

RESEARCH ARTICLE

Open Access



Assessing the effect of a regional integrated care model over ten years using quality indicators based on claims data – the basic statistical methodology of the INTEGRAL project

Dominikus Stelzer^{1*} , Erika Graf¹, Ingrid Köster², Peter Ihle², Christian Günster³, Patrik Dröge³, Andreas Klöss³, Claudia Mehl⁴, Erik Farin-Glattacker¹, Max Geraedts⁴, Ingrid Schubert², Achim Siegel⁵ and Werner Vach⁶

Abstract

Background: The regional integrated health care model “Healthy Kinzigtal” started in 2006 with the goal of optimizing health care and economic efficiency. The INTEGRAL project aimed at evaluating the effect of this model on the quality of care over the first 10 years.

Methods: This methodological protocol supplements the study protocol and the main publication of the project. Comparing quality indicators based on claims data between the intervention region and 13 structurally similar control regions constitutes the basic scientific approach. Methodological key issues in performing such a comparison are identified and solutions are presented.

Results: A key step in the analysis is the assessment of a potential trend in prevalence for a single quality indicator over time in the intervention region compared to the corresponding trends in the control regions. This step has to take into account that there may be a common - not necessarily linear - trend in the indicator over time and that trends can also appear by chance. Conceptual and statistical approaches were developed to handle this key step and to assess in addition the overall evidence for an intervention effect across all indicators. The methodology can be extended in several directions of interest.

Conclusions: We believe that our approach can handle the major statistical challenges: population differences are addressed by standardization; we offer transparency with respect to the derivation of the key figures; global time trends and structural changes do not invalidate the analyses; the regional variation in time trends is taken into account. Overall, the project demanded substantial efforts to ensure adequateness, validity and transparency.

Keywords: Integrated care model, Evaluation, Quality indicator, Statistical analysis

*Correspondence: stelzer@imbi.uni-freiburg.de

¹Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center, University of Freiburg, Freiburg, Germany

Full list of author information is available at the end of the article



© The Author(s). 2022 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

Background

The integrated health care model “Healthy Kinzigtal” was established in 2006. The model addresses the full spectrum of morbidities and health issues for a population defined by a residential area (with the only exception of dental care). It is based on a contract among a regional physicians’ network, a management and holding company specialised in integrated care, and the Allgemeine Ortskrankenkasse Baden-Württemberg (AOK-BW), the largest statutory health insurance fund in the federal state of Baden-Württemberg (BW). The contract is a so-called shared savings contract, that is, the healthcare cost savings achieved are distributed between the contractual partners. (For further details we refer to the study protocol [1] and the main project publication [2].) From an economic perspective the contract was rather successful, as cost savings were achieved. However, a shared savings contract may lead to lower levels of care, that is, an underutilisation of health services. Consequently, an evaluation of the health care quality of the model is highly relevant. The aim of the INTEGRAL project was to conduct such an evaluation.

Comparing quality indicators based on claims data between the intervention region and 13 structurally similar control regions constitutes the basic scientific approach. More precisely, the idea was to determine for a single indicator the prevalence trend in each region and to compare the observed trend in the intervention region with the trends in the control regions. The control regions allowed not only a determination of the expected trend in the absence of the intervention. They also allowed the determination of the natural variation of these trends, i.e. we did not a priori assume that the trends are equal across the control regions. Regional differences in health care are well established in Germany [3], and it cannot be ruled out that there are also regional differences in the prevalence trends of quality indicators, reflecting ongoing changes in the local situation. Hence the fact that a trend in an intervention region differs from a mean trend in the control regions does not necessarily indicate an effect of the intervention. Such differences may also happen by chance, and therefore an evaluation has to take into account the variation in trends observed in the control regions. We hence aimed to conclude a specific role of the intervention only in the case that the observed trend in the intervention region could be regarded as an extreme one relative to the variation observed in the control regions.

The study protocol [1] mentions some statistical challenges in implementing the basic approach for a single indicator: 1) In spite of the structural similarity of all regions, differences in the population composition with respect to important risk factors are to be expected. This may lead to variation in crude prevalences at baseline. 2) Moreover, the populations may change over time, and

they may change differently in different regions. 3) There may be global time trends or structural changes in the prevalence numbers. 4) There may be floor or ceiling effects. 5) To support final decisions, it might be desirable to reduce complex patterns in time trends across regions to a single number. This, however, may hide important data features influencing the final interpretation. 6) There is an interest in analysing an entire set of quality indicators. Consequently, we have to measure or phrase the intervention effect in a manner comparable across all indicators.

The purpose of this paper is to present the conceptual and statistical approach chosen to tackle these issues in the analysis of a single indicator, as well as the approach chosen to assess the overall evidence for an impact of the intervention. In addition, some further ideas and extensions are discussed.

Methods

Selection of control regions

The contract covers the Kinzigtal region, which is located in the Black Forest in Southwest Germany and home to about 70,000 inhabitants. This defines the intervention region.

The control regions were selected in a process described in detail in Additional file 1. As a first step, potential control regions were identified based on the following criteria, attempting to mimic basic features of the intervention region: geographically contiguous area; rural community or small to medium sized town (<50,000 inhabitants); river valley or active physicians’ network existing already in 2005. Regions with an integrated care contract lasting at least until 2015 were excluded. In a second step, the distribution of a series of structural indicators reflecting the social, economic and health services structures were compared among the 29 regions identified in the first step. The aim was to identify the regions most similar to the intervention region. In addition, the size requirement was tightened to include only regions with at least 35,000 inhabitants and to cover only towns with maximally 30,000 inhabitants. Immediate proximity to a hospital with maximum service level or to a major city, common border with Switzerland, and high internal heterogeneity were defined as further exclusion criteria. It was also taken into account that the overall number of insureds in the control regions and the intervention region should not exceed 500,000 in order to fulfill requirements on data protection. Finally, 13 control regions were identified that were regarded as showing a pattern of the structural indicators similar to the intervention region. Seven of the 13 control regions had an active physicians’ network.

We number the regions from 1 to $R = 13$ for the control regions and use either $r = 0$ or KT to designate the intervention region Kinzigtal (KT).

Selection of indicators

The quality indicators to be analysed in this project were selected in a process described in [4]. In the end only indicators referring to a binary event in a specific population were selected. So in the classical terminology of a population-based quality indicator, it consists of a definition of a denominator, i.e. the size of a population of interest, and a numerator, i.e. the number of subjects in this population who experienced the event of interest. All selected indicators referred to populations defined on an annual basis, reflecting a common practice in defining indicators. Not all primarily selected indicators could be operationalized based on the available claims data, and some indicators could be operationalized in different manners [5]. Finally, 119 indicators were candidates for the statistical analyses. However, 13 indicators showed only very few events or non-events, such that the formal analysis described in this paper was not feasible. Among the 106 remaining indicators the prevalence ranged from 0.12% to 96.7% in the BW sample. As part of this selection process, the desired direction of a change in the indicator was also defined, i.e. whether an increase or a decrease is regarded as desirable. For five indicators, such a decision was not possible. Finally, it was decided whether it was reasonable to apply an indicator to subjects covered by a family doctor-centered healthcare contract or not.

The data and the intervention(s)

The project makes use of the routine claims data of the AOK-BW. Of the 70,000 inhabitants of the Kinzigtal region, 33,000 are insured with the AOK-BW. Overall, the control regions cover about 452,000 insured of the AOK-BW.

The shared savings contract and the regional integrated care model evaluated in the INTEGRAL project implied many different concrete interventions initiated by the management company. The interventions aimed at fostering patient self-management and shared decision-making (e.g. by supporting the use of individual treatment plans and goal-setting agreements) and at coordinating the care efforts across different sectors. A more detailed description can be found in [6]. It should be noted that many of these interventions aimed at an improvement at the system level and hence also patients not insured by the AOK-BW may have been affected. Effects in this patient population could of course not be evaluated in the INTEGRAL project.

Time range

The time range included the years 2006 to 2015. It was originally planned to include data from 2005, i.e. the year prior to the start of the intervention. However, it turned out that the data of 2005 were not of sufficient quality to be included in the analysis. The insufficient quality can be

explained by the fact that the data warehouse of the AOK research institute (WIdO) was still in its start-up phase at that time.

Data preprocessing

The indicators were operationalized based on the claims data of the AOK-BW. Details are described in [5]. Roughly speaking, for each calendar year from 2006 to 2015 we identified the population living in the intervention region or one of the control regions who were insured of the AOK and satisfied the denominator definition of the indicator. Then for each member of this population it was determined whether the event of interest happened or not. More details are outlined in Additional file 2.

In addition, a 10% sample of the entire BW population of members of the AOK was drawn as described in Additional file 2. Excluding subjects living in the intervention region (but not those living in one of the control regions), an additional region – called the “BW region” in the sequel – was defined for each year. This was regarded as an alternative control region supplementing the control regions described above. However, when referring to “control regions” in the following, we do not include this additional region. When referring to region numbers, this region has the index $R + 1 = 14$, but also the index “BW” is used. The 10% sample will be also used for the purpose of standardization.

For a single indicator considered in this project, the analysis population is defined as the union of all these populations, i.e. from the intervention region, the preselected control regions, and the BW region. Subjects could be followed over time, and hence the basic data can be described by random variables Y_{it} , denoting the outcome of interest in subject i in year t . The outcome may not be defined for all years in the event a subject enters or leaves the denominator population. We implicitly assume in the following that for each year only the subjects included in the denominator population of the indicator are included in the analysis. In addition, the variable R_{it} denotes the region subject i was living in year t , which may change over time due to moving. The years are numbered from $t = 0$ to T . The value of T was 9 for most indicators, but a few indicators could not be operationalized in the first years and covered a more narrow time range.

Unit of prevalence trends

Independent of the type of event defining an indicator, we refer to the frequency of an event in the denominator population of a year as the (annual) prevalence. Time trends in prevalences, i.e. changes in prevalence over time, will build the main corner stone of the analytical approach. We decided to express trends as absolute differences over time, as stakeholders reading the analyses of single indicators can be expected to have some background knowledge

– including an idea about the prevalence – such that they can interpret absolute differences. When later summarizing results across different indicators, we will change partially to an odds ratio scale to achieve a better comparability across indicators.

We decided to report trends in prevalence with respect to a five years interval to assist in the interpretation. Five years reflect a time period at which one typically expects to see an effect from a population-based intervention. We did not report annual changes, as such numbers tend to be small, and hence invite an overly pessimistic interpretation. Irrespective of this timescale, all trend estimates will be based on all years for which data was available. We further decided to report the five-year trends in percentage points instead of fractions between 0 and 1, again mainly to avoid over-pessimistic interpretations.

Confounders

The comparison regions were selected with the aim of facilitating comparability with the intervention region. Nevertheless, we have to expect differences in the composition of patient populations with respect to variables such as age, gender, comorbidity status and socio-economic status (SES). These differences can also explain differences in the prevalence of indicators. Hence, it is a common approach in population epidemiology to adjust for such differences by appropriate standardization [7, 8].

This is typically done in order to allow a fair comparison across regions with respect to the prevalence. This is to some degree also relevant in our context, as comparability of prevalences across regions at baseline facilitates the interpretation of trends. However, our main interest lies with the time trends themselves. It is thus relevant for us to address confounding with respect to time trends. A typical example would be age demographics with differences in the growth of elderly populations across regions. This will increase the variation in trends for any indicator with the probability of an event increasing with age, e.g. the population prevalence in diabetes. This may even bias the estimation of intervention effects, if the intervention region has a specific speed.

