germany

[4] Institute for Health Economics and Health Care Management, Hochschule Neubrandenburg, Neubrandenburg, Germany

[5] Centre for Research in Health Innovation Management, IESE Business School, Barcelona, Spain

[6] Department of Health Care Management, TU Berlin, Berlin, Germany

[7] Department of Health Services Research and Policy, London School of Hygiene and Tropical Medicine, London, United Kingdom

[8] Gesundes Kinzigtal GmbH, Haslach, Germany

Full author contact information

Alexander Pimperl
University of California, Berkeley
School of Public Health
50 University Hall, #7360
Berkeley, CA 94720, United States
Email: apimperl@berkeley.edu
Phone: +1 510 679 8630

Timo Schulte
OptiMedis AG
Borsteler Chaussee 53
22453 Hamburg, Germany
Email: t.schulte@optimedis.de
Phone: +49 40 22621149-0

Axel Mühlbacher
Institute for Health Economics and Health Care Management, Hochschule Neubrandenburg,
Postfach 11 01 21
17041Neubrandenburg, Germany
Email: muehlbacher@me.com
Phone: +49 395 5693 3301

Magdalena Rosenmöller
Centre for Research in Health Innovation Management, IESE Business School,
Avda Pearson, 21
08034 Barcelona, Spain
Email: MRosenmoller@iese.edu
Phone: +34 932534200

Reinhard Busse
Department of Health Care Management, TU Berlin,
Sekretariat H80, Raum H8112
Straße des 17. Juni 135
10623 Berlin, Germany
Email: rbusse@tu-berlin.de
Phone: +49 30 314 28420

Oliver Groene
London School of Hygiene and Tropical Medicine
15-17 Tavistock Place
London, WC1 9SH, United Kingdom
Email: oliver.groene@lshtm.ac.uk
Phone: +49 40 22621149-0

Hector P. Rodriguez
University of California, Berkeley
School of Public Health
50 University Hall, #7360
Berkeley, CA 94720, United States
Email: hrod@berkeley.edu
Phone: +1 510 642 4578

Helmut Hildebrandt
OptiMedis AG
Borsteler Chaussee 53
22453 Hamburg, Germany
Email: h.hildebrandt@optimedis.de
Phone: +49 40 22621149-0

Conflict of interest

Mr. Schulte, Dr. Groene and Mr. Hildebrandt are employees of the OptiMedis AG. Dr. Pimperl was also employed by the OptiMedis AG, but is currently a Harkness Fellow in Health Care Policy and Practice and does not receive any funding from the OptiMedis AG. The OptiMedis AG is the management partner and shareholder of the Gesundes Kinzigtal GmbH, which is used as study setting for the empirical application of the evaluation design described in this paper. Mr. Hildebrandt is also the CEO of the Gesundes Kinzigtal GmbH. There are no conflicts of interest for the other co-authors.

Funding

At the time this manuscript was developed, Dr. Pimperl was engaged in the Harkness Fellowship in Health Care Policy and Practice, supported by The Commonwealth Fund, a private independent foundation based in New York City. The views presented here are those of the authors and not necessarily those of The Commonwealth Fund, its directors, officers, or staff.

Abstract:

A central goal of accountable care organizations (ACOs) is to improve the health of their accountable population. No evidence currently links ACO development to improved population health. A major challenge to establishing the evidence-base for the impact of ACOs on population health is the absence of a theoretically-grounded, robust, operationally feasible, and meaningful research design. We present an evaluation study design, provide an empirical example, and discuss considerations for generating the evidence-base for ACO implementation.

A quasi-experimental study design using propensity score matching in combination with small-scale exact-matching is implemented. Outcome indicators based on claims data were constructed and analyzed. Population health is measured by using a range of mortality indicators: mortality ratio, age at the time of death, years of potential life lost / gained and survival time. The application is assessed using longitudinal data from Gesundes Kinzigtal, one of the leading population-based ACOs in Germany.

The proposed matching approach resulted in a balanced control of observable differences between the intervention (ACO) and control group. The mortality indicators used, indicate positive results. For example, 635.6 fewer *years of potential life lost* (2,005.8 vs. 2,641.4; T-Test: sig. $p<0.05$*) in the ACO-intervention group (n=5,411), attributable to the ACO, also after controlling for a potential (indirect) immortal time bias by excluding the first half year after enrollment from the outcome measurement.

Our empirical example of the impact of a German ACO on population health can be extended to the evaluation of ACOs and other integrated delivery models of care.

