Why do we use R^2 so often in risk equalization research?

EsCHER Working Paper No. 2024001
16 January 2024

Wynand P. M. M. van de Ven, PhD
Richard C. van Kleef, PhD

Erasmus University Rotterdam
Making Minds Matter

Title
Why do we use R^2 so often in risk equalization research?

Authors
Wynand P.M.M. van de Ven [a] and Richard C. van Kleef [a]
[a] Erasmus University Rotterdam / Erasmus Centre for Health Economics Rotterdam (EsCHER)

Corresponding author: Wynand P.M.M. van de Ven (vandeven@eshpm.eur.nl)

Keywords
R^2, risk equalization, health insurance, Mean Absolute Prediction Error (MAPE), Cumming's Prediction Measure (CPM), Payment System Fit (PSF).

JEL classification
C10, G22, I11, I13.

Cite as
Van de Ven, W.P.M.M. and R.C. van Kleef (2024). Why do we use R^2 so often in risk equalization research? EsCHER Working Paper Series No. 2024001, Erasmus University Rotterdam. Available from: https://www.eur.nl/en/research/research-groups-initiatives/escher/research/working-papers

Erasmus Centre for Health Economics Rotterdam (EsCHER) is part of Erasmus University Rotterdam.
Want to know more about EsCHER? Visit www.eur.nl/escher
Want to contact EsCHER? E-mail escher@eur.nl

Interested in more EsCHER Working Papers? Download from www.eur.nl/escher/research/workingpapers

WHY DO WE USE R^2 SO OFTEN IN RISK EQUALIZATION RESEARCH?

16jan24

Wynand P.M.M. van de Ven [a] and Richard C. van Kleef [a]
[a] Erasmus University Rotterdam / Erasmus Centre for Health Economics Rotterdam (EsCHER)

Abstract

Nearly all empirical studies that estimate the coefficients of a risk equalization formula present the value of the statistical measure R^2. The R^2-value is often (implicitly) interpreted as a measure of the extent to which the risk equalization payments remove the regulation-induced predictable profits and losses on the insured, with a higher R^2-value indicating a better performance. In many cases, however, we do not know whether a model with $R^2 = 0.30$ reduces the predictable profits and losses more than a model with $R^2 = 0.20$.

In this paper we argue that in the context of risk equalization R^2 is ***hard to interpret, can lead to wrong and misleading conclusions*** *when used as a measure of selection incentives, and is therefore* ***not useful*** *for measuring selection incentives. The same is true for related statistical measures such as the Mean Absolute Prediction Error (MAPE), Cumming's Prediction Measure (CPM) and the Payment System Fit (PSF). There are some exceptions where the R^2 can be useful.*

Our recommendation is to either present the R^2 with a clear, valid interpretation or not to present the R^2. The same holds for the related statistical measures MAPE, CPM and PSF.

Key words: R^2, risk equalization, health insurance, Mean Absolute Prediction Error (MAPE), Cumming's Prediction Measure (CPM), Payment System Fit (PSF).

JEL Classification: C10, G22, I11, I13.

1. Introduction

Regulated health insurance markets are often characterized by premium regulation and open enrolment. Although these regulations can help improve the affordability and accessibility of health insurance coverage for high-risk people, they also create predictable profits and losses on insured, and therefore provide insurers with incentives for risk selection. To prevent the adverse effects of risk selection, a risk equalization scheme can be implemented. That is a system of risk adjusted equalization payments to and from the insurers that compensate insurers for differences in individuals' expected healthcare expenses that are not allowed to be explicitly reflected in premium differences. The goal of risk equalization is to eliminate the regulation-induced incentives for risk selection[1] by compensating insurers for predictable profits and losses (Van de Ven et al., 2023).[2] Therefore, an important aspect of the evaluation of risk equalization formulas is the extent to which the risk equalization payments remove the regulation-induced *predictable* profits and losses on the insured.

With *predictable* profits and losses we mean the profits and losses that, given the premium regulation, can be predicted by consumers and/or insurers prior to the contract period. That is, there is an information asymmetry in the sense that consumers and/or insurers have more predictive information about future healthcare expenses than the information that is used for calculating the equalization payments. Insurers have rich panel data that they can use to make better predictions of the future individual healthcare expenses than the predictions based on imperfect risk equalization. Insurers can use their information surplus for risk selection, which can have negative effects in terms of quality of care, efficiency, and affordability (see e.g., Van de Ven and Ellis, 2000). Insurers can learn about predictable profits and losses in subtle ways. For example, any correlation between consumers' preferences regarding health insurance and their health risk can reveal a predictable profit or loss. When insurers know these correlations (e.g., based on revealed preferences in a prior period), they can exploit the associated predictable profits and losses via the design of health insurance products. In addition, insurers can base their selection strategy on the information that is increasingly presented in the literature about which groups of people generate predictable profits or losses, given a certain equalization formula (e.g., Van de Ven and Ellis, 2000; Lamers, 1999 and 2001; Van Barneveld et al., 2000 and 2001; Shen and Ellis, 2001; Van Kleef and Van Vliet, 2022; Van Kleef et al., 2024). And, finally, insurers don't need to know which specific individuals are (un)profitable, it is sufficient for them to know that *there are* (un)profitable individuals. They can then use general selection tools, e.g., different level of deductibles, the copayment structure and benefits design. Consumers will then use their information surplus for choosing the most appropriate health insurance product (adverse selection). Of course, to

[1] *Selection* can be defined as "actions by consumers and insurers to exploit unpriced risk heterogeneity and break pooling arrangements" (Newhouse, 1996).
The insurers' incentives for risk selection are determined by both the regulation-induced *predictable* profits and losses on the insured and the insurer's costs of selection, e.g., a reduction of good reputation and the costs of the selection activities (e.g., finding out who are the (un)profitable people and designing, marketing, and administering new insurance products to deter unprofitable people and attract profitable people). In this paper we will abstract from these costs and only focus on predictable profits and losses.

[2] For reasons of fairness there may be some exceptions, e.g., if predictable profits reflect underutilization (Van de Ven and Ellis, 2000, 783-784; see also McWilliams et al., 2023).

prevent selection the risk equalization needs only to compensate for predictable profits and losses as far as insurers and/or consumers can exploit them by selection actions. But it is hard to think of any predictable profits or losses that cannot be exploited by selection actions.[3]

The coefficients of the risk equalization formula are typically estimated by means of the Ordinary Least Squares (OLS) regression method with individual healthcare expenses as dependent variable and a set of risk adjusters as explanatory variables. OLS finds the set of coefficients that minimizes the "sum of squared residuals", given the set of risk adjusters. The statistical software used to run OLS regressions typically also presents the so-called R^2.[4] The R^2-value indicates the proportion of the total variance in individual expenditures that is explained by the linear influence of the set of risk adjusters, and ranges between zero and one.[5] In empirical evaluations of risk equalization models, researchers routinely present the R^2-value, but often without an explicit interpretation. More specifically, researchers and policymakers often implicitly use the R^2 as a measure to evaluate the predictive performance of a risk equalization model, with a value closer to one indicating a better performance. By predictive performance we mean the extent to which a risk equalization model reduces *predictable* profits and losses.