A further crucial point may be that the intervention itself may change the distribution of covariates in a way, such that adjustment introduces a bias. For example, the intervention may reduce comorbidity, and then adjustment for comorbidity punishes the intervention region for this. Or the intervention may improve the documentation of the comorbidity level, and hence the population in the intervention region seems sicker on average than in reality, giving the intervention region an unwarranted advantage when adjusting. To avoid such problems, we will use the insuree's comorbidity level at the time of entering the study population instead of using the actual value from each year.

The use of claims data restricts the possibility to define confounders. We make use of the following three potential confounders:

- (i) age (in years)
- (ii) gender
- (iii) Charlson comorbidity index at study entry

In addition, we make use of the SES, operationalized by the German Index of Socioeconomic Deprivation (GISD, [9, 10]). While not available at the individual level, the SES was available at the regional level of municipalities associations and could be assigned by postal code. The intervention region and the 13 control regions covered 1687 different post codes overall, such that on average 4518 subjects shared a postal code. For the study population the GISD varied between 4.76 and 8.72, where higher values stand for a more pronounced deprivation and hence a lower SES.

Family doctor-centred healthcare

AOK-insurees in the federal state of BW are offered participation in a specific family doctor-centered healthcare program. Insurees enrolling into this program choose a fixed general practitioner (GP) as their family doctor. Unfortunately, for some health services, this implies a lack of claims data in these insurees, as they are covered by a general fee for the GP. Consequently, if an insuree signs such a contract, certain events will not be visible in the claims data. If the event definition of an indicator could have been affected by this, the patients with this type of contract were removed from the analysis population. This may introduce a biased selection, as insurees signing such a contract may be more (or less) healthy than the general population. However, by adjusting for the above mentioned confounders, we can at least limit a corresponding selection bias.

Relevance limits

The precision at which prevalence trends can be estimated varies highly from indicator to indicator due to differences in the size of the denominator population and the prevalence. Hence it would be highly misleading to regard statistical significance as the solely relevant criterion. The magnitude of the observed trend differences should be taken into account, too.

Originally, the idea was to ask the medical experts involved in the development of the indicators to also define limits for clinical relevance. This turned out to be not manageable due to the inherent arbitrariness and also the large number of indicators involving very different specialities. We hence decided to use a numerical criterion. For this we first considered the range of potential improvement, taking the prevalence π in the BW sample in the initial year as a starting point. For

an indicator with a desired increase in prevalence, the potential improvement was then defined as the difference between 100% and the prevalence. For an indicator with a desired decrease in prevalence, it was defined as the difference between the prevalence and 0%. A consensus process within the research group resulted in setting the relevance limit as 10% of the potential improvement, as well as to require minimum 0.1 percentage point improvement. (Cf. [11] for a similar approach).

The medical practice of a patient

For each patient in the analysis population of an indicator and for each calendar year, the health care provider who was likely responsible for the management of the patient was determined. This was assessed by an algorithm that, roughly speaking, identified the primary responsible provider (depending on the tracer diagnosis) based on the following criteria (in descending order): highest number of treatment quarters, contacts (days with services), and number of services billed. In case of ties, the provider with the first contact in the calendar year was chosen. If still a unique provider could not be identified, a random selection was made. In general, health care provider refers here not to a single individual, but to a practice, possibly constituted by several individuals. We refer to this as the practice of the patient. The precision of the assignment varies from indicator to indicator, as for some indicators it can be challenging to determine a responsible provider. This information will not be used in the main analytic approach, but only in sensitivity analyses and specific extensions.

Main analytic approach

The analysis of a single indicator is based on the following five steps:

- (i) estimating standardized annual region-specific prevalences p_{rt}
- (ii) estimating five-year time trends θ_r in each region
- (iii) estimating the mean trend μ_C in the control regions and the standard deviation σ_C of the true trends across the control regions
- (iv) computing key figures to assess the difference in trend in the intervention region relative to the control regions, in particular $\hat{\Delta}_C = \hat{\theta}_{KT} - \hat{\mu}_C$, i.e., the difference between the trend in the intervention region and the mean trend in the control regions, and a z-score z relating the observed difference $\hat{\Delta}_C$ to the standard deviation $\hat{\sigma}_C$ of the trends in the control regions
- (v) verbal classification of the results as a strong positive (or negative) hint, a regular positive (or negative) hint, a weak positive (or negative) hint, or inconclusive while taking the relevance of the magnitude of the difference Δ_C into account

In principle, the first four steps could be replaced by fitting one complex model integrating all steps. We preferred a step-wise approach, as it allowed us to achieve more transparency about the process from the original data to the final results. To increase this transparency, the numerical results from each step are visualized in a user-friendly manner. This is exemplified in the single indicator “Treatment with acetylsalicylic acid (ASA)” in the sequel. This indicator aimed at the percentage of coronary heart disease patients who received a prescription for ASA within the last 12 months. Further examples will be presented later.

All statistical computations were performed with Stata 15.

Estimating standardized prevalences

Estimates of the annual region-specific prevalences were based on a direct standardization to the 10%-sample of BW. This was achieved by 1) using the maximum likelihood principle to fit a logistic regression model to the data from all regions and time points describing the individual probability of an event as a function of region, calendar year and the four potential confounders, 2) applying the derived (year and region specific) prediction rule to all subjects in the BW sample, and 3) averaging over all the individual probabilities.

In the logistic regression model it is assumed that all confounders have the same effect in all regions and at all time points, but we allow a region and time specific intercept. Due to the large sample size we allow for gender specific effects of age, comorbidity and GISD and also potential non-linear effects of age and comorbidity.

The logistic regression model considered reads

$$\begin{aligned} \text{logit } P_{\alpha, \beta} (Y_{it} = 1 | X_{it}) = & \alpha_{R_{it}} \\ & + \beta_{\text{gender}} \mathbb{1}_{\{\text{gender}_i=1\}} + \beta_{\text{gisd}}^{\text{gender}_i} \text{gisd}_{it} \\ & + f_{\beta_{\text{age}}^{\text{gender}_i}}^{\text{age}}(\text{age}_{it}) + f_{\beta_{\text{comorb}}^{\text{gender}_i}}^{\text{comorb}}(\text{comorb}_{it}) \end{aligned}$$

with α and β denoting parameter vectors and f functions with self-explanatory indexing. The gender specific parametrization is dropped if there are < 100 events (or non-events) in one gender. For the gender-specific age effects we chose $f_{\beta_{\text{age}}^{\text{gender}_i}}^{\text{age}}(x)$ as restricted cubic splines (also called natural cubic splines) with k knots. k is chosen as the difference between the 99% and 1%-ile of the gender specific age distribution, divided by 10 and rounded up to the next integer while allowing a minimum value of 2 and a maximum value of 4. The spline knots are placed at the 10th, 50th and 90th percentile if $k = 3$, and at the 5th, 35th, 65th and 95th percentile if $k = 4$. In the case $k = 2$ the function $f_{\beta_{\text{age}}^{\text{gender}_i}}^{\text{age}}(x)$ is linear, so the knot placement does

not matter. To increase the numerical stability, the splines are based on the mean centered version of the variable age. For the gender-specific effect of comorbidity we consider an alternative approach taking the typical skewed distribution of comorbidity indices into account. We use the function

$$f_{\beta}^{comorb}(x) = \alpha + \left(\sum_{l=0}^{L-1} \beta_l 1_{\{x=l\}} \right) + \beta_L(x - L)1_{\{x \geq L\}}.$$

Here L is chosen as the upper 90%-ile of the gender-specific distribution of the integer-valued variable *comorbidity*, rounding down but choosing at most the second largest observed value. This way it is ensured that the linear part is based on at least 10% of the available observations and that there are at least two different values among these observations. Furthermore, if only two different values are observed in the variable *comorbidity*, L is set to 0 and the model reduces to a simple linear model. The variable is dropped from the model if it is constant. Since there is already an explicit gender effect in the model, we set $\alpha = 0$ to ensure identifiability.

The standardized prevalences are obtained as $\hat{p}_{rt} =$

$$\frac{1}{|S_t^{BW}|} \sum_{i \in S_t^{BW}} \Lambda(\hat{\alpha}_{rt} + \hat{\beta}_{gender} 1_{\{gender_i=1\}} + \hat{\beta}_{gisd}^{gender_i} gisd_{it} + f_{\beta_{age}^{gender_i}}^{age}(age_{it}) + f_{\beta_{comorb}^{gender_i}}^{comorb}(comorb_{it})),$$

with S_t^{BW} denoting the individuals in the BW sample at time t and

$$\Lambda(x) = \text{logit}^{-1}(x) = \frac{1}{1 + e^{-x}}.$$

The standard error of each standardized prevalence is obtained by application of the delta rule. Standard errors for the parameter estimates of the logistic regression model are based on robust standard errors taking clustering within an individual over time into account. Due to the large sample size of the BW sample, the full application of the delta rule was not possible and a Monte Carlo approximation was used as outlined in the [Appendix](#).

The estimated standardized prevalences are visualized by line plots (Fig. 1) distinguishing the intervention region, the control regions, and the BW region by different

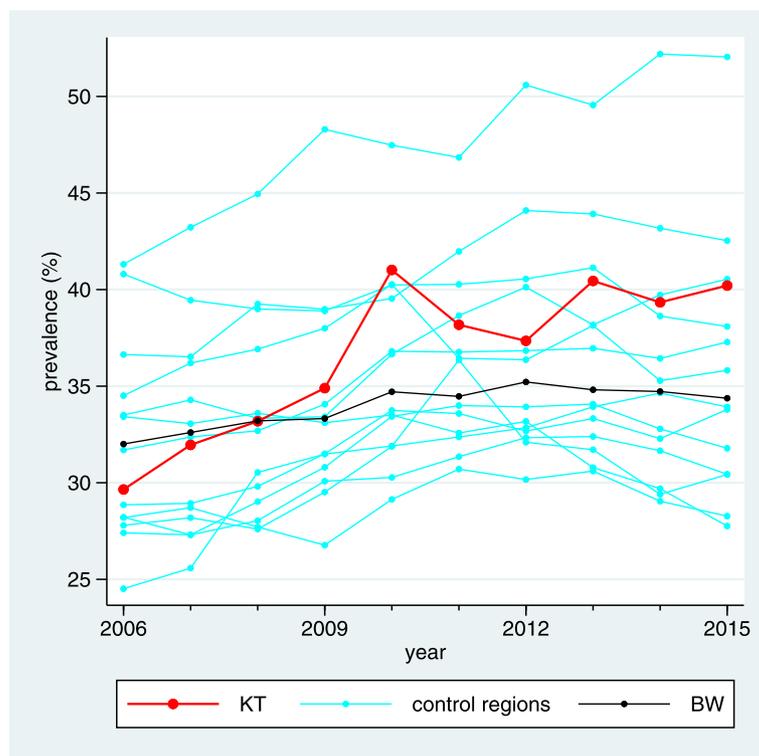


Fig. 1 Visualization of the estimated standardized prevalences in a line plot for the indicator “Treatment with acetylsalicylic acid (ASA)”, for which high prevalences are desired. In the BW region we can observe a slight upwards trend and the trends vary substantially across the control regions. In the intervention region “Kinzigtal” (KT) the prevalence starts below the BW prevalence, but is distinctly above the BW prevalence at the end of the observation period

colors. No confidence intervals are shown in these plots to avoid an overcrowded visualization.