Keywords: program evaluation, integrated delivery systems, Accountable Care Organizations, managed care organizations, evaluation design, health services research, quasi-experiments, administrative data uses, impact assessment, evaluation framework

1 Background

Health care systems are aiming to achieve the '*Triple Aim'*: improving population health, patient experience and cost efficiency. The architects of the Triple Aim[1] highlight that it is achieved by an '*integrator*', who organizes a close collaboration between all actors, such as care providers, professionals or community institutions. An Accountable Care Organization (ACO) can play a central, facilitating role in moving providers and systems towards the Triple Aim. The Centers for Medicare and Medicaid Services (CMS) defines ACOs as "[…] groups of doctors, hospitals, and other health care providers, who come together voluntarily to give coordinated high quality care to the Medicare patients they serve […] When an ACO succeeds in both delivering high-quality care and spending health care dollars more wisely, it will share in the savings it achieves for the Medicare program."[2]. Integrated, accountable care initiatives have been introduced in other countries, such as the UK or the Netherlands, and have moved up the political agenda[3]. In Germany Gesundes Kinzigtal (GKT), one of the country's leading population-based ACOs, has taken on the '*integrator*' role and strives to achieve the Triple Aim[4].

In order to examine the impact of Triple Aim initiatives, the Institute of Healthcare Improvement (IHI) proposes a series of outcome indicators[1,5]. However, limited information is provided on the evaluation design and how causal inference can be enhanced. In practice, a wide variety of approaches may be found[6–8]. Claims data have been used extensively to assess quality and safety of care so far[9] but challenges of their usability in measuring Triple Aim outcomes requires further studies.

In this paper, we aim to provide guidance on a robust evaluation design to measure the impact of ACOs on Triple Aim outcomes by using claims data. We apply the evaluation design to the GKT ACO. Our focus will be on population health indicators, corresponding to the first Triple Aim dimension; the other two Triple Aim dimensions are addressed elsewhere[13,21]. With population health, we mean the health of the population of insurees attributed to the ACO by contract.

Specifically, we aim to:

- Identify an appropriate study design for evaluating population health outcomes of ACOs on the basis of claims data
- Discuss methodological implications and the feasibility of the approach to evaluate ACOs using routine data from the ACO GKT
- Evaluate the impact of the ACO GKT on population health
- Provide guidance to future evaluations of ACO impacts on population health using routine data sources

2 Methods

2.1 Study setting for the empirical application

The ACO GKT is located in a rural area in southwest Germany. Its central entity is the Gesundes Kinzigtal GmbH – created in 2006 and jointly owned by a long-established local physician network and the healthcare management company OptiMedis. A long term (10 years) shared saving contract with two statutory health insurers (SHIs) – the AOK Baden-Württemberg (AOK BW) and the LKK Baden-Württemberg (LKK BW) – ensured financial stability that allowed for long-term planning and implementation of population health interventions. The scheme covers about half of the region's population, corresponding to 32,595 insurees. The GKT concept is based on the cross-sectorial cooperation of physicians, hospitals, social care, nursing staff, therapists and pharmacies, the involvement of all stakeholders in the community and the encouragement of patients to actively participate in prevention and care. The patients' free choice of health care providers remains unrestricted. Patients may seek care services from any legally accredited provider, regardless of whether the provider (e.g. a general practitioner) does or does not have a contract with GKT[10]. The range of GKT activities includes, besides general care management, a

set of community initiatives; specific financial incentives for cooperating providers and about 20 preventive and health promotion programs for specific conditions, as described in detail elsewhere[10].

The GKT focus on population health management towards the Triple Aim is realized through a data driven approach, utilizing internal monitoring[8,11,12] and external evaluations[13,14].

2.2 Data source

We based the study essentially on claims data, as it has been shown that they are valuable to assess quality and safety of care[9] and they have the advantage of being easily and widely accessible in electronic form without the need for additional documentation[5], an important factor in times of excessive external performance reporting requirements[15].

GKT obtains de-identified insuree level master data of the 32,595 insurees (AOK BW = 31,101 in July 2014, LKK BW n=1,494 in January 2014) concerned by the shared savings contract, and associated data on out-patient care, hospital stays, pre- and post-stationary services, outpatient surgery, work incapacities, drugs or non-medicinal remedies and aids, prevention, rehabilitation and long-term care services from the two cooperating SHIs. From this group, 9,568 (AOK BW= 9,130 and LKK BW=438 in October 2014) are enrolled in the ACO (ACO-enrollees, see Figure 1). Enrollment is voluntary and allows for special offers by the ACO and the participation in one or more of the 20 special GKT health programs, where indicated or certain requirements are fulfilled.

In this study, we limited the study group to the ACO-enrollees actively enrolled in the years 2006 to 2009 (n=6,922). 2006 was the first year of enrollment at GKT and 2009 was the last year we included, to ensure four years of after intervention time points for each ACO-enrollee. Overall data for the years 2005 (baseline for 2006 ACO-enrollees) to 2013 (fourth intervention year for 2009 ACO-enrollees) were utilized in this study.

2.3 Study design

2.3.1 Potential evaluation designs for assessing ACO impact

The large number of possible evaluation designs can be divided into three categories: *experimental*, *quasi-experimental* and *non-experimental*. *Experimental* designs typically involve randomization, manipulation of independent variable(s) and control, and are usually longitudinal and prospective. *Quasi-experimental* designs also include manipulation but lack the full control established in experimental designs, as a separate control group might be missing, or not assigned by randomization. Designs not fitting into these two classes can be considered as *non-experimental*; they do not involve manipulation, randomization or control groups[16].