The goal of this paper[6] is to critically review the use of R^2 in risk equalization research. In section 2 we discuss why the R^2 is hard to interpret and can lead to wrong and misleading conclusions regarding the predictive performance of a risk equalization model. In many cases, we do not know whether a risk equalization model with $R^2 = 0.30$ reduces the predictable profits and losses more than a model with $R^2 = 0.20$. In section 3 we argue that part of our critique on using the R^2 as a performance measure also holds for other statistical metrics such as the Mean Absolute Prediction Error (MAPE), Cumming's Prediction Measure (CPM), and the Payment System Fit (PSF). While the R^2 is inappropriate for indicating the extent to which a risk equalization model mitigates predictable profits and losses, it can be useful for other purposes, which will be discussed in section 4. Section 5 summarizes the conclusions, which will be discussed in section 6.
Finally, unfortunately, we cannot answer the question: Why do we use R^2 so often in risk equalization research? We will speculate on some hypothetical answers in section 6.4.

[3] Achieving a level playing field for insurers and achieving maximum cross-subsidies among consumers require that all risk factors that insurers are not allowed to use for premium differentiation should be included in the risk equalization (Van de Ven et al., 2023).
[4] The R^2 can be calculated as: the variance of predicted expenses divided by the variance of actual expenses, or the square of the correlation between predicted expenses and actual expenses, or one minus the ratio of the variance of the prediction error divided by the variance of actual expenses (where the prediction error equals the difference between predicted expenses by the risk equalization model and actual expenses) (Van Veen et al., 2015). These three ways of calculating the R^2 are only equivalent when an OLS-model is used and expenses are predicted on the full sample. If this is not the case, the only correct way of calculating the R^2 is the third method (i.e., the R^2 proposed by Efron, 1978).
[5] Ideally, to prevent overfitting R^2-values should be reported which are based on out-of-sample predictions. In that case Efron's (1978) R^2 should be used.
[6] This paper is partly based on several previous papers by the authors and their colleagues (see the references). The novelty of this paper is that it brings together some relevant R^2-related aspects of these previous papers, builds upon them, extends the analysis, provides several new arguments and insights, and comes with new conclusions and recommendations concerning the R^2 in the context of risk equalization.

2. R^2 is hard to interpret

Nearly all empirical studies that evaluate the predictive performance of risk equalization models present the statistical R^2-value (Van Veen et al., 2015). The R2-value is often (implicitly) interpreted as a measure of the extent to which the risk equalization payments remove the regulation-induced 'predictable profits and losses on the insured' (i.e., the incentives for selection), with a higher R^2-value indicating a better performance. For three reasons, however, the R^2-value is hard to interpret as a measure of selection incentives. First, the maximum R^2-value is typically unknown. Consequently, the R^2-value does not indicate to what extent predictable profits and losses remain after risk equalization. Second, the R^2-value is nonlinearly related to the predicted profits and losses. As a result, a change in R^2-value due to the inclusion of a new risk adjuster does not indicate the extent to which that risk adjuster contributes to the reduction of predictable profits and losses. Third, a high R^2-value does not necessarily imply a better predictive performance than a low R^2-value. Consequently, comparing alternative equalization models solely on the basis of the R^2-value can lead to misleading conclusions about which model is to be preferred. Below, we explain these three problems in more detail.

2.1. No benchmark because the maximum R^2-value is typically unknown

The R^2 indicates the proportion of the *total* variance in individual healthcare expenses that can be predicted by the risk adjusters. A major problem with the interpretation of the R^2-value is that a substantial part of the variance in individual healthcare expenses is *not predictable*, due to e.g., accidents and the unforeseen onset of new diseases, and thus cannot be anticipated on by insurers and consumers. Therefore, when assessing the predictive performance of a risk equalization model we are more interested in the proportion of the *maximum predictable* variance that is predicted by the risk adjusters than in the proportion of the *total* variance that is predicted by the risk adjusters (= R^2-value). However, in general we do not know the maximum R^2-value for a specific setting. This lack of a benchmark makes R^2 hard to interpret and an inappropriate measure of the extent to which the risk equalization payments remove the regulation-induced *predictable* profits and losses. More specifically, the R^2-value does not indicate to what extent predictable profits and losses remain after risk equalization.

Some studies have dealt with the question what the maximum R^2 is that can be achieved by a set of prospective[7] risk adjusters in a specific setting.[8] These studies are based on different assumptions about the anatomy of the variance of the actual expenses. Newhouse et al. (1989)

[7] The R^2 of *prospective* equalization models captures only predictable variance, in contrast to the R^2 of *concurrent* equalization models that also captures unpredictable variance. To our knowledge there is no study on the maximum R^2 for concurrent models.

[8] Other studies have analyzed the extent of predictable profits and losses that remain after risk equalization by simulating a 'best available' prediction model for insurers. For different models that the regulator could use, these studies present the predictable profits and losses that insurers could make with their 'best available' prediction model. Although these studies are informative, they underestimate the potential profits (losses) that insurers could make (avoid) because in practice insurers and consumer most likely have more predictive information than is included in these 'best available' models. Examples of such studies are Lamers (1999 and 2001), Van Barneveld et al. (2000 and 2001), and Shen and Ellis (2002).

and Newhouse (1996) assumed that the variance of the actual expenses can be divided into three components:[9]

1. The first component is a *fixed* effect, i.e., the constant above or below average expenses of a person over the last years. For the analyzed data set this component has been estimated to be at most 15-20 percent of total variance.
2. The second component is a predictable, but *time-varying* effect, i.e., the above or below average expenses of a person in the next period but that would not persist in later periods. This component is estimated to be another 3-5 percent of total variance.
3. The third component is the *random* component, i.e., the unpredictable variance.

Newhouse (1996) concludes that with risk factors that are based on observed individual utilization/expenses in prior periods one can explain around 20-25 precent of the annual actual expenses. However, this percentage is a lower bound of the maximum predictable variance because there may be additional predictive information that is not reflected in past utilization/expenses (e.g., giving birth in the next period). Therefore, consumers or insurers could potentially predict more than the 20-25 percent, but how much more is unknown. Van Vliet (1992) compared four different error component models and concluded that for the analyzed data set 'no more than 20 percent of annual costs variations are predictable at the individual level'.

One should be careful in using the maximum R^2 estimated in one setting as a benchmark in other settings because the maximum R^2 is context specific and may depend on the following determinants (Van de Ven and Ellis, 2000, 790-793):

- *The type of services under analysis.*
  For example, outpatient care and pharmaceutical care are typically more predictable than inpatient care, and long-term care is typically more predictable than acute care. (Van de Ven and Ellis, 2000, 790). This implies that for one set of services the maximum R^2 can be lower/higher than for another, ceteris paribus.
- *The (sub)population under analysis.*
  Kronick et al. (1995) concluded that expenditures are much more predictable among persons with chronic conditions than for other populations. They hypothesize that among persons with chronic conditions (disabilities or diseases) a much greater portion of resource utilization results from chronic problems and their complications that persist from year to year, and a smaller portion from acute episodes that lead to short-term spikes in resource use but are not followed by long-term needs. The same holds for age, with elderly people having on average more chronic diseases than young people. Newhouse et al. (1989,1993) and Van Vliet (1992) found empirical evidence that differences in health expenditures for older individuals are more predictable than those for young people. This implies that for one population the maximum R^2 can be lower/higher than for another, ceteris paribus.
- *The price of healthcare services*
  A change of the price of healthcare services may influence the R^2-value. For example, Van

[9] For a nontechnical description of this method see Newhouse et al. (1997).

Kleef et al. (2018, p. 414-5) show that a substantial reduction of the price for kidney dialysis, which was included as a risk adjuster in the risk equalization formula, ceteris paribus resulted in a substantial decrease in the R^2-value with 0.05 (from about 0.25 to 0.20). This implies that for one set of healthcare prices the maximum R^2 can be lower/higher than for another, ceteris paribus.