Estimating time trends

If one could be sure that a clear linear trend were visible in the prevalences within each region, assessment of a trend would be rather straightforward. However, we have to anticipate that for some indicators there may be a non-linear overall trend, for example a flattening decrease/increase or a structural change due to administrative reasons. In such a situation, the trend of interest is the trend on top of the common pattern across all regions, i.e. how a region is developing compared to the overall trend. For example, when leaving aside the general trend, a prevalence trend of -0.01 over 5 years for a region should reflect that the prevalence decreased by one percentage point faster in five years as compared to the general trend. We refer to this definition of a trend as the “on-top-trend”, which is used independent of the observed pattern in the general trend. However, as on-top-trends do not reflect a general linear trend, they do not coincide with the trend visible in the inspection of the line plots mentioned above. To avoid this conflict, an estimate of the overall linear trend is added.

To obtain estimates for the on-top-trend $\tilde{\theta}_r$, we fit a meta-analytic model of the type

$$\hat{p}_{rt} = \alpha_t + \beta_r + \tilde{\theta}_r t + \gamma_{rt} + \epsilon_{rt} \tag{1}$$

with $\gamma_{rt} \sim N(0, \tau)$ and $\epsilon_{rt} \sim N(0, \hat{\sigma}_{rt})$

with the side conditions $\sum_t \alpha_t = 0$ and $\sum_r \tilde{\theta}_r = 0$. Here τ denotes the unexplained variation across regions and time points (estimated in fitting this model), whereas $\hat{\sigma}_{rt}$ denotes the previously obtained estimate of the standard error of \hat{p}_{rt} for $r = 0, \dots, R + 1$ and $t = 0, \dots, T$, which is plugged in.

Estimates of the 5-year trends θ_r are then obtained by adding an estimate of the (linear) overall trend and multiplication with 5, i.e.

$$\hat{\theta}_r := 5(\tilde{\theta}_r + \hat{\theta})$$

with $\hat{\theta}$ derived as the ordinary least squares (OLS) estimate from fitting the model

$$\hat{\alpha}_t = \mu + \theta t + \epsilon_t.$$

Note that we have for $\hat{\theta}$ the explicit representation

$$\hat{\theta} = \frac{\sum_{t=0}^T (t - \bar{t})(\alpha_t - \bar{\alpha})}{\sum_{t=0}^T (t - \bar{t})^2} = \frac{\sum_{t=0}^T t \alpha_t}{\sum_{t=0}^T (t - \bar{t})^2}.$$

Consequently, we have an explicit representation of $\hat{\theta}_r$ as a function of $(\hat{\theta}_r)_{r=0, \dots, R+1}$ and $(\hat{\alpha}_t)_{t=0, \dots, T}$, allowing us to compute easily its standard error based on the variance-covariance matrix of $(\hat{\theta}_r)_{r=0, \dots, R+1}$ and $(\hat{\alpha}_t)_{t=0, \dots, T}$. The results of this step – i.e. the region specific trend estimates

with 95% confidence intervals – are visualized by a forest plot as shown in Fig. 2.

The choice of the structure of the model (1) is basically motivated by what can be achieved with standard software for meta regression – in particular the `metareg` command of Stata that we used. For this reason we could not incorporate the (estimable) correlations between the different \hat{p}_{rt} (in particular within one region) and could not allow region specific variances for γ_{rt} .

Estimating the mean and standard deviation of the true trends across the control regions

Estimates for the mean μ_C and the standard deviation σ_C of the true trends are obtained by considering a meta-analytic model for the estimated trends, i.e.

$$\hat{\theta}_r = \gamma_r + \epsilon_r \text{ with } \gamma_r \sim N(\mu_C, \sigma_C) \text{ and } \epsilon_r \sim N(0, \hat{\sigma}_r)$$

with $\hat{\sigma}_r$ denoting an estimate of the standard error of $\hat{\theta}_r$. We made use here of Stata’s `metan` command using the `random` option, implementing the method of [12].

Computing key figures

The first key figure of interest is the difference between the trend in the intervention region and the mean trend in the control regions:

$$\Delta_C = \theta_{KT} - \mu_C$$

This informs us about the magnitude of the difference, and allows us in particular to judge the relevance. However, this does not address the question of which degree the difference can be regarded as exceptional or whether it is within the variation of trends to be expected when considering small local regions. To address this question, we relate the observed difference to the standard deviation of the trends in the control regions in the spirit of a z-score:

$$z = \frac{\theta_{KT} - \mu_C}{\sigma_C} = \frac{\Delta_C}{\sigma_C}$$

As a third supportive figure we consider the difference between the trend in the intervention region and the trend in the BW-region:

$$\Delta_{BW} = \theta_{KT} - \theta_{BW}$$

This is mainly included to allow a comparison with previous analyses based on a comparison with figures from BW [13]. The expectation is that the estimates of Δ_C and Δ_{BW} are similar. Discrepancies may remind us that the control regions differ substantially from the whole federal state of BW, which has to be taken into account when extrapolating effects of the intervention to the whole state.

For all three key figures we report estimates and confidence intervals. For the differences we also report p-values

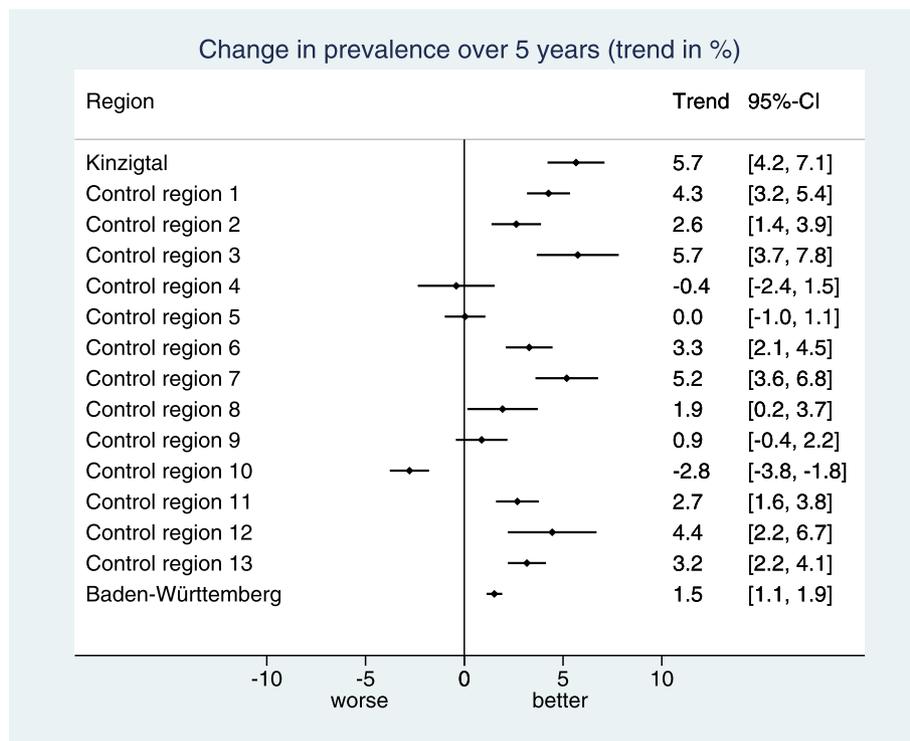


Fig. 2 Visualization of the estimated trends per region in a forest plot for the indicator “Treatment with acetylsalicylic acid (ASA)”. We can observe a positive trend for most control regions, i.e. also the control regions are able to move into the desired direction. However, the intervention region “Kinzigtal” shows together with one control region the most pronounced trend

for a test of the null hypothesis of no difference. Confidence intervals for Δ_C and Δ_{BW} are computed as Wald-type intervals assuming independence between $\hat{\theta}_{KT}$ and $\hat{\mu}_C$ and between $\hat{\theta}_{KT}$ and $\hat{\theta}_{BW}$, respectively. Confidence intervals for z are based on Fieller’s method assuming independence between $\hat{\Delta}_C$ and $\hat{\sigma}_C$ [14].

Conceptually, the assumption of independence between the different $\hat{\theta}_r$ can be justified by the fact that they reflect mainly the data from the different regions, i.e. non-overlapping sources of information. However, as they are based on fitting one model, slight correlations cannot be excluded. Independence between $\hat{\Delta}_C$ and $\hat{\sigma}_C$ may be justified by the general result on independence between estimates of fixed effect parameter and random effect variances in mixed models.

In order to facilitate the understanding of the background of these key figures, we suggest a two-dimensional visualization of all trend estimates and the estimated mean and standard deviation of the true trends in the control regions. This way the complete input used in computing the key figures becomes visible. In this visualization (Fig. 3) the x-axis refers to the trend and each trend estimate is reflected by a vertical line. Different colors are used for the control regions, the intervention region and the BW region, respectively. $\hat{\mu}_C$ and $\hat{\sigma}_C$ are transformed

to a normal density superposed over the whole graph. $\hat{\Delta}_C$ and $\hat{\Delta}_{BW}$ are then visible as differences between lines of different colors, and \hat{z} corresponds to the relative position of $\hat{\theta}_{KT}$ to the density function reflecting the variation of trends in the control regions.

Verbal classification of the results

To come to a final conclusion about a potential specific role of the intervention region with regard to a single indicator, it is definitely not sufficient to look at the p -value to reject the null hypothesis $H_0 : \Delta_C = 0$. This would ignore the relevance of the magnitude of the difference, and the relative position of the trend as expressed by the z -score. Since there is some freedom in how to summarize these different aspects in a final conclusion, we suggest a formal rule for verbalization of the results. This seems to be useful here, as we have to judge a large number of indicators, and a pre-specified formal rule helps to ensure a uniform, objective and transparent handling of all indicators. This can also serve as a basis for a formal assessment of the evidence about the role of the intervention across all indicators.

In the following, we denote with $\bar{\Delta}_C$ and \bar{z} the directed version of the estimates $\hat{\Delta}_C$ and \hat{z} , i.e. in the case where a numerically negative trend implies an improvement in

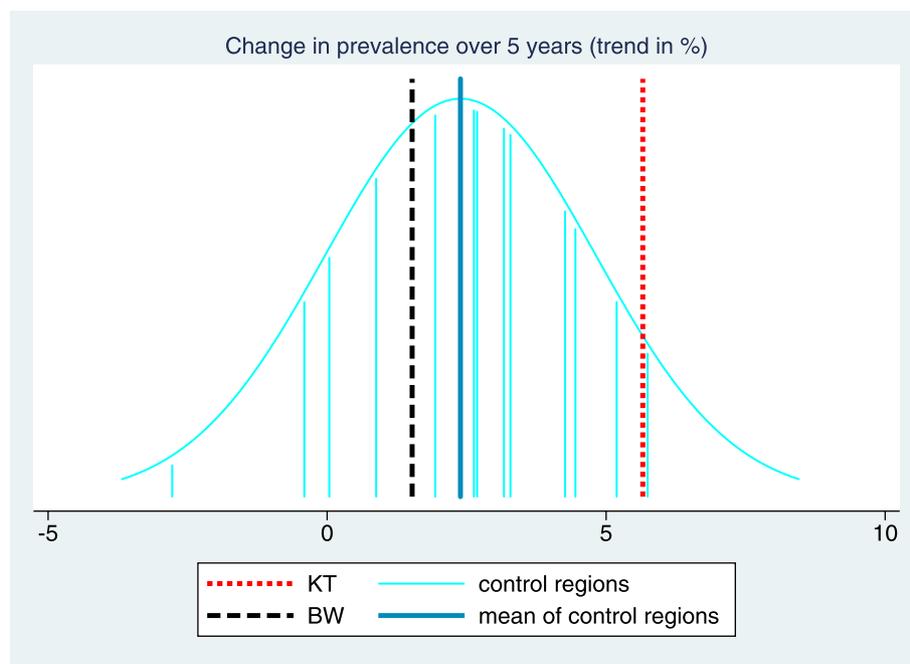


Fig. 3 Visualization and tabulation of the key figures for the indicator “Treatment with acetylsalicylic acid (ASA)”. We can observe that the trend estimate of the intervention region “Kinzigtal” is at the upper bound of the trend estimates of the control region. However, several other control regions have only a slightly less pronounced trend, and we obtain consequently a z-score of 1.35 only. The difference of the trend in the intervention region to the mean of the trends in the control regions is 3.27 with a confidence interval from 1.31 to 5.22. This is similar to the difference to the BW region, as the trend in the BW region is rather close to the mean of the trends in the control region

quality, we just change the sign of the estimates. With b_l we denote the lower bound of the 95% confidence interval for $\bar{\Delta}_C$ and with b_u the upper bound. l_{rel} denotes the relevance limit. In case of $\bar{\Delta}_C$ to be positive, we apply the verbal grading shown in Table 1. (The table is to be read sequentially: As soon as a condition is satisfied, the corresponding verbalization is applied.)