The suitability of an evaluation design is mostly dependent on the research question and the context of application. Non-experimental designs are most appropriate for exploratory, descriptive or correlational research questions[16]. This study aims to analyze the cause-effect relationships of ACO intervention and its impact on population health, in a counterfactual approach to causality[17]; *non-experimental* designs are inadequate and can be excluded from the present discussion.

Increased confidence of cause-and-effect relationships requires *true* or *quasi-experimental* designs[18,19]. True experimental designs, also referred to as *Randomized Controlled Trials (RCT)*, are regarded as the gold standard for evaluating healthcare interventions[18]. *Individually-randomized parallel group designs (i-RCTs)* are often not feasible in population level interventions because of the risk of contamination; i.e. it is practically impossible for an ACO to train physicians (e.g. in shared-decision making) and then limit the application of the new knowledge and capacity to a selected group of patients for the same medical office. As a result, the control group will be impacted by the experience of the physician or his/her organization[18,19].

In case of a high likelihood of such contamination, *Cluster-Randomized Controlled Trials (c-RCTs)* could be an option to separate intervention and control group, where randomization is realized on the level of organizations or individual physicians[18,19]. However, c-RCTs are much more complex, and the number of cases must be much higher to account for the design factor resulting from the assumed intra-cluster correlation. Also, in practice it is difficult to recruit enough suitable organizations for comparison[18]. Other options for randomization exist, such as *stepped wedge designs, preference trials, randomized consent designs* or *N-of-1 designs*[19]. In addition to these

options - all variants of the traditional RCT that seeks to maximize internal validity - pragmatic trials put stronger emphasis on the external validity (generalizability) of results. To achieve this, interventions are conducted under conditions that resemble normal practice (moving away from ideal settings and highly selected participants) and that allow leeway in implementing the intervention, rather than strictly enforcing a study protocol[20].

Despite this multitude of RCT designs, the majority are not adequate enough to evaluate the impact of ACOs, the main reason being the irreversibility of the intervention, its application to the entire population at the same time and not in different steps. In addition, usually evaluators have no access to sufficient numbers of suitable organizations or regions for the purpose of comparison. Ethical reasons or costs of RCTs can also be an obstacle[18,19]. ACOs aim to improve efficiency in comparison to standard care and are not designed as research projects. ACOs with a shared-savings contract, such as GKT, must be economically viable for providers and the sickness fund. Here, complex RCTs are not always the preferred option[21,22].

Quasi-experimental designs share the purpose of true experiments – testing the cause-effect relationships, but because they lack full experimental control and randomization, they are more vulnerable to threats of internal validity[16,18,19]. A multitude of quasi-experimental designs try to control these biases, and – while discussing them in detail goes beyond the scope of this paper – a few design characteristics might be worth noting, as they allow a better understanding of our selection of the evaluation design, and the possibilities of reducing biases (as can be seen in the scheme by Shadish, Cook, and Campbell[16]): *designs without control groups*, *designs without pretest indicators*, and *combination designs*[23].

Designs without control groups: A post-test-only design (observation only after the intervention) can be improved by pre-test indicators if no control group is available[23]. Such a design with one measurement point before and after an intervention implemented in the same study site(s) is usually called an *uncontrolled before and after (UBA)* study. In *interrupted time series (ITS),* additional pre- and post-test indicators are added. Although ITS, combined with appropriate statistical methods (time series regression models, autoregressive integrated moving averages modelling), reduce the threat of statistical regression or misinterpretation of underlying secular trends or cyclical effects, they do not provide protection against distortive external effects[18,23], such as systematic, inexorable advances in (medical) technology or changes in the external provider structure.

Designs without pretest indicators: On the other hand, studies where ACOs lack data for pretest indicators could improve post-test-only designs through a control group. In such a situation, a *nonequivalent control group* might be used. However, these designs are prone to a selection bias, i.e. subjects in the intervention group differ from those in the control group in regard to morbidity, social status, etc.[23,24].

Combination designs: Whenever possible, a simultaneous use of pre-test indicators and control groups is desirable. A study, where data are collected at one time point before and after the intervention, is referred to as a *controlled before and after (CBA)* study. Multiple data collection points are preferable[18,23] (*control group interrupted time series*). The GKT claims database allows for such a design. Nonetheless, these combined designs are also susceptible to a selection bias, because they are not based on true experimental (randomized) data. Researchers rely on sophisticated statistical methods of bias control in quasi-experimental studies[24]. Commonly used are '*matched pair*' approaches (e.g. exact matching, propensity score matching, genetic matching), where possible distortions of a selection bias are minimized by taking into consideration observable risk factors. Insurees of the intervention group are compared to a control group with "statistical twins" of the same age, gender, morbidity, costs, social status etc.[25].