- *The level of effective medical technology*
  Van de Ven and Ellis (2000, p. 791) hypothesize a positive relation between R^2 and the level of medical technology. An increase of the level of effective medical technologies may keep alive at-risk patients who otherwise would have died, and thereby increases the proportion of chronically ill persons. Their healthcare expenses are better predictable than those of healthy people without chronic problems (see above). This implies that for one level of medical technology the maximum R^2 can be lower/higher than for another, ceteris paribus.
- *The length of the period being predicted*
  By using a longer period for predicting healthcare expenses (e.g., a year rather than a month) some of the random variation is averaged out, which - under the assumption that the systematic variation largely remains unchanged - increases the R^2. This implies that for one length of period the maximum R^2 can be lower/higher than for another, ceteris paribus.

Because of these many determinants, the maximum R^2-value can substantially differ across settings and over time. For example, the above-mentioned results about the maximum R^2-values presented by Newhouse (1996) and Van Vliet (1992) are based on data sets from the 1970s. These estimated maximum R^2-values seem no longer relevant, as recent studies that estimate a prospective risk equalization model based on data sets from forty years later find R^2-values somewhere in the range of 0.30-0.35. For example, Van Kleef et al. (2020) found an R^2 of 0.30, which is clearly below the maximum R^2 because they also found that the equalization model studied in their paper undercompensates insurers for chronically ill people.

So, due to a lack of a benchmark one should be careful in using the R^2 as a measure to indicate the extent to which a risk equalization model reduces the regulation-induced predictable profits and losses. Moreover, one should be careful to compare the R^2-values across different settings and over time. Ideally, to have a benchmark researchers should estimate the maximum R^2 on the same (longitudinal) data base that is used for estimating the coefficients of the risk equalization model. In practice, however, this is hardly done.[10]

2.2. R^2 is nonlinearly related to the predicted profits and losses

Even if we knew the maximum R^2, the R^2-value would still be hard to interpret. An x-percent decrease of the difference between R^2 and the maximum R^2, as a result of an improvement of the equalization formula, cannot be interpreted as an x-percent decrease of the predictable profits and losses, because the relation between R^2 and the predicted profits and losses is nonlinear. The amount of profit an insurer can make by exploiting its private information on risk is a nonlinear function of the amount of that information, and, from the point of view of

[10] For an example of such analysis, see Lamers (2001, p. 426).

the regulator, the nonlinearity can be in exactly the wrong direction because the last percentage point of variance explained may have a larger effect on profit than the earlier changes (Newhouse et al., 1989; Newhouse 1996; Van Barneveld et al., 2000). For example, Van Barneveld et al. (2000, p. 134) found that the first five percentage points increase in the R^2 lead to a reduction of the mean absolute predicted result (=profit/loss) of 360 Dutch guilders, whereas the last five percentage points increase in the R^2 lead to a reduction of 720 Dutch guilders. Newhouse et al. (1989) found similar results.

The quadratic weighing of residuals (i.e., the insurers' actual profits and losses) in the R^2 results in a nonlinear relation between R^2 and the predicted profits/losses, and can lead to misleading conclusions in the case of *small* predictable profits and losses. Although large residuals may be more problematic than small residuals, it is not obvious that quadratic weighting is better for the evaluation of the regulation-induced profits and losses than alternative forms of weighting. According to Layton et al. (2017) the argument in favor of squaring the residuals, such as in R^2, has a sound basis in welfare economics, where it is normally assumed that the welfare loss of a distortionary incentive is proportional to the square of the distortionary incentive. Although this squaring-argument may hold in the familiar context of a tax (Harberger, 1964), it is not clear that it is a valid argument for squaring the residuals as is done in R^2 in the context of risk equalization. First, the distortionary incentives are the (by consumers and insurers) *predictable* profits and losses and not the *actual* profits and losses (i.e., the residuals), which are much larger and have more extreme outliers (that weigh heavily in the squaring) than the *predictable* profits and losses. The R^2 is a function of the squared *actual* profits/losses and not of the *predictable* profits/losses. The R^2 attaches enormous weight to large outliers: the one person in a sample with an actual loss of a million euros will add as much to the variance as 1,000,000 people with an actual loss of 1,000 euro (Van de Ven and Ellis, 2000, p. 810). Second, in the case of selection incentives one could hypothesize that above a certain level of the distortionary incentive the welfare loss looks more like a flat curve than a quadratic function of the distortionary incentive. The higher the undercompensations and the more stable the undercompensations are over time, the more insurers may respond to the distortionary incentives. However, it could be hypothesized that at a certain level of undercompensation U the insurers have achieved their 'maximum selection actions', i.e., they do their utmost best to get rid of unprofitable patients and there are no additional cost-effective selection actions. This would imply that with undercompensations larger than U the welfare loss resulting from risk selection is a flat curve. So, it is not obvious that quadratic weighting is better than alternative forms of weighting.[11]
The quadratic weighing of residuals can lead to misleading conclusions in the case of *small* predictable profits and losses. For example, Van Kleef (2022) present results that the addition of socio-economic characteristics to a relatively good performing risk equalization model increases the R^2 from 0.3438 to 0.3440. The elimination of the relatively small predictable profits and losses on the different socioeconomic groups results in only a seemingly

[11] Alternatively, one can apply measures such as the Mean Absolute Prediction Error (MAPE) and Cumming's Prediction Measure (CPM), which both are a function of absolute deviations, rather than squared deviations (see section 3.2 and 3.3).

negligible increase of the R^2. If a regulator would base his decision *solely* on the R^2-value rounded to three decimal places (0.344) one might conclude not to include these risk adjusters in the risk equalization model. However, a regulator might find the addition of these risk adjusters to be relevant as they eliminate the incentives for risk selection against socio-economic groups and thereby eliminate the negative welfare effects of that type of risk selection. This example illustrates that evaluating (potential) risk adjusters solely on the basis of R^2 can lead to misleading conclusions.

2.3. *A high R^2 does not necessarily imply a better predictive performance than a low R^2*
According to Layton et al. (2018b, p. 141-142) it is intuitive that a higher R^2 should improve the performance of a payment system with respect to selection problems. This reflects the general perception that a high R^2 indicates a better predictive performance of the risk equalization model than a low R^2. However, this general perception is not necessarily correct. Below we give some counterexamples.

A first example is the comparison of the R^2-values for risk equalization models that explain different types of healthcare expenses. For example, one model explains outpatient expenses and the other model explains inpatient expenses. Although the first model may have a higher R^2 than the second model, it does not necessarily mean that it reduces the selection incentives more than the second model. The reason is that outpatient expenses are much more predictable than inpatient expenses (and therefore has a higher maximum R^2). For similar reasons a higher R^2-value in setting A than in setting B does not necessarily mean that selection incentives are smaller in A than in B (see our discussion in section 2.1).

A second example is given by Van Barneveld et al. (2000, p. 136-137). They simulated on the same data set both several risk equalization models that a regulator could use and a selection model that an insurer could use. Their empirical results showed two regulator-models such that the model with the higher R^2 resulted in higher incentives for selection (and not in lower incentives). They concluded that R^2-values can be misleading if they are used as an indicator of an insurer's incentives for selection.

A third example is the comparison of the R^2 of prospective and retrospective risk equalization models. *Prospective* equalization models include risk adjusters based on information that is known before the prediction period (e.g., diagnoses in the prior year), while *concurrent* equalization models use risk adjusters based on information that becomes known during the prediction period (e.g., diagnoses in the current year). The R^2 of concurrent models is typically higher than that of prospective models because the correlation between healthcare expenses and current-year diagnoses is inherently higher than the correlation between healthcare expenses and prior-year diagnoses. However, it would be incorrect to conclude that *because of their higher R^2* concurrent models reduce the incentives for risk selection more than prospective models. Comparison of prospective and concurrent models on the basis of R^2 alone is problematic because the R^2-value of a concurrent model does not only capture 'predictable' variance in healthcare expenses but also some 'unpredictable' variance related to occurrence of new health problems that could not have been predicted ex-ante. So, without

further information and without a relevant benchmark the difference between the R^2 for prospective and concurrent models is hard to interpret.