Both the condition on the directed estimate as well as the condition on the directed z-score have to be fulfilled. The condition on the z-score, however, makes only sense if there is some evidence for a variation of the true trends across the control regions. Hence the condition is only applied if $\hat{\sigma}_C$ is above 0.001 and if the confidence interval for z is not degenerated (as it can happen when using the Fieller approach). In case $\bar{\Delta}_C$ is negative, we apply the corresponding scheme to define strong negative,

regular negative, or weak negative hints or the choice of the verbalization “inconclusive”.

In the example considered in Figs. 1, 2 and 3 the desired direction is an increase in prevalence. The overall baseline prevalence in the BW sample is 32.0%, hence the relevance limit is 6.8%. The estimated trend difference $\hat{\Delta}_C$ between the intervention region and the control regions is 3.27%, so below the relevance limit. Consequently, the verbalization is “inconclusive”.

Results

Analysing an indicator with a global trend

It is well known that the use of statins in patients with coronary heart disease (CHD) has increased during the study period. This is also visible in Fig. 4 when considering the prevalence of the indicator “Treatment with Statins I” aiming at the prevalence of prescribing statins within the previous 12 months in CHD patients. In the intervention region the increase seems to be more pronounced than in the BW region, and in Fig. 5 we observe that it is also more pronounced than in many control regions. However, there are still several control regions with a similar (or even more pronounced) trend such that the z-score reaches only a value of 1.12 (Fig. 6). Moreover, the relevance limit is here 4.92%, and $\hat{\Delta}_C$ is below this value. Hence this result is verbalized as “inconclusive”.

Table 1 The verbal grading used to classify hints

Verbalization	Trend	Z-score
strong positive hint	$\bar{\Delta}_C > l_{rel}$ AND $b_l(\bar{\Delta}_C) > 0.5 * l_{rel}$	$\bar{z} > 1.96$ AND $b_l(\bar{z}) > 1$
regular positive hint	$\bar{\Delta}_C > l_{rel}$ AND $b_l(\bar{\Delta}_C) > 0$	$\bar{z} > 1.96$
weak positive hint	$\bar{\Delta}_C > l_{rel}$	$\bar{z} > 1$
inconclusive	otherwise	

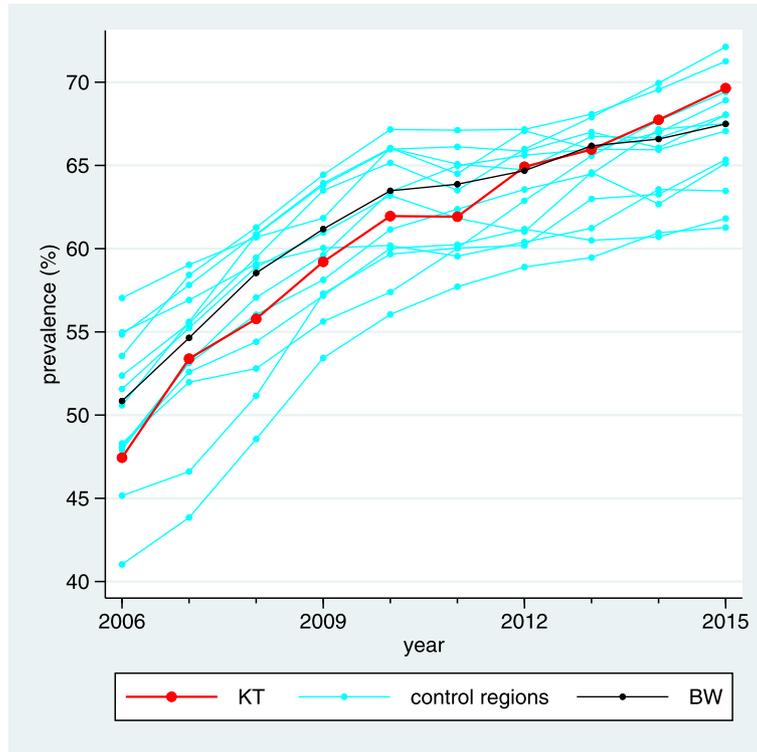


Fig. 4 Visualization of the estimated standardized prevalences in a line plot for the indicator “Treatment with Statins I”

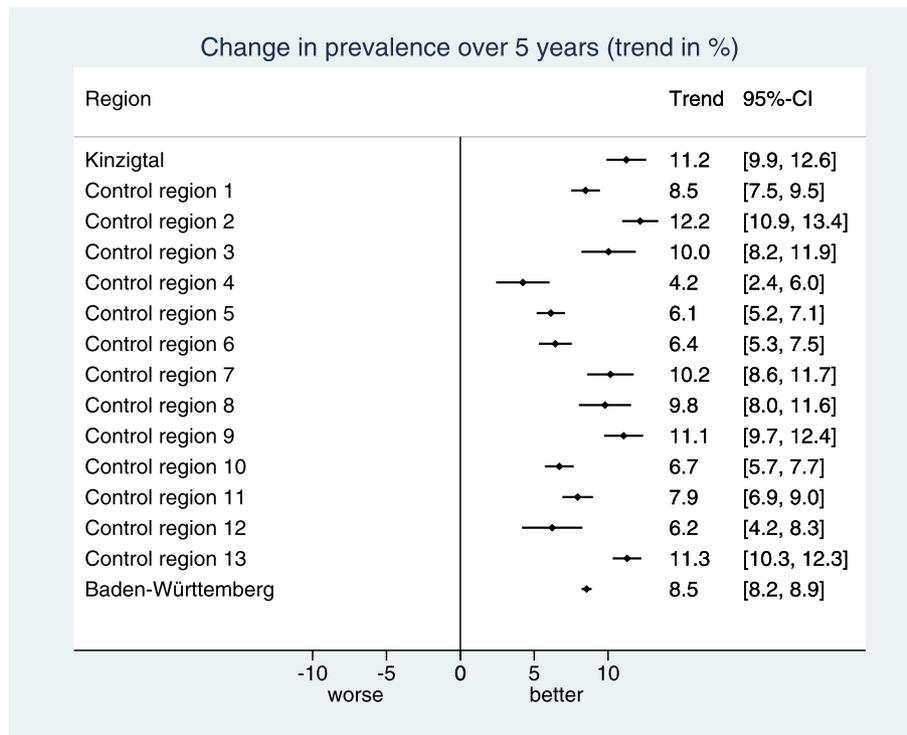
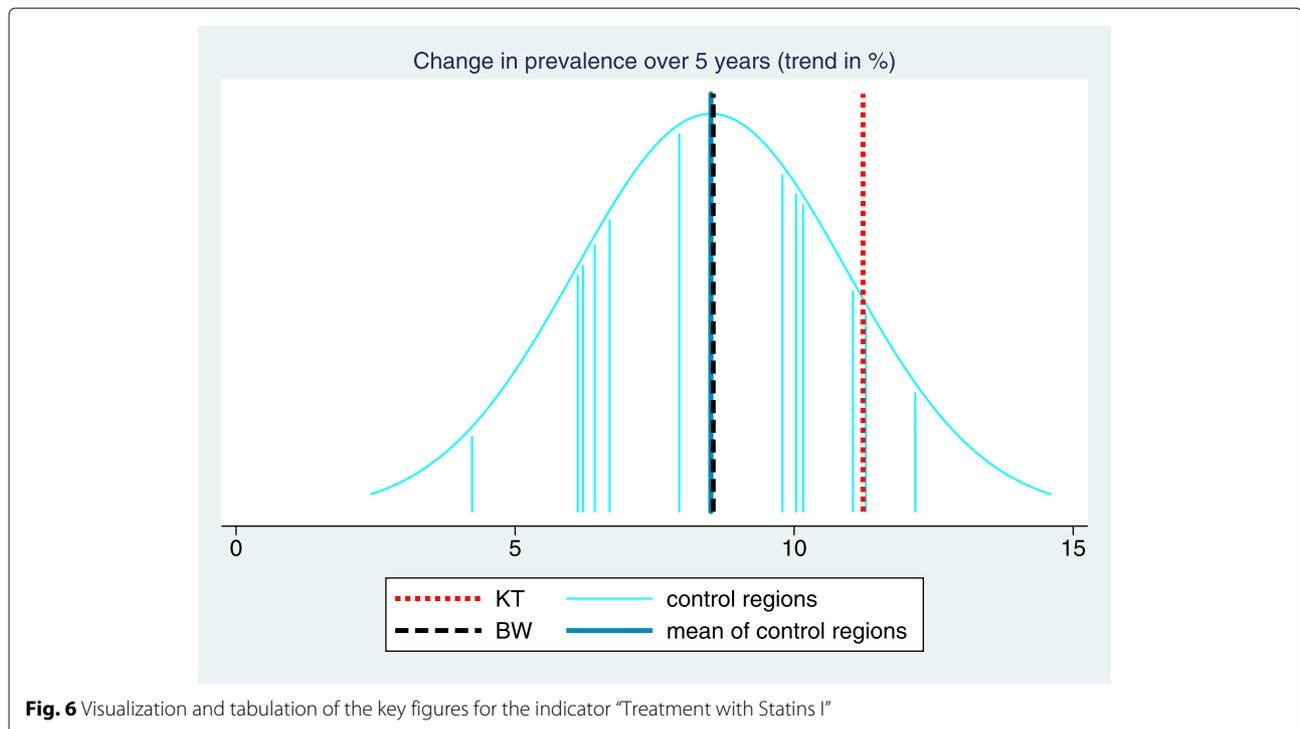


Fig. 5 Visualization of the estimated trends per region in a forest plot for the indicator “Treatment with Statins I”



Analysing an indicator with a change point in the global trend

During the study period, new drugs for the treatment of diabetes type II came into use. Hence an overall trend to more formulary concordant diabetes medication reversed around 2011 into a decreasing trend (Fig. 7). Also the intervention region followed this reversal, but in a less pronounced degree. Consequently, we observe in the intervention region a positive trend in contrast to most control regions (Fig. 8). Actually, the trend is higher than in all control regions (Fig. 9), resulting in a z-score of 1.76. Since the baseline prevalence of this indicator in the BW sample is already as high as 91.4%, the room for improvement is limited. The relevance limit is 0.86, and the lower bound 2.88 of the confidence interval for Δ_C is above this value. Consequently, the requirements for a weak positive hint are fulfilled. However, the z-score is not above 1.96, and hence the requirements for a regular positive hint are not fulfilled.