For our study at GKT, we also adopted a *matched pair* approach; the empirical application will be described in more detail in the following section.

2.3.2 Evaluation design chosen for empirical application in GKT

For the GKT population health study, we chose a matched pair approach, in the light of the practical, economical and ethical shortcomings of RCTs. Existing claims data allowed for multiple before- and after-intervention time points, and also the integration of a control group. For the purpose of this study, we limited the study group to the ACO-enrollees, actively enrolled in the years 2006 to 2009 (n=6,922). The "untreated" matched pairs (Non-ACO-control group) are drawn

from the AOK and LKK BW insurees also living in the region of Kinzigtal but not enrolled in the ACO and not primarily treated (less than 49% of their physician cases) by GKT contractual primary care physician network members (Non-enrolled insurees); cf. Figure 1. Nonetheless, Non-ACO-control group subjects might also access some ACO interventions, such as specialist treatment, seminars for patients on literacy and patient empowerment, occupational health management, and additional physician activities offered in sports clubs and gyms, a bias that needs to be considered.

Due to the limited available data set from which “statistical twins” may be drawn, a propensity score matching (PSM) approach was preferred over an exact matching approach. In the case of a limited data set, exact matching approaches, otherwise often the better option[26], can lead to an exclusion of a relevant number of cases in the matching process and/or to the necessity to abandon covariates. The potential resulting bias can cause a more important distortion than less exact matched pairs (by PSM) but a more complete set of insurees[26,27]. But exact matching and PSM can be used effectively in combination[26], therefore in the GKT study we combine the PSM with small-scale exact matching, thus a better matching can be achieved with an acceptable level of exclusion of cases, as already seen in an earlier GKT study[12].

The PSM is based on a logistic regression which estimates the conditional probability (propensity score) of an insuree to be an ACO-enrollee on a scale from 0 to 1 using multiple predictors (see Table 1) from a base year (the year preceding ACO enrollment). The calculated propensity score is used to find adequate pairs in a nearest neighbor approach. A maximal difference of ± 0.01 (approx. 0.2 standard deviations) in caliper is thereby tolerated between the propensity score of the insuree in the ACO-intervention group and its "statistical twin" in the Non-ACO-control group[27,28].

Following the recommendation of Stuart[26], the PSM is also combined with a small-scale exact matching, including the (socio-)demographic covariates of age, sex and the person's SHI and insurance status; and the morbidity-related covariate Charlson-score[29] (see Table 1). As the LKK BW insured represent a special community including mainly farmers (active insured or pensioner) and their family members, we considered it appropriate to do the matching only within the same SHI. Also, the social gradient of health[30] has been taken into account in the exact matching by the person’s insurance status. The insurance status provides certain information about income, considered one of the best predictors for socio-economic status related to health[30]. For the matching, four classes were formed: compulsorily-insured persons, voluntarily insured persons, unemployed, and others. Whereas the salaries of voluntarily insured persons usually exceed the annual contribution income threshold (53,550 Euro p.a. in the year 2014), compulsorily-insured persons must have earnings below this threshold, and unemployed are a special income group.

Because of the *low-prevalence* criteria rare diseases may be excluded from the PSM. As they are often associated with high morbidity and potential mortality, we used the Charlson-score in the exact matching approach to adjust for that. A maximum difference of ± 1 is allowed in the matching.

We used a 1:1 matching, meaning that one ACO-intervention group subject is matched to one Non-ACO-control group subject(s). In our limited data set, this will lead to an improved bias reduction. The matching is done without replacement as ‘greedy matching’: the first nearest Non-ACO “statistical twin” is selected, even though this twin might be a better match for another ACO subject. In an *optimal matching approach* this would be taken into account[28]. However, since optimal matching does not produce better balanced matched samples than greedy matching[31], and greedy matching can be realized in an easier and faster way, we decided to use greedy matching. To avoid the immortal time bias[32], only Non-ACO subjects who are still alive at the time of the enrollment of the ACO subjects are matched. Additionally, we applied data quality criteria to prevent a data quality bias (see Table 1). These criteria resulted in a loss of 501 ACO-enrollees (7.2%). Additional 1,010 (14.6%) ACO-enrollees had to be excluded from the analysis, because no adequate matched pair could be found, resulting in an analytic sample of 5,411 insurees in both the ACO-intervention group and the control group (see appendix).

To assess the quality of the matching, we have compared the categorical variables between groups of patients concerned with diagnosis or medication pre- and post-matching[33].

For metric variables, it is recommended that arithmetic mean and standard deviation pre and post matching are compared, and standardized differences between the groups are indicated. The

standardized difference can serve as an indicator for the matching balance: the lower the value, the better the balance[33,34], where ±10 is assumed to be good balance[34].

To estimate the impact of the ACO GKT on population health, we constructed and analyzed outcome indicators derivable from claims data.

2.4 Outcome indicators

Based on the IHI recommendations[5] we have adapted the following outcome indicators for the population health dimension for our study, taking into account restrictions related to the use of claims data.