These examples illustrate that the classical interpretation that a high R^2 indicates a better predictive performance than a low R^2, can be misleading in the context of risk equalization.

2.4. Conclusion

Most empirical studies that estimate the coefficients of a risk equalization regression model (routinely) present the R^2-value, which is often (implicitly) interpreted as a measure of the extent to which the risk equalization payments remove the regulation-induced predictable profits and losses on the insured. However, it is hard to interpret the R^2-values and to compare the R^2-values across different settings for the following reasons:

1. In nearly all studies there is no benchmark for interpreting the R^2, because the maximum R^2-value is unknown. Moreover, there are many determinants of the maximum R^2, which are usually not taken into account.
2. The R^2 is nonlinearly related to the predicted profits and losses.
3. A high R^2 does not necessarily imply a better predictive performance than a low R^2. In many cases we do not know whether a model with $R^2 = 0.30$ reduces the predictable profits and losses more than a model with $R^2 = 0.20$.

In sum, in the context of risk equalization as a tool to reduce the predictable profits and losses the R^2-value is hard to interpret, can lead to wrong and misleading conclusions[12], and is therefore not useful for measuring selection incentives.

3. Related measures-of-fit: similar problems

Mutatis mutandis some or all the above-mentioned problems with the R^2-values hold similarly for related statistical measures-of-fit such as the Mean Absolute Prediction Error (MAPE), Cumming's Prediction Measure (CPM) and the Payment System Fit (PSF).

3.1. Mean Absolute Prediction Error (MAPE)

As discussed above, the conventional R^2 attaches enormous weight to large outliers. A measure that does not weigh prediction errors differently, is the Mean Absolute Prediction Error (MAPE). The MAPE is calculated as the mean of the absolute value of 'predicted expenses minus actual expenses' across all individuals. The MAPE is less sensitive to

[12] The Dutch government repeatedly tried to convince the Parliament that the risk equalization works well by showing that the R^2 in analyses explaining the "cost variation *among insurers' portfolios*" is 0.98 (Tweede Kamer 2011a, Tweede Kamer 2011b). However, this is an incorrect argument. A problem with the R^2 and other measures based on the variation in (residual) spending at the *insurer level* is that the outcomes of these measures heavily depend on the distribution of risk types across insurers' portfolios and on the cost structure in these portfolios (Van Kleef et al., 2022). For example, consider the worst risk equalization formula with for each person the predicted expenses equal to the mean per person expenses. Such a risk equalization model has maximum predictable profits and losses. Nevertheless, if the risk composition and cost structure are identical across the insurers' portfolios, the R^2 at the *insurer level* is 1.0. So, the R^2 at the *insurer level* is misleading because it is not a valid measure of predictable profits and losses on the insured.

extreme values in the distribution of expenses than the R^2. However, it is unknown what the minimum value of the MAPE is that can be achieved with risk equalization models. Mutatis mutandis the problems with the interpretation of R^2-values as mentioned in the sections 2.1 and 2.3 hold similarly for the MAPE-values.

3.2. Cumming's Prediction Measure (CPM)

Another measure that does not weigh prediction errors differently, is the Cumming's Prediction Measure (CPM), which is equal to one minus the ratio of the MAPE to the mean absolute deviation from the mean expenses (Cumming et al., 2002, p.51). Like the R^2, its value ranges between zero and one. One could interpret the CPM as the proportion of the mean absolute deviation from the mean expenses that is explained or predicted by the linear influence of the set of risk adjusters. However, the maximum value of CPM that can be achieved with risk equalization models is unknown, because the minimum value of the MAPE is unknown.

Mutatis mutandis the problems with the interpretation of R^2-values as mentioned in the sections 2.1 and 2.3 hold similarly for the CPM-values.

3.3. Payment System Fit (PSF)

Zhu et al. (2013) introduced the measure Payment System Fit (PSF) as a generalized form of R^2 to capture the fit of the *entire* payment system, i.e., equalization payments plus other payments, e.g., from risk sharing (i.e., ex-post cost-based payments to the insurers) and premiums. Their PSF is an R^2-type statistical measure-of-fit that quantifies the portion of the variance in healthcare expenses that is explained by the *entire* payment system.[13] This PSF is analogous to Efron's R^2 (1978), which is equal to one minus the ratio of the variance of the error term to the variance of the dependent variable. Similar arguments why the R^2 is hard to interpret (see above) also apply to the PSF. Although Zhu et al. (2013) did not give an explicit interpretation of the calculated PSF-values when presenting their empirical results, they wrongly gave the impression that maximizing PSF (just like R^2) is an objective of the payment policy.[14]

The PSF has also been applied in other studies, e.g., Geruso and McGuire (2016), Beck et al. (2020), McGuire et al (2021a), Schmid and Beck (2016), Van Kleef and Van Vliet (2022), and Henriquez et al. (2023). In contrast to Zhu et al. (2013) these other studies explicitly use the PSF to quantify *selection incentives*. More specifically, these studies associate a higher PSF-value with reduced incentives for risk selection. However, in the case of risk sharing (as

[13] To our knowledge Zhu et al. (2013) were the first to introduce this Payment System Fit (PSF).

[14] Zhu et al. (2013, p. 216) motivated the introduction of the Payment System Fit (PSF) as follows: "One objective of payment policy is to match payments to expected costs for individuals. Risk adjustment systems are commonly graded by their R-squared, a statistic reporting how much of the variation in health care costs is explained by the variables in the regression underlying the risk adjustment formula. Our generalization of the statistical R-Squared metric reflects how much of the total variation in plan-paid costs is captured by all payment-system features." They rightly stated that an objective of payment policy is to match payments to *expected* costs for individuals. However, as argued above, the R^2 measures the fit between payments and *actual* costs, and the same holds for the PSF. Therefore, Zhu et al. (2013) wrongly gave the impression that maximizing PSF (just like R^2) is an objective of the payment policy.

in these studies) this doesn't have to be true (for a counterexample see Table 1). The PSF-value indicates the fit between the payments that an insurer receives and the *actual* expenses, while for an indicator of the incentives for risk selection the fit between these payments and the *predictable* expenses is relevant.

In general, it is unknown which portion of actual spending is predictable. Consequently, it is unknown whether and to what extent risk sharing (i.e., payments based on actual spending) reduces the incentives for risk selection. E.g., in the hypothetical situation that risk equalization fully compensates for predictable profits and losses, the addition of risk sharing payments can only lead to compensation of unpredictable losses and therefore does not reduce the incentives for risk selection. Nevertheless, the PSF-value (strongly) increases.[15] The same conclusion holds in the case of a very high threshold in the risk sharing scheme such that the risk sharing payments only compensate for unpredictable losses.

Van Kleef et al. (2022) present an example that clearly illustrates that in the case of risk sharing a higher PSF-value does not necessarily reduce the predictable profits and losses. A 'payment system with 50% proportional risk sharing and *without* risk equalization' has a PSF-value of 0.75 and reduces the predictable profits/losses for subgroups of interest by 50% (assuming that premiums are community-rated). Although systems with state-of-the-art prospective risk equalization have a PSF-value that is much lower (somewhere in the range of 0.30-0.35), these risk equalization models have been proven to better mitigate predictable profits and losses than 50% proportional risk sharing. For example, Van Kleef et al. (2020) find that the prospective risk equalization model used in the Netherlands in 2016 has a PSF of 0.298 and reduces predictable profits/losses for subgroups of interest by 84% (see Table 1). So, when comparing different 'entire payment systems' the model with the highest PSF-value does not necessarily reduce the predictable profits and losses the most, even if they are estimated on the same data.