Analysing an indicator with a structural change

Until October 2013, GP-led geriatric assessments required a corresponding additional certificate of the GP in order to be reimbursed. After this date, all GPs could expect reimbursement if the indication for such an assessment (defined by the presence of certain diagnoses) was given. This implied a substantial change in the corresponding indicator "GP-led geriatric assessment", aiming at the prevalence of this assessment among all eligible patients. This is clearly visible in Fig. 10, and the

intervention region follows this general trend as well as most control regions. Consequently, the trend in the intervention region is within the distribution of the trends of the control regions (Fig. 11), and the verbalization is just "inconclusive".

Assessing the overall evidence

Since the main analytic approach is applied to a large number of indicators, the question how to summarize the results in an adequate manner naturally arises. In particular, the question of the overall evidence for a specific role of the intervention region has to be addressed. In general we can try to summarize the evidence by computing a statistic for each indicator and to sum up these values. We consider here three approaches to compute such a summary statistic S :

- (i) S_{hint} : Strong positive / negative hints are counted as +/- 5 points, regular hints as +/- 3 points, weak hints as +/- 1 points, inconclusive as 0 points, and the average is taken over all directed indicators.
- (ii) S_{diff} : The directed estimates $\bar{\Delta}_C$ are transformed into log odds ratios comparing the two probabilities $\bar{\pi} + \bar{\Delta}_C$ and $\bar{\pi}$, (i.e. $\log(\frac{\bar{\pi} + \bar{\Delta}_C}{1 - (\bar{\pi} + \bar{\Delta}_C)} / \frac{\bar{\pi}}{1 - \bar{\pi}})$) and the average is taken over all directed indicators.
- (iii) $S_{|\hat{z}|}$: The average of $|\hat{z}|$ over all indicators.

The first two approaches aim to investigate whether the results in the intervention region (compared to the control regions) go on average either in the desired or

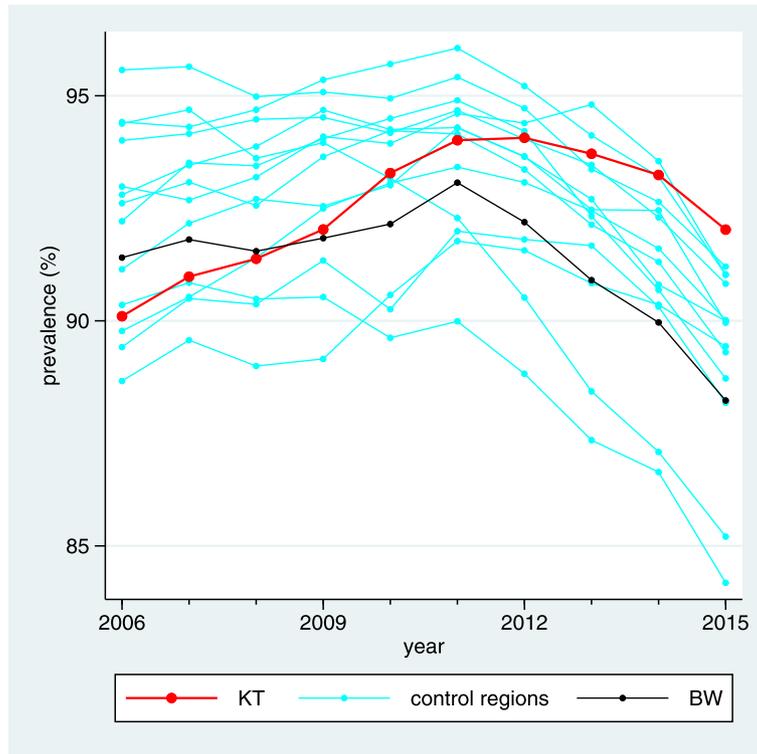


Fig. 7 Visualization of the estimated standardized prevalences in a line plot for the indicator ‘Formulary concordant diabetes medication’

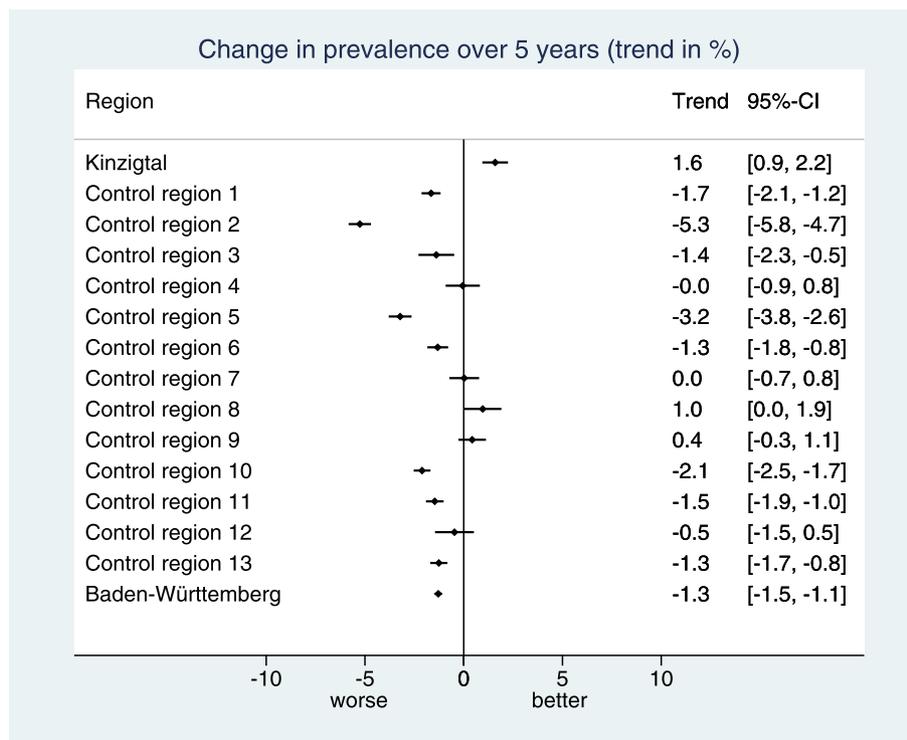


Fig. 8 Visualization of the estimated trends per region in a forest plot for the indicator ‘Formulary concordant diabetes medication’

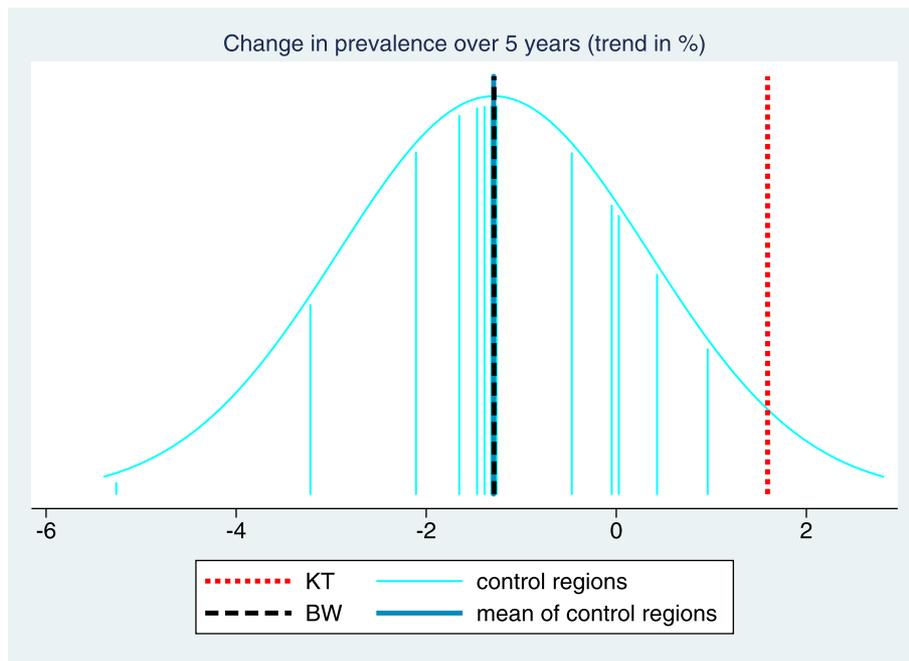


Fig. 9 Visualization and tabulation of the key figures for the indicator "Formulary concordant diabetes medication"

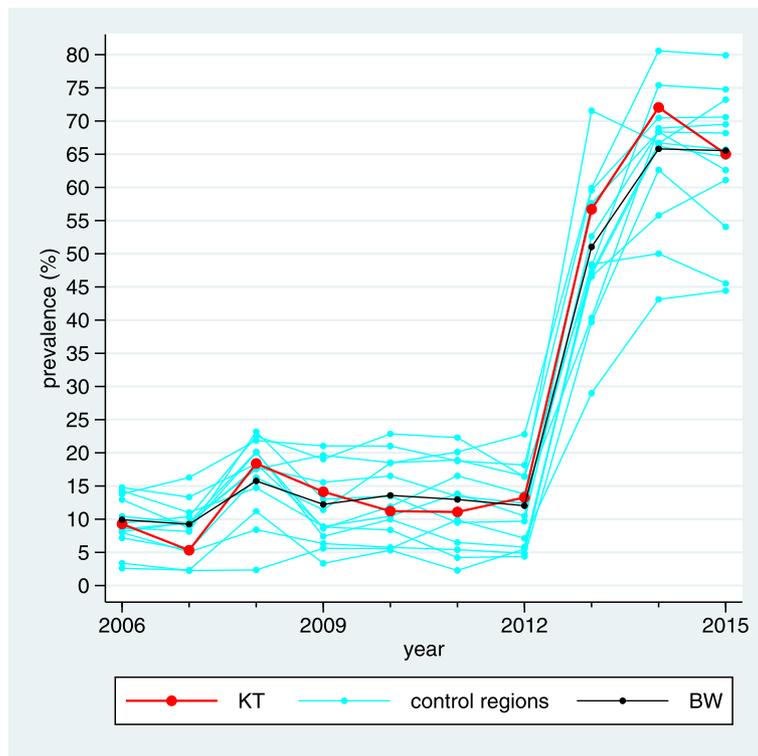


Fig. 10 Visualization of the estimated standardized prevalences in a line plot for the indicator "GP-led geriatric assessment"