Mortality ratio (observed number of deaths / total of subjects in the studied population) is widely recognized as a simple, manipulation-resistant surrogate parameter for outcome quality and patient benefit[5,35]. With a fixed cohort study design (no new subjects to enter the study), an increase of the mortality ratio over time should be anticipated: members of the group are to die anyway at some point in time, and the intervention can only postpone the time of death. This issue is better dealt with by other mortality indicators as they either compare the age of death to the expected time (life expectancy/LE) and deduct the potential lost lives or look directly at survival time.

Age at the time of death (statistically expected number of years of life in the studied population) is used to predict LE[5].

Years of potential life lost and gained (YPLLG) is an adapted individually age-adjusted YPLL indicator. YPLL measures potential life lost due to premature death, i.e. a person with a mean life expectancy of 75 years dying aged 65 represents 10 years lost[5,36]. For the YPLLG indicator LE is calculated for individuals using the generations' LE tables of the German Federal Statistical Office[37], accounted also for potential life years gained. A person with an individual LE of 98 dying at the age of 100 will contribute 2 years gained, a fact not reflected in the YPLL. Thus, this adapted YPLLG indicator aims to improve accuracy.

Survival time (time between the start of the observation (enrollment in the ACO) and the end of the study period or an event (death)) estimates the probability of a study insuree's survival in a given time interval, measured by the Kaplan-Meier-method[38].

A few considerations have to be made regarding the ***mortality indicators***. As death is an irreversible end point, logically no pre-test data can be collected (although it might be useful to have other pre-test data for the matching approach and analysis). The status of the subject as being alive at the time of recruitment[23] is taken into consideration for Non-ACO subjects in the control group to avoid an immortal time bias[32], as only Non-ACO subjects still alive at the time of enrollment of the ACO subject may be matched[32]. In addition, an indirect immortal time bias may occur: this is the case for GKT physicians usually deciding against enrollment of terminally ill patients in the ACO, as it represents additional stress and little benefit for these patients[12], leading to the fact that patients with a high risk of imminent death are present in the control group rather than the intervention group. To control for this indirect immortal time bias it is recommended to exclude the first half year after the start of the intervention. We allowed for this bias by differentiating the results of the first year after intervention in the analysis.

3 Results: Empirical application: Gesundes Kinzigtal

3.1 Effectiveness of the matching approach

To assess the quality of the matching, we have compared the categorical variables between groups of patients concerned with diagnosis or medication. For the top 40 ICDs the inter-group differences could be reduced from a maximum of 31.3% pre-matching to a maximum of 2.1% post-matching.

For metric variables we compared arithmetic mean and standard deviation pre and post matching, and standardized differences between the groups. The before and after matching comparison confirms the requested equalization between the ACO-intervention group and the Non-ACO-control group in all variables and all enrollment years. No standardized difference post-matching is bigger than ±10. The maximum standardized difference post-matching is -8.3 for "*insured person days*" in the 2009 cohort. Table 2 and Table 3 show the results exemplary in detail for the year 2006.

Tables for all years and further analysis can be found at Schulte et al.[8].

3.2 Impact of the ACO GKT on population health

Table 4 summarizes the differences of the population health outcome indicators *mortality ratio*, *age at the time of death* and *YPLLG* for the ACO-intervention group vs. the Non-ACO-control group. Mortality rates are lower in year one to three, but higher in the year four in the ACO-intervention group. In total over the four years 222 insurees die in the ACO-intervention group (4.1%) and 266 in the Non-ACO-control group (4.9%), thus 44 insurees less die in the ACO-intervention group. Differences are not significant (chi-square: $p>0.05$). Also after exclusion of the first six months after enrolment (an adjustment to avoid an indirect immortal time bias) the results stay similar (33 less deaths).

Because of the shortcomings of the mortality ratio indicator discussed above we will further concentrate on the other proposed outcome indicators. The average age at the time of death is 1.4 years higher in the ACO-control group (78.9 vs. 77.5). Over the considered period of four years (excluding the first two quarters to avoid an indirect immortal time bias) the YPLLG-indicator showed 635.6 fewer *years of potential life lost* in the ACO-intervention group (2,005.8 vs. 2,641.4 years of potential life lost; T-Test: sig. $p<0.05$*).

Figure 2 shows that the survival time estimated by using the Kaplan-Meier-method is 6.7 days higher in the ACO-Intervention group (1,433.8; 95% CI: 1,430.3 – 1,437.3) than in the Non-ACO-control group (1,427.1; 95% CI: 1,423.0 – 1,431.2) when censoring of deceased within the first 182 days and SHI changers (not significant at log-rank>0.05, log rank=0.082). When not censoring the first half year, then the results are significant at the 0.05 level (log rank =0.03). The ACO-Intervention group (1,430.1; 95% CI: 1,426.2 – 1,434.1) has a 9.4 days higher survival time than the Non-ACO-control group (1,420.7; 95% CI: 1,415.9 – 1,425.5) in that case.