Table 1. A high PSF may not be preferred to a low PSF

	PSF	Reduction of the predictable profits/losses for subgroups of interest
Model 1 *	0.750	50%
Model 2 **	0.298	84%

* Model 1: A payment system with 50% proportional risk sharing and *without* risk equalization.
** Model 2: A payment system with risk equalization and *without* risk sharing. This payment system mimics the risk equalization model used in the Netherlands in 2016. Source: Van Kleef et al. (2020)

In sum, in the context of 'risk equalization and risk sharing as tools to reduce the predictable profits and losses' the PSF is hard to interpret as a measure of selection incentives, is an

[15] In fact, in the case of perfect risk equalization risk sharing might *increase* the incentives for selection depending on the financing of the risk sharing payments. For different forms of risk sharing and different forms of financing the risk sharing payments see e.g., Van de Ven and Ellis (2000).

inappropriate measure of the extent to which the payment system reduces predictable profits and losses, and is therefore not useful for measuring selection incentives.

4. Applications in which R^2 can be useful

Despite the previously discussed problems, in the following cases R^2 can be useful.[16]

4.1. F-test when estimating different equalization formulas on the same data

In section 2.1 we discussed several determinants of the (maximum) R^2 that make it hard to compare the R^2-values across different settings and over time. When comparing different risk equalization formulas that are defined by adding new risk adjusters to the previous formula and that are estimated by means of OLS on the *same data*, this problem does not occur. However, as discussed above, it is then still hard to interpret an increase of the R^2-value in terms of a reduction of the selection incentives, because there is no benchmark (if, as usual, the maximum R^2 is unknown) and because of the non-linear relation between R^2 and the predicted profits/losses.[17] However, the R^2-values can be useful for calculating the F-value to perform an F-test to test whether the set of additional risk adjusters, given the set of original risk adjusters, jointly make a statistically significant contribution to further explain the variance in healthcare expenses (see e.g., Kmenta, 1971, p. 371).

4.2. Evaluating the validity of data used for calibration of risk equalization models

The R^2-value can be informative at the calibration stage of risk equalization models. For example, when recalibrating a given risk equalization formula on a more recent data year, the R^2 might be a helpful indicator to detect changes in cost patterns in the estimation data. A substantial decrease or increase in R^2-value for the same risk equalization model when going from one data year to another might point at a change in cost structures. Researchers might want to find the source of such a change when evaluating the validity of the estimation data. For example, Cattel et al. (2022) find a substantial decrease in R^2-value of the Dutch risk equalization model when moving from data-year 2019 to data-year 2020. After further research, Cattel et al. (2022) conclude that this decrease was caused by the COVID-19 pandemic. More specifically, the pandemic lead to a disproportionate increase in unpredictable spending variation. After some careful checks, however, the impact on the coefficients of the risk equalization model seemed to be limited and the researchers advised to use the 2020-data for calibration of the risk equalization model for 2023.

4.3. Evaluating the extent to which 'constraints' on the estimated coefficients are binding

Risk equalization models can include constraints on the estimated coefficients. For example, the Dutch risk equalization model-2024 for somatic care includes the restriction that for a specific risk class of healthy and unhealthy people (which is not included as a risk adjuster

[16] Mutatis mutandis the same holds for MAPE and CPM; not for PSF.
[17] As discussed in section 2.2 this non-linear relation can lead to misleading conclusions in the case of *small* predictable profits and losses. For another interpretation problem see the example given by Van Barneveld et al. (2000) discussed in section 2.3.

variable) mean predicted spending equals mean actual spending.[18] From a statistical point of view, the R^2-value can be an informative measure to indicate the extent to which a restriction is binding, i.e., the extent to which the constraint requires deviations from the original OLS-estimates. If the constraint is not binding (because it is already fulfilled in the original OLS without a constraint), the R^2 remains the same. The more a constraint is not fulfilled with the original-OLS-estimated coefficients, the more will the constrained-OLS-estimated coefficients of the risk adjusters that are correlated with the constraint-factor, (have to) deviate from the original-OLS-estimated coefficients, and therefore the lower will be the R^2 of the constrained regression. If the unconstrained OLS and the constrained OLS are estimated on the same data, as we usually do, the R^2 is a monotonous (although non-linear) measure that allows us to conclude that the constraint is more binding the larger the reduction of R^2 is (see e.g., Van Kleef et al., 2017, Table 3).

4.4. Economic rather than statistical measures-of-fit

The statistical R^2 measure does not give any information about (1) whether in practice selection will occur, (2) and if so, which forms of selection, and (3) what negative effects selection will have. Therefore, in addition to the *statistical* R^2 measure Layton et al. (2017) developed an *economic* R^2-type measure to indicate the expected welfare effects of risk selection. In doing so, they made assumptions about how financial incentives influence the behavior of insurers and consumers, how this can result in price and benefits distortions, and how these distortions result in welfare loss. Although this economic R^2-type measure is useful, in our opinion it should always be carefully described and interpreted in the light of the underlying economic model and assumptions (to avoid misinterpretations by readers who are not yet familiar with that measure).

5. Conclusion

Nearly all empirical studies that estimate the coefficients of a risk equalization formula present the value of the statistical measure R^2, mostly without a clear interpretation. Often the R^2-value is implicitly interpreted as a measure of the extent to which the risk equalization payments remove the regulation-induced *predictable* profits and losses on the insured, i.e., a measure of selection incentives.

In this paper we have argued that in the context of risk equalization R^2 is hard to interpret, can lead to wrong and misleading conclusions when used as a measure of selection incentives, and is therefore not useful for measuring selection incentives. In many cases, we do not know whether a model with $R^2 = 0.30$ reduces the predictable profits and losses more than a model with $R^2 = 0.20$. So, although in the context of risk equalization research the R^2 is the most used measure-of-fit (Van Veen et al., 2015) and is often implicitly interpreted as a measure of selection incentives, it is not useful for measuring selection incentives. Similar problems hold for related statistical measures-of-fit such as the MAPE, CPM, and the PSF. There are some exceptions where the R^2 can be useful.

[18] This type of restriction has been tested in several papers, including Van Kleef et al. (2017), Withagen-Koster (2020), McGuire et al. (2021b).

Our conclusion is that one should be careful with presenting and interpreting R^2-values in risk equalization research.

6. Discussion

6.1. How to evaluate the predictive performance of risk equalization models?

Given our conclusion that the R^2 is not useful for measuring selection incentives, one could raise the question: Which evaluation measure can be used to appropriately evaluate the predictive performance of risk equalization models?

Van Veen et al. (2015) and Layton et al. (2018b) give an overview of statistical[19] measures for evaluating the predictive performance of risk equalization models. Van Veen et al. (2015) give an overview of statistical measures-of-fit that have been used for evaluating risk equalization models since 2000 and discussed the properties of 71 measures-of-fit. They concluded that the R^2 is the most commonly used measure. This underlines the relevance of our paper. Van Veen et al. (2015) conclude that "if the objective is measuring financial incentives for risk selection, the only adequate evaluation method is to assess the performance for selected nonrandom groups of interest" (different from the groups formed by the risk adjusters). If the groups are sufficiently large, the difference between the average equalization-based predicted expenses and the average actual expenses for relevant non-random groups (e.g., the chronically ill or healthy people) can be interpreted as the over/undercompensations (i.e., predictable profit or loss) for an individual in that group. Therefore, a good evaluation measure of the predictive performance of risk equalization models is the over/undercompensations for relevant non-random groups.[20] An option is to present a weighted average of the absolute over/undercompensations of a set of relevant

[19] In addition to the *statistical* R^2 measure Layton et al. (2017) developed an *economic* R^2-type measure to indicate the expected welfare effects of risk selection.