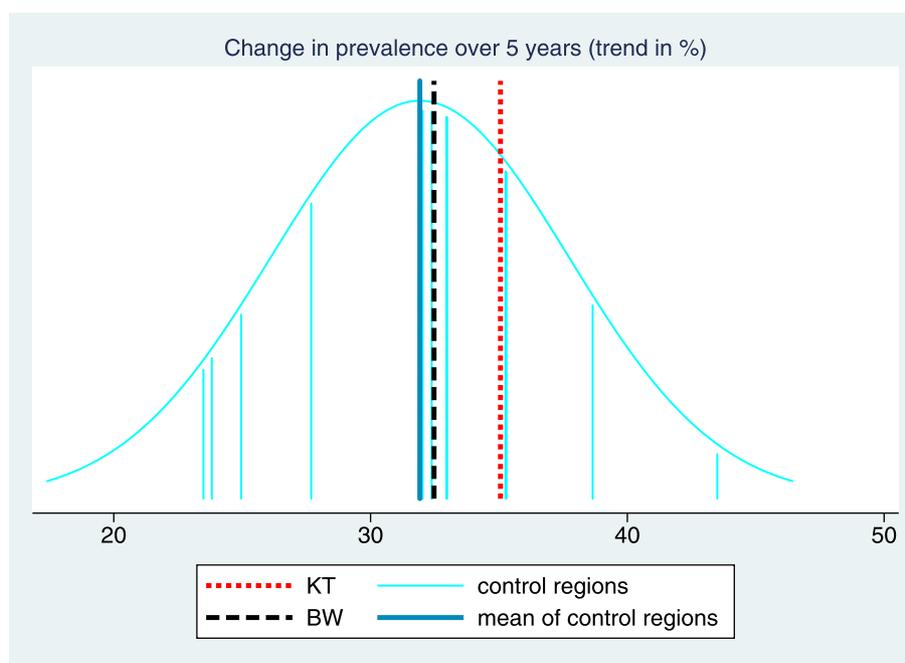


Fig. 11 Visualization and tabulation of the key figures for the indicator “GP-led geriatric assessment”

in the opposite direction. The third approach looks only at whether anything special has happened in the intervention region, independent of whether it goes into the desired or the opposite direction. The approaches can be interpreted as a sequential procedure. The first approach takes the relevance of the differences into account, and requires a rather distinct effect for the single indicators. If this way some evidence for an overall positive or negative effect can be generated, the investigation can be stopped. However, if this way no effect is found, we may still obtain some evidence from the second approach, if there are many indicators with a (small) move into the desired direction – or if the opposite scenario holds. If this also fails, we can reach evidence in the third approach, if there are distinct trend changes (compared to the control regions) in many indicators, possibly in opposite directions. This could be interpreted in the way that the intervention had an overall impact, but sometimes in the desired and sometimes in the undesired direction.

In all three approaches it remains the question how to add inference statements to the computed statistics in a meaningful manner. Computing standard errors is not straightforward, as this has to take into account potential correlations across different indicators. Such correlations are not unlikely, as some indicators are conceptually closely related. Moreover, it is unclear which value the summary statistics should be compared to. For example, if we consider a null hypothesis of the type “There is no difference between the intervention region and the control regions for any indicator” (which can be operationalized

in different manners), the expectation of S needs not to be zero. This follows from the fact that if the prevalence of an indicator is close to 0 or 1 (which has to be expected for a substantial number of indicators), it is not equally likely to obtain a positive or a negative hint of the same degree under such a null hypothesis. Sometimes, the requirements on relevance may even make it impossible to obtain a hint in one direction.

We hence suggest an alternative approach to obtain inferential statements. We can systematically exchange the role of the intervention region with any of the control regions and compute the value of S for any of these R scenarios. If we always observe a less extreme value in these alternative scenarios, then we have some evidence that the intervention region really plays a specific role. Formally, this can be regarded as a permutation test approach and a p -value can be computed by counting how often the originally observed value exceeds or falls below the values under the alternative scenarios. (In the third approach, only exceeding values are of interest.) However, based on $R = 13$ control regions, this method leads to obtaining only a few different possible p -values. In particular, even if the intervention region is far away from all control regions, this does not imply a small p -value. We hence suggest to base the p -value on approximating the distribution of the 13 values under the alternative scenarios by a normal distribution. Subsequently all observed values are presented in a simple plot together with the approximating normal distribution, thus allowing the reader to judge the basis and adequateness

of such a p -value. Figure 12 illustrates this approach using the first summary measure S_{hint} as an example.

This approach can also be used to analyze different subgroups of indicators, e.g. the program specific indicators. However, this rises additional multiplicity issues, which have to be addressed.

Discussion

Sensitivity and subgroup analyses

The overall analytical strategy of a single indicator is rather complex. Consequently, there may be concerns about specific assumptions or decisions made in the analysis process that may have an undesirable impact on the results. This motivates sensitivity analyses to investigate the stability of the results with respect to such assumptions and decisions. We consider in the sequel three different types of analyses that may be useful in this context.

Technical sensitivity

This refers to assumptions and decisions made in the statistical modeling and computations. If such aspects are varied, we expect to see a very small impact on the results. Examples for technical sensitivity are

- (i) Varying the size of the Monte Carlo samples described in the [Appendix](#).
- (ii) Varying the number of knots for the splines used to estimate the effect of age.

- (iii) Consider practices instead of patients as clusters in computing robust standard errors.
- (iv) Using the on-top-trends $\hat{\theta}_r$ as input in the meta analysis.
- (v) Estimating θ_r by a meta regression within each region. These estimates are consistent for the same quantities, but less efficient, as variation due to an overall trend is not modeled. On the other hand, this approach allows a region specific error variance.

It would also be of interest to combine the different analytical steps described above into one model. In principle, this can be accomplished by incorporating the structure assumed for p_{rt} and for θ_r , respectively, into the logistic model used to compute the standardized prevalences. In other words, α_{Rit} is replaced with a random effects model for the trend. However, the parameters of such a model would refer to probabilities on the logit scale, and not to probabilities on a probability scale. Hence, they would be conceptually different. Alternatively, a generalized linear model with Bernoulli variance and identity link may be used. This may, however, be problematic with respect to modeling covariate effects correctly. In any case, it is not completely straightforward to combine the different steps in one model.

Conceptual sensitivity

This refers to approaches that use the available data in a slightly different way. There is still the expectation to

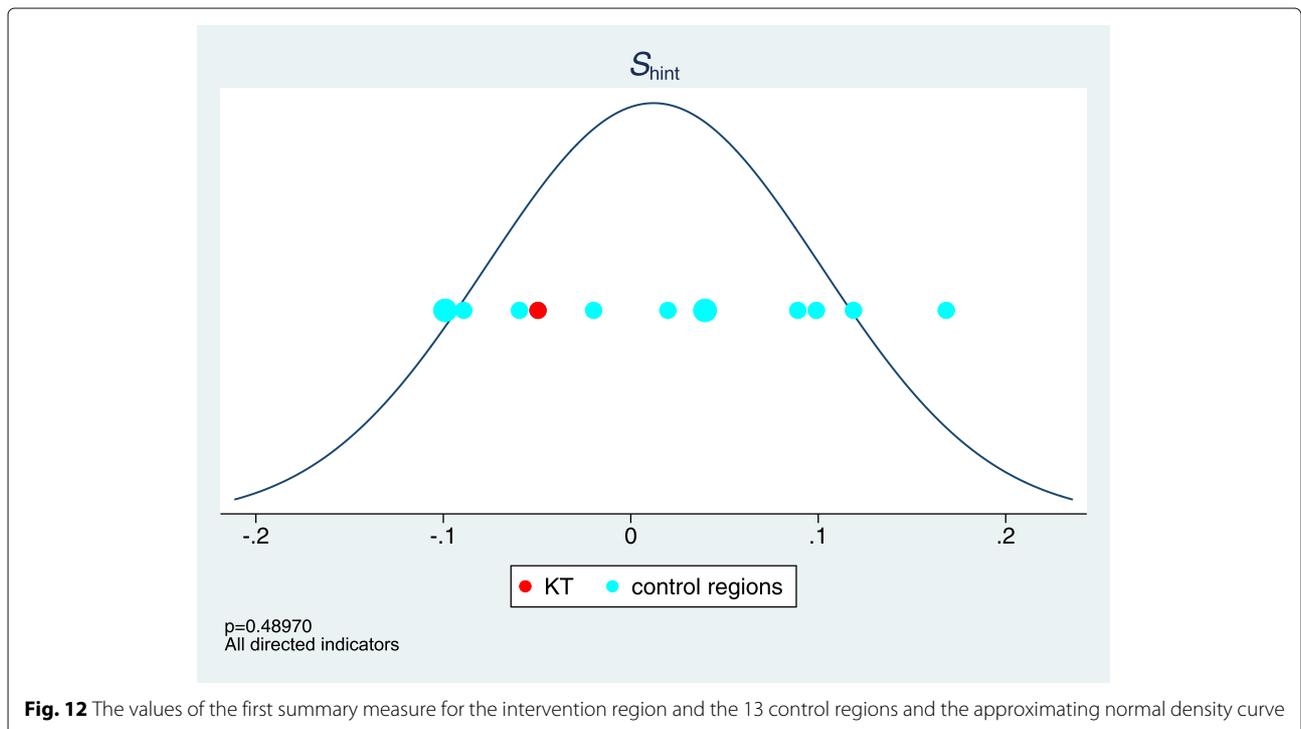


Fig. 12 The values of the first summary measure for the intervention region and the 13 control regions and the approximating normal density curve

get similar results, which may be even more adequate. Examples for conceptual sensitivity analyses are:

- (i) Restricting the analysis period to the first five years. In case the effect of the intervention was saturated after some time, this may result in more distinct effects. Otherwise, we would just lose efficiency due to a lower sample size.
- (ii) Removing practices with prevalences already close to the desired prevalence over the whole period. Patients treated in these practices cannot improve any further, and this may mask intervention effects.
- (iii) Using the intervention region instead of the BW sample for standardization. This may make the intervention effect more visible, if the intervention is best to a population close in composition to that of the intervention region.
- (iv) Restricting the control regions to those with a regional physicians' network. This comparison might be more fair, but less efficient.

Subgroup analyses

Similar results in different subgroups of patients may underline that the intervention is really effective for all subjects. Differing results may inform us that the intervention is particularly helpful in certain subgroups. However, additional multiplicity issues may arise. Candidate variables for subgroup analyses are the confounders already available in all subjects, i.e. age, gender, comorbidity and SES.

Identifying indicators with high likelihood for a specific role of the intervention region

When considering so many indicators, it is desirable to focus on the interpretation of those indicators with the strongest positive or strongest negative intervention effects, respectively. It is well known that just ordering the effects by size is a poor approach [15], as this invites over-interpretation of extreme results. In the last decades there has been substantial progress in developing methods to identify signals of interest in a more reliable manner based on the fact that a large number of indicators allows the estimation of the variation of the true effects. Consequently, the posterior probability to be above the clinical relevant threshold (cf. [16]) or characteristics of the posterior distribution of the rank can be considered [17] for each effect. However, the necessary computations are complicated by the fact that the effect estimates are not independent.

Assessing potential acceleration and deceleration

It is one of the overall questions of the project, whether the effects of the interventions are long lasting. Some effects may vanish over time or they may become more distinct

after some time. Investigating the stability of the trend may give an idea about this.

A simple approach is to consider within the intervention region estimates for the on-top-trend $\tilde{\theta}_{KT}^E$ in the first half of the time period, for the on-top-trend $\tilde{\theta}_{KT}^L$ in the second half of the time period, and for the difference $\Delta_{KT} = \tilde{\theta}_{KT}^L - \tilde{\theta}_{KT}^E$ as an expression of the acceleration/deceleration. Estimates for these quantities can be reported together with confidence intervals and p -values.

Estimates of these quantities can be easily obtained by replacing in model (1) the linear trend in the intervention region by a linear spline with a knot exactly at the middle of the observation period. This would again take into account that there might also be a change in the overall trend.

A verbal classification of the results is here more challenging, as there are a lot of possible patterns such as *vanishing*, *reversal*, *late start*, *acceleration* or *deceleration*. Furthermore, it is not straightforward to define rules to distinguish these patterns, in particular if relevance should also be taken into account.