4 Discussion

The evaluation approach, described in this paper, considers the complexity of evaluating the impact of ACOs on the *population health* Triple-Aim dimension. A claims-data-based quasi-experimental study design using PSM in combination with a small-scale exact matching approach is proposed, in order to control a possible bias caused by the non-randomized group assignment. In the application to the GKT context, it could be shown that intervention and control group can be balanced by adopting such an approach.

While matched pair approaches simulate the randomization balance (intervention vs. control group) in experimental studies, they can do so only for observable risk factors. Claims data might not include important factors such as in our GKT case for example out-of-pocket medication or health service utilization, or the health-consciousness of the patients and their treating physicians. Excluding these 'unobservable factors' might lead to a bias, a fact to be considered when discussing results[24,39]. Thus, calling, in particular if just small effects are observable, for validation through supporting evidence, where possible[19]. Also the applicability of the high-dimensional propensity score (HDPS) methodology should be tested in future studies, as it has shown potential to control for residual and unmeasured confounding in recent claims data based studies[40].

We also demonstrated that claims data offer, because of their good electronic availability and low collection costs, as well as their comprehensive longitudinal and cross-healthcare-provider-view, a good base for population health outcome measurement. Mortality indicators that can be measured

on the basis of claims data have been applied to the GKT model: mortality *ratio*, *age at the time of death*, *survival time* and an adapted individually age-adjusted years of potential life lost indicator (YPLLG). The mortality indicators used, show positive results towards the ACO, also after controlling for a potential (indirect) immortal time bias by excluding the first half year after enrollment from the outcome measurement. As we had to draw our Non-ACO-control group from the AOK and LKK insurees living in Kinzigtal and not enrolled in the ACO – because of the limited data set – the impact of the ACO may be underestimated. This is because the Non-ACO-control group, as part of the shared-savings contract, can also access some ACO interventions. In an ideal case, all 32,595 GKT insurees from the Kinzigtal region enclosed in the ACO-shared savings contract could be compared to a standard care group from other regions. We then could move more towards measuring population health in the sense of the health of the population in a geographic area and not just the enrolled population of ACO patients. This is actually intended by the GKT ACO contract. Also the amount of ACO enrollees that had to be excluded (n=1,010; 14.6%) because no adequate matched pair could be drawn from the limited data set, may be reduced with such an approach. In further studies these excluded ACO-enrollees have to be investigated in more detail.

Taking into account that the time span of this case study (four years) might be too short to perceive any realistic effects on mortality, further studies with longer observation periods will have to confirm the intermediate results presented here. Also the question, whether *years of potential life gained* are healthy years has to be addressed in future studies.

In view of the limitations of the chosen evaluation design from 'unobservable factors' in the matching procedure, a validation of the impact of the ACO on population health with supporting evidence through other quantitative and qualitative methods is recommended. In the case of GKT, other studies with different study designs, as well as external scientific evaluations support the results of this paper. For example Koester et al.[14] have shown improvements concerning overuse, underuse and misuse of care; and respective quality improvements in GKT. Siegel et al.[13] highlighted positive health-related behaviour changes in GKT in their study.

We describe here the application to a specific ACO, in this instance GKT. However, we believe that for its simplicity and ease of application, this evaluation approach can also be applied in other forms of ACOs and integrated care systems. We would thereby, in general, strongly recommend long observation periods for the evaluation of such population-based interventions, as effects from, for example, prevention programs may unfold their impact only over longer time periods. However, in many circumstances contractual, funding or other restrictions may prohibit such an approach. There may also be a need for more timely feedback to stimulate rapid learning processes. In these cases additional shorter term intermediate outcome measures of population health will be needed, in addition to the mortality indicators presented here in this paper. In future studies we plan to explore further claims-data-based population health indicators, and the transferability and usability of the presented evaluation design and mortality measures for comparative evaluations.

References

1. Berwick DM, Nolan TW, Whittington J. The triple aim: care, health, and cost. *Health Aff (Millwood)*. 2008;27(3):759-769. doi:10.1377/hlthaff.27.3.759.

2. Centers for Medicare. Accountable Care Organizations (ACOs): General information | Center for Medicare & Medicaid Innovation. http://innovation.cms.gov/initiatives/aco/. Published 2014. Accessed October 15, 2014.

3. Stein V, Barbazza ES, Tello J, Kluge H. Towards people-centred health services delivery: a Framework for Action for the World Health Organisation (WHO) European Region. *Int J Integr Care*. 2013;13(4):1-3.

4. Barnes AJ, Unruh L, Chukmaitov A, van Ginneken E. Accountable care organizations in the USA: Types, developments and challenges. *Health Policy*. 2014;118(1):1-7. doi:10.1016/j.healthpol.2014.07.019.