[20] Alternatively, predictive ratios for relevant non-random groups are presented in the literature (Layton et al., 2018b). A predictive ratio equals the *ratio* of the average equalization-based predicted expenses to the average actual expenses for relevant non-random groups. For two reasons we prefer the over/undercompensations over the predictive ratio to assess the predictive performance of a risk equalization model for selected non-random groups. First, a predictive ratio is hard to interpret as a single measure of selection incentives because the selection incentive (i.e., the over/undercompensation per group) depends on both the predictive ratio and the average actual expenses of the group. So, one should also know the latter to calculate a measure of selection incentives. Second, one could hold the view that when comparing the predictive ratios across groups a higher predictive ratio is associated with a larger selection incentive. However, this needs not to be true because when comparing the predictive ratios across groups there is not necessarily a monotonic relation between the predictive ratio and the size of selection incentives. See the following example:

Group	Predictive ratio	Average actual expenses (in euro)	Selection incentive = average predictable loss = (predictive ratio - 1) * average actual expenses
A	1.1	4,000	400
B	1.2	4,000	800
C	1.3	1,000	300

nonrandom groups as a single measure of the predictive performance of a risk equalization model (see e.g., Van Kleef et al., 2020).

6.2. Why maximizing R^2?

Given our conclusions about R^2, one could raise the question: Does it make sense, as we usually do, to maximize the R^2-value (i.e., using the OLS regression method) when estimating the coefficients of a risk equalization formula? Our answer is: yes. It is important to make a distinction between the role of R^2 in the *estimation method* and the R^2 as an *evaluation measure* of the predictive performance of a risk equalization model. Our criticism of the R^2 relates only to the R^2 as an evaluation measure of the selection incentives.

If we specify a *linear* relation between individual health expenses and the set of risk adjusters, it turns out that under the classical assumptions[21] both the Best Linear Unbiased (BLU) estimators and the Maximum Likelihood (ML) estimators are equivalent to the OLS estimators of the regression parameters (see e.g., Kmenta, 1971, p. 205-216). This implies that under the classical assumptions the OLS estimators are unbiased, have the smallest variance among all linear unbiased estimators, are consistent and asymptotically efficient. Therefore, it is important to emphasize that *maximizing the R^2 is not an end in itself, it is a means to get estimators with very desirable statistical properties (BLU and ML).*[22]

Another question could be: *Why* do we specify a *linear* relation between individual health expenses and the set of risk adjusters? There are two good arguments to do so. First, by specifying a linear relation we stay close to the cell-based approach. Assume that there is only one risk adjuster, e.g., age with, say, 24 classes. Then most regulators would use the cell-based approach, i.e., use the average expenses per age group as the basis for calculating the equalization payments (as in e.g., Colombia (Bauhoff et al., 2018), Israel (Brammli-Greenberg et al, 2018), and in Switzerland till 2020 (Schmid et al., 2018)). If in addition to age a second risk adjuster is relevant, e.g., yes/no healthy, most regulators would use the cell-based approach with 24x2=48 cells. The results of these cell-based approaches are exactly the same as the results of OLS with these 24 respectively 48 binary (0/1) risk adjusters as dependent variables and no intercept. When there are many additional risk adjusters the number of cells may soon become too large. A problem is then that in many cells there are too few observations to get reliable results. In addition, many coefficients may not be stable over time. A solution is to make the assumption (which is implicitly made in nearly all risk equalization models in practice) that most or all interaction terms between the risk adjusters are zero. For example, the assumption can be made that the effect of health on expenses is the same for each age group. The weights of the equalization formula can then be estimated by applying OLS to a linear relation between individual expenses and a manageable number of binary (0/1) risk adjusters.

[21] For each individual the error is normally distributed, with expectation zero and variance sigma (homoskedasticity), and is not correlated with the error terms of other individuals (see e.g., Kmenta, 1971).
[22] Even if the classical assumptions (see previous footnote) are not fulfilled, the OLS-estimators have desirable properties. If the error is not normally distributed, the OLS-estimators are still the BLU estimator, i.e., they are unbiased and have the smallest variance among all linear unbiased estimators. In the case of heteroscedasticity, the OLS estimators are still unbiased and asymptotically efficient (see e.g., Kmenta, 1971, Chapter 8).

A second argument to specify a *linear* relation between individual health expenses and the set of risk adjusters is that a linear relation avoids the complicated retransformation problems with making predictions of the individual health expenses when using a *nonlinear* relation (Duan et al, 1983). Examples of nonlinear models include two-part models[23] or nonlinear transformations of the health expenses such as the logarithm of health expenses (see e.g., Manning et al., 1987). However, with the two-part models the predicted expenses are seriously biased in the case of heteroscedasticity, and the use of the simple transformation log(expenses+1) has very poor statistical properties when calculating the predicted individual expenses by retransforming to the original scale (e.g., euros rather than log-euros). (Mullahy, 1998; Manning, 1998). According to Mullahy (1998) applying OLS to individual health expenses may be sufficient when sample sizes are large.

6.3. A high R^2 is not necessarily preferred to a low R^2

In section 2.3 we concluded that a high R^2 does not necessarily imply a better predictive performance than a low R^2. However, there are also other reasons why a high R^2 is not necessarily preferred to a low R^2.

A first example is constrained regression. The goal of risk equalization is to eliminate the regulation-induced incentives for risk selection by compensating insurers for predictable profits and losses (Van de Ven et al., 2023). However, as long as the risk equalization is imperfect, the regulator may analyze the potential effects of different selection incentives (that result from the over/undercompensations for relevant non-random groups) on the insurers' and consumers' behavior, analyze what types of selection actions are possible and realistic, and what the negative effects of these selection actions are. For example, regulators might be particularly concerned about risk selection via the distortion of insurance products and via quality skimping (Van de Ven et al., 2015). The regulator could then give priority to reduce the predictable profits and losses that induce the selection actions with the most negative effects. One tool to do so is constrained regression, which, by definition, results in a lower R^2 than OLS. For example, Van Kleef et al. (2017) present results showing that adding specific constraints to an OLS-regression model reduces the R^2, but strongly reduces the predictable losses for subgroups of chronically ill people and thereby reduces the incentives for quality skimping.[24] The regulator may prefer the model with a worse *statistical* performance (in terms of R^2) but better *economic* performance (i.e., fewer selection incentives regarding groups of interest) to the model with better *statistical* performance but worse *economic* performance. That is, the regulator may prefer the model-with-low-R^2 to the model-with-high-R^2.

[23] In the first part, a yes–no expenses equation is estimated (e.g., a probit, logit, or linear probability model). In the second part, for those with positive expenses, the expenses are regressed (a linear, log-linear or square root model) on the risk adjusters (see e.g., Manning et al. 1987; Manning, 1998; and Mullahy, 1998). Conventionally, both parts of the two-part model are estimated independently and a smearing transformation (see Duan et al, 1983) is used to generate unbiased estimates of the second part of the model in the common situation in which nonlinear transformations of the dependent variable are used.

[24] This example shows that even when comparing different risk equalization formulas that are estimated on the same data, it is hard to interpret the R^2 without further relevant information, because a low R^2 may be preferred to a high R^2.