Assessing inter-practice variation

Improving quality indicators at the population level is the ultimate aim of any intervention of the type considered in this project. However, we can also investigate other aspects that may inform us how the intervention has (or has not) reached this aim. One of these aspects is the inter-practice variation. A high inter-practice variation is often regarded as an indicator of poor quality, as this may reflect a fundamental dissent among the health care providers about the optimal management of patients [18, 19]. Reducing inter-practice variation is hence also an aim of an integrated care model. However, such a reduction does not automatically imply an improvement in care; practices may also converge to a poor management strategy.

In any case it is rather simple to analyse trends in inter-practice variation with a similar analytic strategy. We just have to replace the estimates \hat{p}_{rt} for the standardized prevalences in each region and year by estimates of the inter-practice variation. These can be easily obtained by fitting a logistic regression model with the practice as random intercept in each region and year with adjustment for confounders at the individual level. Thus, differences in the patient populations across practices are taken into account. Furthermore, by comparison with the control regions it can be taken into account that increasing or decreasing inter-practice variation in a region may also happen without an intervention.

Inter-practice variation can also occur with respect to the trends in the intervention region. Some practices may improve fast and some slow. This can be analysed with models including random intercepts and random slopes

for each practice fitted to the data from the intervention region.

Incorporating differences at baseline

It is natural that the prevalence of an indicator moves up and down within a region to some degree over time, either due to random variation or to temporal changes in the average patient management. Therefore, the intervention region may by chance have a relatively high or a rather low prevalence level at the baseline year $t = 0$. This may lead to an increasing or decreasing trend over time, either due to regression to the mean or due to specific measures taken in the region to address the issue reflected by an indicator. In both cases a comparison with all control regions can become unfair, and it is desirable to compare the intervention region only to control regions with a similar level at baseline.

In the model (1) the parameter β_r reflects this baseline level for region r . So a naive implementation of this idea would be to restrict the comparison to control regions with an estimated baseline level $\hat{\beta}_r$ similar to $\hat{\beta}_{KT}$. This requires, however, defining some threshold distinguishing *similar* from *not similar*. This can be avoided by assuming a joint normal distribution for $(\beta_r, \theta_r)_{r=0, \dots, R}$ and to use this to predict θ_r based on β_r . Then μ_C can be replaced by the predicted trend for a control region with a baseline value identical to that of the intervention region.

Using automatically generated indicators

One limitation of the project is the restriction to a preselected albeit large set of quality indicators. These indicators may not cover all aspects of the health care system and hence there is a risk to overlook important trends. On the other hand, the claims data includes information on diagnoses, prescriptions, and medical procedures. They are based on well-established coding systems such as ICD, ATC, OPS, or EBM [20–23]. Hence we can also define indicators by defining events according to making a certain diagnosis, to prescribing a specific drug, or to receiving a certain medical procedure. Moreover, the hierarchical structure of the coding systems allows us to define meaningful groups of diagnoses, drugs or procedures, e.g. drugs with the same therapeutic intention. Hence thousands of events of interest can be defined for which time trends can indicate a change in patient management. However, the crucial point is to define the populations of interest. Just taking the whole population is possible and valid, but this may imply a lack of power in detecting trends. In principle, it is possible to also define risk populations in an automatic manner. For example, for a certain drug or drug group, we can identify all diagnoses that typically appear close to the prescriptions, and can define the population at risk as those suffering from such a diagnosis. Such an approach would be similar to exist-

ing data mining techniques in pharmaco-epidemiological data bases covering spontaneous reports of adverse events [24, 25].

Code development

The analysis to be performed for a single indicator is rather complex. It involves the application of several statistical procedures in a sequential manner using the output of one method as input for the next method. The simultaneous application on 106 indicators (and the use of resampling procedures) makes it impossible to check the validity of each application by inspecting outputs and log files. Consequently, it was essential to develop code that was both robust and valid. The approach was hence split further into very small steps reflecting the application of single statistical procedures or certain data management actions. Each step was tested using input for which the desired output could be determined by independent means. A collection of such tests could be executed and compared to previous outputs in an automatic manner, allowing us to test the whole code after each development cycle.

The overall code could not be tested this way, as the desired output could not be determined by independent means. We hence generated data sets according to our overall model and tested the consistency of the resulting estimates by applying these to very large datasets. Furthermore, we explored the validity of the inference by checking the coverage of confidence intervals in simulation studies.

Reporting

The application on 106 indicators (including descriptive tables to report basic properties of the data) and the systematic conduct of sensitivity analyses resulted in a huge amount of information. We hence generated a series of automated reports addressing different levels of interest. A first report (with nearly 10,000 pages) presented several items for each indicator. First, a summary of the results from the main analysis and all sensitivity analyses. Second, the complete results including some descriptive tables of the raw prevalences and the distribution of the covariates (stratified by region, year and both). Further reports focused on the acceleration or deceleration of trends, the analysis of the overall evidence, or provided an overview of all indicator specific results using tables and graphs presenting the numbers generated in a manner facilitating comparisons of interest. All these reports are available for the public (on request) at <https://www.pmvforschungsgruppe.de/projekte/integral.html>.

Conclusions

The methodological approach developed in this project addresses general challenges in the evaluation of integrated care models such as accountable care organiza-

tions. This latter model for delivery and payment of health care has attracted much attention in the US [26] but also in Europe [27]. Recommendations for the evaluation of accountable care organizations were already developed some years ago [28, 29]. Since randomization is rarely feasible when introducing accountable care organizations, comparing patient groups exposed and unexposed to the new model in an observational pre-post design with adjustment for potential confounders can be regarded as providing the best available evidence [26]. This approach can be also called a difference-in-difference approach [30, 31]. The methodology developed for this project can also be seen as a difference-in-difference approach with adjustment for potential confounders. However, we consider the specific situation of evaluating one region-specific accountable care organization. This specific situation made it necessary to take into account that regional time trends can happen by chance. A comparison with several control regions lacking such an organization made is possible to tackle this challenge.

At first sight, the idea to compare prevalence time trends in an intervention region with corresponding trends in control regions sounds like a rather standard epidemiological analysis. However, implementation of this idea was rather complex, when taking into account the need for adjustment and the need to assess a specific role of the intervention region. We believe that the approach presented here can handle the major statistical challenges: population differences (also over time) are addressed by standardization; we offer transparency with respect to the derivation of the key figures; global time trends and structural changes do not invalidate the analyses; the regional variation in time trends is taken into account in the final judgement of the intervention effect. The latter three points are hopefully well illustrated by our examples. And it is worth mentioning that we could indeed observe a regional variation in time trends for the majority of indicators.

The simultaneous application on over one hundred indicators added further complexity, both computationally and conceptually. Overall, the project demanded substantial intellectual as well as organizational efforts to ensure adequateness, validity and transparency – as described in this paper – just to prepare the first main publication. Funders and scientists may tend to prioritize projects that can address the question of interest within a fully established methodological framework and an existing computational environment, such that results and publications can be generated with limited efforts. However, certain questions of high relevance such as the one addressed in this project require more efforts, and we should continue to plan and execute projects with an advanced methodological complexity.

Appendix: Monte Carlo approximation of the standard errors of the standardized prevalences

To obtain the standard errors of \hat{p}_{rt} , we make use of the fact that we can express \hat{p}_{rt} as

$$\hat{p}_{rt} = \frac{1}{|S_t^{BW}|} \sum_{i \in S_t^{BW}} \Lambda \left(\sum_l \hat{\beta}_l z_{il} \right),$$

ignoring in the notation that β_l may depend on r and t and z_{il} may depend on t . Consequently, we have

$$\text{Var}(\hat{p}_{rt}) = \frac{1}{|S_t^{BW}|^2} \left(\sum_{i \in S_t^{BW}} v_i + \sum_{i \neq i'} c_{ii'} \right),$$

with

$$v_i = \text{Var} \Lambda \left(\sum_l \hat{\beta}_l z_{il} \right) \approx l_i^2 \cdot \tilde{v}_i \quad (\text{Delta method: Var}f(X) \approx \text{Var}(X) \cdot f'(X)^2)$$

$$l_i = \Lambda' \left(\sum_l \hat{\beta}_l z_{il} \right) \quad \text{with } \Lambda'(\cdot) = \Lambda(\cdot)[1 - \Lambda(\cdot)]$$

$$\tilde{v}_i = \text{Var} \left(\sum_l \hat{\beta}_l z_{il} \right) = \sum_{l, l'} C_{ll'} z_{il} z_{il'}$$

and

$$\begin{aligned} c_{ii'} &= \text{Cov} \left(\Lambda \left(\sum_l \hat{\beta}_l z_{il} \right), \Lambda \left(\sum_l \hat{\beta}_l z_{i'l} \right) \right) \\ &\approx l_i l_{i'} \cdot \text{Cov} \left(\sum_l \hat{\beta}_l z_{il}, \sum_l \hat{\beta}_l z_{i'l} \right) \quad (\text{Delta method}) \\ &= l_i l_{i'} \sum_{l, l'} C_{ll'} z_{il} z_{i'l'} \end{aligned}$$

where $C_{ll'}$ denotes the covariance between $\hat{\beta}_l$ and $\hat{\beta}_{l'}$, and we can obtain an estimate for $\text{Var}(\hat{p}_{rt})$ by plugging in the estimated covariances.

The full enumeration of all pairs of subjects is, however, cumbersome, so we make use of a Monte-Carlo approximation, based on expressing the variance of \hat{p}_{rt} as

$$\frac{1}{|S_t^{BW}|} \left(\frac{1}{|S_t^{BW}|} \sum_i v_i + (|S_t^{BW}| - 1) \frac{1}{(|S_t^{BW}| - 1) |S_t^{BW}|} \sum_{i, i'} c_{ii'} \right),$$

allowing us to replace the two averages by averages based on subsamples. So we draw a subsample S_1 of size n_1 of S_t^{BW} and approximate the first average by

$$\frac{1}{|S_t^{BW}|} \sum_i v_i \approx \frac{1}{n_1} \sum_{i \in S_1} v_i$$

and a subsample S_2 of size n_2 of order pairs of S_t^{BW} , and approximate the second average by

$$\frac{1}{(|S_t^{BW}| - 1) |S_t^{BW}|} \sum_{i,i'} c_{ii'} \approx \frac{1}{n_2} \sum_{(i_1, i_2) \in S_2} c_{i_1 i_2}.$$

We choose $n_1 = \min(n, 10000)$ and $n_2 = \min(\lfloor n/2 \rfloor, 20000)$, with $n = |S_t^{BW}|$.

Abbreviations

AOK: Allgemeine Ortskrankenkasse (health insurance fund); ASA: acetylsalicylic acid; BW: Baden-Württemberg (federal state in Germany); CHD: coronary heart disease; CI: confidence interval; DRKS: Deutsches Register Klinischer Studien (German Clinical Trials Register); GISD: German Index of Socioeconomic Deprivation; GP: general practitioner; KT: Kinzigtal (region in BW); OLS: ordinary least squares; SES: socio-economic status; WIdO: Wissenschaftliches Institut der AOK (AOK research institute).

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12913-022-07573-7>.

Additional file 1: The selection of structurally similar control regions for the intervention region in the INTEGRAL study. This pdf document includes additional information on the selection of the control regions.

Additional file 2: Constructing the 10% sample Baden-Württemberg (BW) and defining study populations. This pdf document included additional information on the process of constructing the 10% sample and the definition of study populations.