5. Stiefel M, Nolan K. *A Guide to Measuring the Triple Aim: Population Health, Experience of Care, and Per Capita Cost*. Cambridge, Massachusetts: Institute for Healthcare Improvement; 2012. http://thehanc.org/Events/HANC%20QI%20Peer%20Network%20Day%20Sept%2020%202013/IHIGuidetoMeasuringTripleAimWhitePaper2012.pdf. Accessed January 14, 2014.

6. Grumbach K, Grundy P. *Outcomes of Implementing Patient Centered Medical Home Interventions - A Review of the Evidence from Prospective Evaluation Studies in the United States*. Washington, DC: Washington, DC: Patient-Centered Primary Care Collaborative; 2010. http://forwww.pcpcc.net/files/evidence_outcomes_in_pcmh_2010.pdf. Accessed November 7, 2014.

7. McCarthy D, Klein S. The triple aim journey: improving population health and patients' experience of care, while reducing costs. *Commonw Fund Pub*. 2010;48(1421). http://mobile.commonwealthfund.org/~/media/Files/Publications/Case%20Study/2010/Jul/Triple%20Aim%20v2/1421_McCarthy_triple_aim_journey_overview.pdf. Accessed November 7, 2014.

8. Schulte T, Pimperl A, Fischer A, Dittmann B, Wendel P, Hildebrandt H. Ergebnisqualität Gesundes Kinzigtal - quantifiziert durch Mortalitätskennzahlen: Eine quasi-experimentelle Kohortenstudie: Propensity-Score-Matching von Eingeschriebenen vs. Nicht-Eingeschriebenen des Integrierten Versorgungsmodells auf Basis von Sekundärdaten der Kinzigtal-Population. June 2014. http://www.optimedis.de/files/Studien/Mortalitaetsstudie-2014/Mortalitaetsstudie-2014.pdf. Accessed November 11, 2014.

9. Romano PS, Geppert JJ, Davies S, Miller MR, Elixhauser A, McDonald KM. A National Profile Of Patient Safety In U.S. Hospitals. *Health Aff (Millwood)*. 2003;22(2):154-166. doi:10.1377/hlthaff.22.2.154.

10. Hildebrandt H, Schulte T, Stunder B. Triple Aim in Kinzigtal, Germany: Improving population health, integrating health care and reducing costs of care – lessons for the UK? *J Integr Care*. 2012;20(4):205-222. doi:10.1108/14769011211255249.

11. Pimperl A, Dittmann B, Fischer A, Schulte T, Wendel P, Hildebrandt H. Wie aus Daten Wert entsteht: Erfahrungen aus dem Integrierten Versorgungssystem "Gesundes Kinzigtal." In: Langkafel P, ed. *Big data in der Medizin und Gesundheitswirtschaft: Diagnose, Therapie, Nebenwirkungen*. Heidelberg, Neckar: medhochzwei Verlag; 2014:83-102.

12. Schulte T, Pimperl A, Dittmann B, Wendel P, Hildebrandt H. Drei Dimensionen im internen Vergleich: Akzeptanz, Ergebnisqualität und Wirtschaftlichkeit der Integrierten Versorgung Gesundes Kinzigtal. 2012. http://www.optimedis.de/images/docs/aktuelles/121026_drei_dimensionen.pdf. Accessed March 13, 2013.

13. Siegel A, Stößel U, Zerpies E. *GEKIM – Gesundes Kinzigtal Mitgliederbefragung: Bericht Zur Ersten Mitgliederbefragung 2012/13*. Freiburg im Breisgau: Universität Freiburg; 2013. www.ekiv.org. Accessed March 14, 2014.

14. Köster I, Ihle P, Schubert I. *Evaluationsbericht 2004-2011 für Gesundes Kinzigtal GmbH: hier: AOK-Daten*. Köln: Universität Köln, PMV -Forschungsgruppe; 2014. www.ekiv.org. Accessed August 14, 2014.

15. Meyer GS, Nelson EC, Pryor DB, et al. More quality measures versus measuring what matters: a call for balance and parsimony. *BMJ Qual Saf*. 2012;21(11):964-968.

16. Shadish WR, Cook TD, Campbell DT. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. 2 edition. Boston: Houghton Mifflin; 2001.

17. Morgan SL, Winship C. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. 2nd ed. Cambridge: Cambridge University Press; 2014.

18. Eccles M, Grimshaw J, Campbell M, Ramsay C. Research designs for studies evaluating the effectiveness of change and improvement strategies. *Qual Saf Health Care*. 2003;12(1):47-52. doi:10.1136/qhc.12.1.47.

19. Craig P, Dieppe P, Macintyre S, Michie S, Nazareth I, Petticrew M. *Developing and Evaluating Complex Interventions: New Guidance*. Medical Research Council; 2008. http://www.mrc.ac.uk/documents/pdf/complex-interventions-guidance/. Accessed November 6, 2014.

20. Treweek S, Zwarenstein M. Making trials matter: pragmatic and explanatory trials and the problem of applicability. *Trials*. 2009;10:37. doi:10.1186/1745-6215-10-37.