A second example is optimal risk adjustment. In the literature two different approaches for estimating the weights of a risk equalization model can be discerned given a set of risk adjusters and risk classes (that result in imperfect risk equalization): the so-called 'conventional risk adjustment' and 'optimal risk adjustment'. Layton et al. (2018a) describe *conventional* risk adjustment as a *two-step* "estimate-then-evaluate" approach: first the weights for a given risk equalization model are estimated and then second the risk equalization model is evaluated based on certain criteria that reflect the regulator's objective. *Optimal* risk adjustment is described as a *one-step* "estimate-to-maximize-the-objective" approach, where the payment weights are estimated such that the regulator's objective function is maximized.[25] If the R^2 for optimal risk adjustment is lower than the R^2 for conventional risk adjustment with the weights estimated by OLS[26], the regulator will prefer the low R^2 to the high R^2.

6.4. Why do we use R^2 so often?

Because most of the problems with the R^2 (as mentioned in this paper) are already known for decades, the question comes to mind: *Why do we use R^2 so often in risk equalization research?* Unfortunately, for most studies[27] we do not know the answer to this question. Possible answers could be that the R^2 is presented in risk equalization research (1) because most researchers do it, or (2) because R^2 is a simple comprehensive measure characterizing the entire payment system, or (3) because in other research areas than risk equalization most researchers present R^2 together with their OLS-results, or (4) because the statistical software used to run OLS regressions routinely presents the R^2, or (5) because we have been so successful in 'selling the R^2-value' to policy makers and politicians (in the early days when, due to data limitations, adequate evaluation criteria were absent) that they ask for it, or (6) because there is no other performance measure available. In the latter case it is recommended to invest in the collection of appropriate data to calculate useful statistical measures such as

[25] If the regulator's objective is perfectly specified, maximizing this objective function under realistic assumptions would be sufficient and would make a further evaluation superfluous. However, several potential objectives of the regulator can be discerned, which in 'optimal risk adjustment' (so far) are maximized without considering the other objectives. Glazer and McGuire (2002) and Layton et al. (2018a) develop a framework where the objective is to minimize the efficiency loss from service-level distortions due to adverse selection (i.e., a form of *direct* selection). This objective is relevant for countries such as the United States of America and the Netherlands, where *individual* insurers purchase or deliver the care to their own enrollees, but this objective seems less relevant for countries such as Germany and Switzerland, where the insurers *collectively* purchase the care. Lorenz (2017) analyzes optimal risk adjustment where the objective is to eliminate *indirect* selection (e.g., selective advertising, or avoiding people living in high-cost areas). In practice, however, regulators might also be concerned with other objectives, such as a level playing field for insurers, the desired level of cross-subsidies, no adverse selection into/out the market by consumers, avoiding other negative effects of risk selection, and basing equalization payments only on acceptable costs (see e.g., Van de Ven and Ells, 2000). According to Layton et al. (2018a) a formal incorporation of *all* regulator's objectives within a single social welfare function is probably unrealistic. Therefore, also in the case of optimal risk adjustment, after the *first* step of estimating the payment weights, a *second* step is desirable to evaluate the predictive performance of the risk equalization model.

[26] As is the case in the empirical analyses by Glazer and McGuire (2002), Lorenz (2017) and Layton et al. (2018a).

[27] In section 4 we discussed some applications in which R^2 can be useful. However, most studies on risk equalization that present the R^2 fall outside the scope of these applications.

the average over/undercompensations per relevant selected non-random group (e.g., chronically ill or healthy people)" (Van Veen et al., 2015).
In general, our recommendation is to either present the R^2 with a clear, valid interpretation or not to present the R^2. The same holds for the related statistical measures MAPE, CPM, and PSF.

Acknowledgements We would like to thank Tom McGuire, Normann Lorenz and the participants at the Risk Adjustment Network Conference 2023 (1-4 November 2023, The Hague) for their valuable comments. The views presented in this paper do not necessarily reflect those of the mentioned persons. Only the authors are responsible for the content of this paper.

Funding There was no external funding for this project.

Conflict of interest The authors have no relevant financial or non-financial interest to disclose.

REFERENCES

Bauhoff, S.,I. Rodriguez-Bernate, D. Gopffarth, R.Guerrero, I. Galindo-Henriquez and F. Nates. (2018) Health plan payment in Colombia. Chapter 10 in T.G. McGuire and R.C. van Kleef *Risk adjustment, risk sharing, and premium regulation in health insurance markets, Theory and Practice*. Elsevier, Academic Press, 279-294.

Beck, K., L. Kauer, T.G. McGuire, C.P.R. Schmid. (2020). Improving risk-equalization in Switzerland: Effects of alternative reform proposals on reallocating public subsidies for hospitals. *Health Policy* 124: 1363-1367.

Brammli-Greenberg, S., J. Glazer, A. Shmueli. (2018) Regulated competition and health plan payment under the National Health Insurance Law in Israel – The unfinished story. Chapter 13 in T.G. McGuire and R.C. van Kleef *Risk adjustment, risk sharing, and premium regulation in health insurance markets, Theory and Practice*. Elsevier, Academic Press, 365-395.

Cattel, D., F. Eijkenaar, M. Oskam, A. Panturu, R.C. van Kleef & R.C.J.A. van Vliet. (2022). *Onderzoek Risicoverevening 2023: Overall toets.* (Research for the risk equalization model 2023: calibrating the model on the new data year.) Erasmus University Rotterdam. Research report.

Cumming, R.B., D. Knutson, B.A. Cameron, and B. Derrick. (2002) *A Comparative Analysis of Claims Based Risk Assessment for Commercial Populations*, Society of Actuaries, USA. https://www.researchgate.net/publication/228378875_A_Comparative_Analysis_of_Claims_Based_Risk_Assessment_for_Commercial_Populations

Duan, N. et al., 1983, Smearing estimate: a nonparametric retransformation method. *Journal of the American Statistical Association*. 78, 605-690.

Efron, B. (1978). Regression and ANOVA with zero-one data: Measures of residual variation. *Journal of the American Statistical Association*, 73, 113-121.

Geruso, M., & T.G. McGuire. (2016). Tradeoffs in the design of health plan payment systems: Fit, power and balance. *Journal of Health Economics* 47: 1-19.

Glazer, J., and T.G. McGuire (2000) Optimal Risk Adjustment in Markets with Adverse Selection: An Application to Managed Care. *American Economic Review* 90(4), 1055-1071.

Glazer, J., and T.G. McGuire (2002) Setting health plan premiums to ensure efficient quality in health care: minimum variance optimal risk adjustment, *Journal of Public Economics* 84: 153–173.

Harberger, A.C., (1964), Principles of efficiency; the measurement of waste *American Economic Review* 54(3): 58-76

Henriquez, J., R.C. van Kleef, A. Matthews, T. G. McGuire, F Paolucci (2023) *Combining risk adjustment with risk sharing in health plan payment systems: private health insurance in Australia.* NBER Working paper.
Kmenta, J., *Elements of Econometrics*, Macmillan Publisher Co, New York, 1971

Kronick, R., Z. Zhou and T. Dreyfus, (1995) Making risk adjustment work for everyone. *Inquiry* 32, 41-55.

Lamers, L.M., (1999) Risk-Adjusted Capitation Based on the Diagnostic Cost Group Model:
An Empirical Evaluation with Health Survey Information. *Health Services Research* 33(6): 1717-1728

Lamers, L.M., (2001) Health-based Risk adjustment: is inpatient and outpatient diagnostic information sufficient? *Inquiry* 38(4): 423-431.

Layton, T.J., R.P. Ellis, T.G. McGuire, R.C. van Kleef (2017) Measuring efficiency of health plan payment systems in managed competition health insurance markets.
Journal of health Economics 56, 237-255.