Acknowledgements

The authors are grateful to Derek Hazard for support on language and style. The authors acknowledge the permission of the Deutsche Ärzteblattverlag for reusing the Figures 7 and 12 previously published in [2].

Authors' contributions

DS, EG and WW developed the statistical methodology. DS implemented the statistical methodology and performed all analyses. AS and WV selected the control regions. AK, AS, CG, CM, DS, EF, EG, IK, IS, MG, PD, PI and WV participated in the development of the general research strategy. WV drafted the first version of the manuscript. All authors participated in editing the final version and approved it.

Funding

The INTEGRAL research project is funded by the Innovation Committee of the Joint Federal Committee under the grant number 01VSF16002. Open Access funding enabled and organized by Projekt DEAL.

Availability of data and materials

The raw data of the project is not available due to data protection reasons.

Declarations

Ethics approval and consent to participate

The INTEGRAL project has been approved by the Ethic Commission of the Faculty of Medicine, Philipps-University Marburg (ek_mr_geraedts_131117). Permission to use the data has been granted by the Research Institute of the Local Health Care Funds (WIdO). The project has been registered at the German Clinical Trials Register (DRKS) at the registration number DRKS00012804.

Consent for publication

Not applicable.

Competing interests

AS declares involvement in former studies on Gesundes Kinzigtal GmbH (2006–2015) and an employment at Gesundes Kinzigtal GmbH (1 June 2015 until 31 December 2015). IS, IK and PI declare that they were involved in one

former study evaluating the start-up phase (2006–2011) of the integrated care model 'Gesundes Kinzigtal'. All authors report grants from the Innovation Committee of the Joint Federal Committee, during the conduct of the study.

Author details

¹Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center, University of Freiburg, Freiburg, Germany. ²PMV research group at the Department of Child and Adolescent Psychiatry, Psychotherapy and Psychosomatics, University of Cologne, Köln, Germany. ³Health Services and Quality Research, Research Institute of the Local Health Care Funds (WIdO), Berlin, Germany. ⁴Institute for Health Services Research and Clinical Epidemiology, University of Marburg, Marburg, Germany. ⁵Institute of Occupational and Social Medicine and Health Services Research, University of Tübingen, Tübingen, Germany. ⁶Basel Academy for Quality and Research in Medicine, Basel, Switzerland.

Received: 30 June 2021 Accepted: 1 February 2022

Published online: 24 February 2022

References

1. Schubert I, Siegel A, Graf E, Farin-Glattacker E, Ihle P, Köster I, Stelzer D, Mehl C, Schmitz J, Dröge P, Günster C, Klöss A, Vach W, Geraedts M. Study protocol for a quasi-experimental claims-based study evaluating 10-year results of the population-based integrated healthcare model 'Gesundes Kinzigtal' (Healthy Kinzigtal): the INTEGRAL study. *BMJ Open*. 2019;9(1):e025945. <https://doi.org/10.1136/bmjopen-2018-025945>.
2. Schubert I, Stelzer D, Siegel A, Köster I, Mehl C, Ihle P, Günster C, Dröge P, Klöss A, Farin-Glattacker E, Graf E, Geraedts M. Ten-Year Evaluation of the Population-Based Integrated Health Care System "Gesundes Kinzigtal". *Deutsches Ärzteblatt Int*. 2021;18(27-28):465–72. <https://doi.org/10.3238/arztebl.m2021.0163>.
3. Salm M, Wübker A. Sources of regional variation in healthcare utilization in Germany. *J Health Econ*. 2020;69:102271. <https://doi.org/10.1016/j.jhealeco.2019.102271>.
4. Geraedts M, Mehl C, Schmitz J, Siegel A, Graf E, Stelzer D, Farin-Glattacker E, Ihle P, Köster I, Dröge P, Günster C, Haas N, Gröne O, Schubert I. Entwicklung eines Indikatorensets zur Evaluation der Integrierten Versorgung Gesundes Kinzigtal. *Z Evidenz Fortbild Qualität Gesundheitswesen*. 2020;150-152:54–64. <https://doi.org/10.1016/j.zefq.2020.04.001>.
5. Köster I, Ihle P, Schubert I. Bericht Innovationsfonds – Projekt: INTEGRAL – AP 4: Operationalisierung der Indikatoren mittels Routinedaten und Deskription. 2019. Teil A: Material und Methode. Internal publication of the PMV Forschungsgruppe, Universität zu Köln.
6. Hildebrandt H, Hermann C, Knittel R, Richter-Reichhelm M, Siegel A, Witzenthath W. Gesundes Kinzigtal Integrated Care: improving population health by a shared health gain approach and a shared savings contract. *Int J Integr Care*. 2010;10:e046. <https://doi.org/10.5334/ijic.539>.
7. Inskip H, Beral V, Fraser P, Haskey J. Methods for age-adjustment of rates. *Stat Med*. 1983;2(4):455–66. <https://doi.org/10.1002/sim.4780020404>.
8. Roalfe AK, Holder RL, Wilson S. Standardisation of rates using logistic regression: a comparison with the direct method. *BMC Health Serv Res*. 2008;8:275. <https://doi.org/10.1186/1472-6963-8-275>.
9. Kroll LE. German index of socioeconomic deprivation (GISD) version 1.0 (version: 1.0.09). 2017. <http://doi.org/10.7802/1460>.
10. Kroll LE, Schumann M, Hoebel J, Lampert T. Regional health differences – Developing a socioeconomic deprivation index for Germany. *J Health Monit*. 2017;2(2):98–114. <https://doi.org/10.17886/RKI-GBE-2017-048.2>.
11. Dreier D, Blagorazumnaya O, Balicer R, Dreier J. National initiatives to promote quality of care and patient safety: achievements to date and challenges ahead. *Isr J Health Policy Res*. 2020;9:62. <https://doi.org/10.1186/s13584-020-00417-x>.
12. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Control Clin Trials*. 1986;7:177–88. [https://doi.org/10.1016/0197-2456\(86\)90046-2](https://doi.org/10.1016/0197-2456(86)90046-2).
13. Schubert I, Siegel A, Köster I, Ihle P. Evaluation of the population-based Integrated Health Care System Gesundes Kinzigtal (IHGK). Findings on health care quality based on administrative data. *Z Evidenz Fortbild Qualität Gesundheitswesen*. 2016;117:27–37. <https://doi.org/10.1016/j.zefq.2016.06.003>.

14. Fieller EC. Some problems in interval estimation. *J R Stat Soc Ser B (Methodol)*. 1954;16(2):75–185. <https://doi.org/10.1111/j.2517-6161.1954.tb00159.x>.
15. Goldstein H, Spiegelhalter DJ. League tables and their limitations: Statistical issues in comparisons of institutional performance. *J R Stat Soc Ser A (Stat Soc)*. 1996;159(3):385–409. <https://doi.org/10.2307/2983325>.
16. Diggle P, Moraga P, Rowlingson B, Taylor BM. Spatial and spatio-temporal log-gaussian cox processes: Extending the geostatistical paradigm. *Stat Sci*. 2013;28(4):542–563. <https://doi.org/10.1214/13-STS441>.
17. Laird NM, Louis TA. Empirical Bayes ranking methods. *J Educ Stat*. 1989;14(1):29–46. <https://doi.org/10.3102/10769986014001029>.
18. Davis P, Gribben B, Scott A, Lay-Yee R. The “supply hypothesis” and medical practice variation in primary care: testing economic and clinical models of inter-practitioner variation. *Soc Sci Med*. 2000;50(3):407–18. [https://doi.org/10.1016/s0277-9536\(99\)00299-3](https://doi.org/10.1016/s0277-9536(99)00299-3).
19. Corallo AN, Croxford R, Goodman DC, Bryan EL, Srivastava D, Stukel TA. A systematic review of medical practice variation in OECD countries. *Health Policy*. 2013;114(1):5–14. <https://doi.org/10.1016/j.healthpol.2013.08.002>.
20. DIMDI. ICD-10-GM Version 2012. Internationale statistische Klassifikation der Krankheiten und verwandter Gesundheitsprobleme. 10. Revision. German Modification. Version 2012. 2012a. <https://www.dimdi.de/static/de/klassifikationen/icd/icd-10-gm/kode-suche/htmlgm2012>. Accessed 10 Feb 2022.
21. DIMDI. Operationen- und Prozedurschlüssel. Version 2012. 2012b. <https://www.dimdi.de/static/de/klassifikationen/ops/kode-suche/opshtml2012>. Accessed 10 Feb 2022.
22. DIMDI. Anatomisch-therapeutisch-chemische Klassifikation mit Tagesdosen. Amtliche Fassung des ATC-Index mit DDD-Angaben für Deutschland im Jahre 2012. 2012c. <https://www.dimdi.de/dynamic/downloads/arszneimittel/atcddd/atc-ddd-amtlich-2012.pdf>. Accessed 10 Feb 2022.
23. Kassenärztliche Bundesvereinigung. Einheitlicher Bewertungsmaßstab (EBM). 2020. <https://www.kbv.de/html/ebm.php>. Accessed 10 Feb 2022.
24. Stephenson WP, Hauben M. Data mining for signals in spontaneous reporting databases: proceed with caution. *Pharmacoepidemiol Drug Saf*. 2017;16(4). <https://doi.org/10.1002/pds.1323>.
25. Harpaz R, DuMouchel W, Shah NH, Madigan D, Ryan P, Friedman C. Novel Data-Mining Methodologies for Adverse Drug Event Discovery and Analysis. *Clinical Pharmacology & Therapeutics*. 2012;91(6):1010–21. <https://doi.org/10.1038/clpt.2012.50>.
26. Kaufman BG, Spivack BS, Stearns SC, Song PH, O'Brien EC. Impact of Accountable Care Organizations on Utilization, Care, and Outcomes: A Systematic Review. *Med Care Res Rev*. 2019;76(3):255–290. <https://doi.org/10.1177/1077558717745916>.
27. McClellan M, Udayakumar K, Thoumi A, Gonzalez-Smith J, Kadakia K, Kurek N, Abdulmalik M, Darzi AW. Improving Care And Lowering Costs: Evidence And Lessons From A Global Analysis Of Accountable Care Reforms. *Health Affairs*. 2017;36(11):1920–7. <https://doi.org/10.1377/hlthaff.2017.0535>.
28. Fisher ES, Shortell SM, Kreindler SA, Van Citters AD, Larson BK. A framework for evaluating the formation, implementation, and performance of accountable care organizations. *Health Aff*. 2012;31(11):2368–78. <https://doi.org/10.1377/hlthaff.2012.0544>.
29. Damberg CL, Sorbero ME, Lovejoy SL, Martolf GR, Raaen L, Mandel D. Measuring Success in Health Care Value-Based Purchasing Programs: Findings from an Environmental Scan, Literature Review, and Expert Panel Discussions. *Rand Health Q*. 2014;4(3):9.
30. Morciano M, Checkland K, Billings J, Coleman A, Stokes J, Tallack C, Sutton M. New integrated care models in England associated with small reduction in hospital admissions in longer-term: A difference-in-differences analysis. *Health Policy*. 2020;24(8):826–33. <https://doi.org/10.1016/j.healthpol.2020.06.004>.
31. Coe NB, Ingraham B, Albertson E, Zhou L, Wood S, Grembowski D, Conrad D. The one-year impact of accountable care networks among Washington State employees. *Health Serv Res*. 2021;56(4):604–14. <https://doi.org/10.1111/1475-6773.13656>.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