21. Pimperl A, Schreyögg J, Rothgang H, Busse R, Glaeske G, Hildebrandt H. Ökonomische Erfolgsmessung von integrierten Versorgungsnetzen – Gütekriterien, Herausforderungen, Best-Practice-Modell. *Gesundheitswesen*. September 2014. doi:10.1055/s-0034-1381988.

22. Black N. Why we need observational studies to evaluate the effectiveness of health care. *BMJ*. 1996;312(7040):1215-1218. doi:10.1136/bmj.312.7040.1215.

23. Radosevich DM. Designing an Outcomes Research Study. In: Kane R, ed. *Understanding Health Care Outcomes Research*. 2nd ed. Sudbury MA, Mississauga Ontaria, London: Jones & Bartlett Publishers; 2006:23-57.

24. Legewie J. Die Schätzung von kausalen Effekten: Überlegungen zu Methoden der Kausalanalyse anhand von Kontexteffekten in der Schule. *KZfSS Köln Z Für Soziol Sozialpsychologie*. 2012;64(1):123-153. doi:10.1007/s11577-012-0158-5.

25. Angrist JD, Pischke J-S. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton: Princeton University Press; 2008.

26. Stuart EA. Matching methods for causal inference: A review and a look forward. *Stat Sci Rev J Inst Math Stat*. 2010;25(1):1-21. doi:10.1214/09-STS313.

27. Rosenbaum PR, Rubin DB. Constructing a Control Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score. *Am Stat*. 1985;39(1):33-38. doi:10.2307/2683903.

28. Austin PC. An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. *Multivar Behav Res*. 2011;46(3):399-424. doi:10.1080/00273171.2011.568786.

29. Sundararajan V, Henderson T, Perry C, Muggivan A, Quan H, Ghali WA. New ICD-10 version of the Charlson comorbidity index predicted in-hospital mortality. *J Clin Epidemiol*. 2004;57(12):1288-1294. doi:10.1016/j.jclinepi.2004.03.012.

30. Knesebeck O von dem, Lüschen G, Cockerham WC, Siegrist J. Socioeconomic status and health among the aged in the United States and Germany: A comparative cross-sectional study. *Soc Sci Med*. 2003;57(9):1643-1652. doi:10.1016/S0277-9536(03)00020-0.

31. Gu XS, Rosenbaum PR. Comparison of Multivariate Matching Methods: Structures, Distances, and Algorithms. *J Comput Graph Stat*. 1993;2(4):405-420. doi:10.1080/10618600.1993.10474623.

32. Levesque LE, Hanley JA, Kezouh A, Suissa S. Problem of immortal time bias in cohort studies: example using statins for preventing progression of diabetes. *BMJ*. 2010;340:b5087. doi:10.1136/bmj.b5087.

33. Austin PC. A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. *Stat Med*. 2008;27(12):2037-2049. doi:10.1002/sim.3150.

34. Murray PK, Singer M, Dawson NV, Thomas CL, Cebul RD. Outcomes of rehabilitation services for nursing home residents. *Arch Phys Med Rehabil*. 2003;84(8):1129-1136. doi:10.1016/S0003-9993(03)00149-7.

35. Schneider EC. Measuring mortality outcomes to improve health care: rational use of ratings and rankings. *Med Care*. 2002;40(1):1-3.

36. Dranger E, Remington P. *YPLL: A Summary Measure of Premature Mortality Used in Measuring the Health of Communities*. University of Wisconsin Public Health and Health Policy Institut; 2004. http://uwphi.pophealth.wisc.edu/publications/issue-briefs/issueBriefv05n07.pdf. Accessed August 18, 2014.

37. Statistisches Bundesamt. *Generationensterbetafeln Für Deutschland: Modellrechnungen Für Die Geburtsjahrgänge 1896-2009*. Wiesbaden: Statistisches Bundesamt; 2011. https://www.destatis.de/DE/ZahlenFakten/GesellschaftStaat/Bevoelkerung/Sterbefaelle/Tabellen/GenerationensterbetafelMethoden.pdf?__blob=publicationFile. Accessed March 14, 2014.

38. Kaplan EL, Meier P. Nonparametric Estimation from Incomplete Observations. *J Am Stat Assoc*. 1958;53(282):457. doi:10.2307/2281868.

39. Schlesselman JJ. Assessing effects of confounding variables. *Am J Epidemiol*. 1978;108(1):3-8.

40. Garbe E, Kloss S, Suling M, Pigeot I, Schneeweiss S. High-dimensional versus conventional propensity scores in a comparative effectiveness study of coxibs and reduced upper gastrointestinal complications. *Eur J Clin Pharmacol*. 2013;69(3):549-557. doi:10.1007/s00228-012-1334-2.

Address correspondence to:

Dr. Alexander Pimperl

University of California, Berkeley

School of Public Health

50 University Hall, #7360

Berkeley, CA 94720