Layton, T.J., T.G. McGuire, R.C. van Kleef (2018a) Deriving risk adjustment payment weights to maximize efficiency of health insurance markets *Journal of Health Economics* 61: 93–110.

Layton, T.J., R.P. Ellis, T.G. McGuire, R.C. van Kleef. (2018b) Evaluating the Performance of Health Plan Payment Systems. Chapter 5 in T.G. McGuire and R.C. van Kleef *Risk adjustment, risk sharing, and premium regulation in health insurance markets, Theory and Practice*. Elsevier, Academic Press, 133-167.

Lorenz, N., (2017) Using quantile and asymmetric least squares regression for optimal risk adjustment *Health Economics* 26: 724–742.

Manning, W.G., 1998, The logged dependent variable, heteroskedasticity, and the retransformation problem. *Journal of Health Economics* 17(3) 283-296.

Manning, W.G., et al., 1987, Health Insurance and the demand for medical care: evidence from a randomized experiment. *American Economic Review.* 77, 251-277.

McGuire, T.G., S. Schillo, R.C. van Kleef (2021a) Very high and low residual spenders in private health insurance markets: Germany, The Netherlands and the U.S. Marketplaces. *The European Journal of Health Economics* 22:35-50.

McGuire, T.G., A.L. Zink & S. Rose. (2021b). Improving the Performance of Risk Adjustment Systems: Constrained Regressions, Reinsurance, and Variable Selection. *American Journal of Health Economics* 7: 497-521.

McWilliams, J.M., G. Weinreb, L. Ding, C.D. Ndumele, and J. Wallace. (2023) Risk adjustment and promoting health equity in population-based payment: concepts and evidence. *Health Affairs*, 42(1) 105-114.

Mullahy, J. 1998, "Much ado about two: reconsidering retransformation and the two-part model in health econometrics." *Journal of Health Economics*, 17(3), p247-282.

Newhouse, J.P., 1996, Reimbursing health plans and health providers: efficiency in production versus selection. *Journal of Economic Literature* 34, 1236-1263.

Newhouse, J.P., W.G. Manning, E.B. Keeler and E.M. Sloss. (1989) Adjusting capitation rates using objective health measurers and prior utilization. *Health Care Financing Review*, 10-3, pp. 41-54.

Newhouse, J.P., E.M. Sloss, W.G. Manning and E.B. Keeler. (1993) Risk adjustment for a children's capitation rate. *Health Care Financing Review*, Fall 1993, volume 15, No.1, 39-54.

Newhouse, J.P., M.B. Buntin, and J.D. Chapman (1997) Risk adjustment and Medicare: taking a closer look. *Health Affairs* 16(5): 26-43.

Schmid, C.P.R., and K. Beck. (2016) Re-insurance in the Swiss health insurance market: Fit, power and balance. *Health Policy* 120:848-855.

Schmid, C.P.R., K. Beck and L. Kauer. (2018) Health plan payment in Switzerland. Chapter 16 in T.G. McGuire and R.C. van Kleef Risk adjustment, risk sharing, and premium regulation in health insurance markets, Theory and Practice. Elsevier, Academic Press, 453-489.

Shen, Y., and R.P. Ellis. (2002) How profitable is risk selection? A comparison of four risk adjustment models. *Health Economics* 11: 165–174.

Tweede Kamer (2011a) Herziening zorgstelsel. Brief van de minister aan de Tweede Kamer, 29689(350).

Tweede Kamer (2011b) Herziening zorgstelsel. Verslag, 29689(358).

Van Barneveld, E.M., LM Lamers, RCJA van Vliet and WPMM van de Ven. (2000) Ignoring small predictable profits and losses: a new approach for measuring incentives for cream skimming, *Health Care Management Science* 3 131-140.

Van Barneveld, EM, LM Lamers, R.C.J.A. van Vliet, W.P.M.M. van de Ven. (2001) Risk sharing as a supplement to imperfect capitation: a tradeoff between selection and efficiency. *Journal of Health Economics* 20(2): 147-168.

Van de Ven, W.P.M.M. & R.P. Ellis. (2000). Risk adjustment in competitive health insurance markets. In: Culyer AJ, Newhouse JP, editors. *Handbook of Health Economics* (Chapter 14) Amsterdam: Elsevier, 755-845.

Van de Ven, W.P.M.M., G. Hamstra, R. van Kleef, M. Reuser and P. Stam. (2023). The goal of risk equalization in regulated competitive health insurance markets. *The European Journal of Health Economics* 24:111-123.

Van de Ven, W.P.M.M., R.C. van Kleef, and R.C.J.A. van Vliet. (2015). Risk Selection Threatens Quality Of Care For Certain Patients; Lessons From Europe? Health Insurance Exchanges. *Health Affairs* 34: 1713-1720.

Van Kleef, R.C. (2022). *Socioeconomic Variables in Risk Adjustment Experiences from the Netherlands.* Presentation at the RAN-Meeting, Berlin, September 2022.

Van Kleef, R.C. & R.C.J.A. van Vliet (2022). How to deal with persistently low/high spenders in health plan payment systems? *Health Economics* 31: 784-805.

Van Kleef, R.C., T.G. McGuire, R.C.J.A. van Vliet and W.P.M.M. van de Ven (2017) Improving risk equalization with constrained regression. *The European Journal of Health Economics*, 18, 1137-1156.

Van Kleef, R.C., F. Eijkenaar, R.C.J.A. van Vliet and W.P.M.M. van de Ven (2018). Health plan payment in the Netherlands. In: T.G. McGuire and R.C. van Kleef (eds.) *Risk Adjustment, Risk Sharing and Premium Regulation in Health Insurance Markets, Theory and Practice (*Chapter 14) Academic Press, Elsevier, London, 397-429.

Van Kleef, R.C., F. Eijkenaar & R.C.J.A. van Vliet (2020). Selection Incentives for Health Insurers in the Presence of Sophisticated Risk Adjustment. *Medical Care Research and Review*, 77: 584-595.

Van Kleef, R.C., M. Reuser, P.J.A. Stam and W.P.M.M. van de Ven. (2022) Positive and negative effects of risk equalization and risk sharing in regulated competitive health insurance markets, draft 30 November 2022, Erasmus University Rotterdam.

Van Kleef, R.C., R.C.J.A. van Vliet, M. Oskam. (2024) Risk Adjustment in Health Insurance Markets: Do Not Overlook the "Real" Healthy. *Medical Care* forthcoming.

Van Veen, S.C.H.M., R.C. van Kleef, W.P.M.M. van de Ven and R.C.J.A. van Vliet. (2015) Is There One Measure-of-fit that Fits All? A Taxonomy and Review of Measures-of-fit for Risk-Equalization Models. *Medical Care Research and Review*, 72: 220–243

Van Vliet, R.C.J.A. (1992) Predictability of individual health care expenditures. *The Journal of Risk and Insurance* 59 (3), 443-460.

Withagen-Koster, A.A., R.C. van Kleef & F. Eijkenaar. (2020). Incorporating self-reported health measures in risk equalization through constrained regression. *The European Journal of Health Economics*, 21: 513–528.

Zhu, J.M., T. Layton, A. D. Sinaiko, and T. G. McGuire (2013) The Power of Reinsurance in Health Insurance Exchanges to Improve the Fit of the Payment System and Reduce incentives for Adverse Selection. *Inquiry*, 50(4) 255–274.

Erasmus University Rotterdam
Erasmus Centre for Health Economics Rotterdam
Burgemeester Oudlaan 50
3062 PA Rotterdam, The Netherlands
T +31 10 408 8555
E escher@eur.nl
W www.eur.nl/escher