Injury & Disability Schemes Seminar
Insights and Outcomes
10-12 November 2019 • QT Canberra

Performance enhancing: A look at modern operational performance rating models

Prepared by Hugh Miller

Presented to the Actuaries Institute
Injury and Disability Schemes Seminar
10-12 November 2019

Institute of Actuaries of Australia
ABN 69 000 423 656
Level 2, 50 Carrington Street, Sydney NSW Australia 2000
t +61 (0) 2 9239 6100 **f** +61 (0) 2 9239 6170
e actuaries@actuaries.asn.au **w** www.actuaries.asn.au

Performance enhancing: A look at modern operational performance rating models

Hugh Miller[1]

Abstract: Performance models are a feature of many programs, including injury and disability schemes. They are particularly relevant when a program function is outsourced, and the performance of different providers needs to be compared. Unless caseloads are completely randomised, there are many design questions that have statistical implications and can generate poor provider rankings. Key findings include the difference between 'residual' approaches to performance estimation versus estimating parameters within the model, the impact of correlation between predictors and providers, and additional diagnostics that should be considered when undertaking a provider comparison exercise.

1 Introduction

Performance measurement is a core concept in business and is often the main mechanism for driving improvements in efficiency and outcomes. We are interested in situations where the 'market' for a service is divided between several agents or 'providers', with each provider taking on a set of separate cases. For example, the claims management of a workers' compensation scheme might be divided between a set of organisations for claims management, or services to a disability scheme might be provided by a set of approved organisations. Identifying which providers are doing a good job on their caseload is important:

- Often performance is tied to rewards for an organisation, such as incentive fees or greater future market share
- Finding high performing providers helps to build evidence for what approaches work
- People (cases) may have choice in which provider they choose and knowing which are the better-performing providers may help inform that decision.

Performance measurement will often be presented as league tables or rankings to enable comparison.

Many aspects of performance frameworks are, by necessity, tailored to the underlying program. However, there are many considerations that are common. This paper explores these commonalities.

We illustrate many of the issues with simulated and real-data examples. Rather than interspersing this with the commentary we have collected numerical results in section 5. We hope this approach makes the commentary more accessible to non-technical readers.

There is significant existing literature on performance measurement and ranking, although relatively less on the use of risk adjustment to improve comparisons. Goldstein & Spiegelhalter (1996) discuss rankings, particularly the importance of quantifying the uncertainty of a ranking. A major UK review of performance measures in government (for example House of Commons Public Administration Select Committee, 2002) considers the specific application of performance monitoring in a variety of government contexts. Bird et al. (2005) give a broad overview of performance measures and some of the risks, including statistical issues. Bevan & Hood (2006) examine the impact of performance monitoring on the English public health system. Hall and Miller (2009, 2010) explore bootstrap properties of rankings

[1] Taylor Fry, Level 22 45 Clarence St Sydney 2000, hugh.miller@taylorfry.com.au.

and the circumstances where rankings can be expected to be reliable. Paruolo et al (2013) explore challenges with ranking using composite weights, particularly when underlying components are correlated.

2 The case for risk-adjusted performance

The simplest type of performance comparison is to choose a measure and report average rates directly. For example, many states in the USA have introduced surgeon-specific, or hospital-specific, surgery outcome rates, reported as straight averages. This should help patients – they can set appropriate expectations for their surgery and have information for which doctors have historically had high surgery success rates.

However, in many cases the reporting of simple averages led to massive distortions in surgeon behaviour and deterioration in patient outcomes. Doctors were conscious of their ratings (perhaps more so than patients). As a result they avoided taking on difficult cases and treated those who required simpler operations in an effort to boost average survival rates. Further, the doctors considered the 'best' reputationally who didn't alter their caseloads in this way were rated poorly since their patients had significantly lower expected rates of surgery success to begin with. Dranove et al (2003) reported higher patient readmission as a result; the behavioural changes lowered overall health outcomes.

There are circumstances where simple averages work. The most obvious is when the caseload is randomly allocated across providers – although this is relatively rare. In other situations, the caseloads may be similar enough that such comparisons are reasonable, even if the allocation is not formally randomised. For example, ombudsmen complaint rates in the telecommunications sector are taken as a measure of service, on the assumption that service offering and customer compositions are homogenous enough that more detailed comparison is unnecessary. This, in our experience, tends to be the exception rather than the norm. For example, even if programs are run in different locations under exactly the same setup and guidelines, with different providers servicing different regions, regional factors will tend to create differences in caseloads.

Another current example relates to performance-based funding for universities[2]. The proposal at the time of writing, which has itself evolved over time, is that universities would be able to access additional funding based on a set of measures (most likely related to student attrition, completion, satisfaction and post-graduation outcomes). The variation among student cohorts across universities was a significant concern, with no detailed models to correct for it (some relatively crude adjustments were proposed).

This leads naturally to the need for risk-adjustment, where performance recognises the characteristics of a provider's caseload. While this could be done at a provider-level, we deal with the more common situation where adjustment is done at the case (individual) level. If a case can be categorised as 'easy' or 'hard', then the penalty for not achieving an outcome can be recognised in the appropriate context. This obviously aligns with general principles of risk-adjustment in insurance pricing, so it seems an area that actuaries can meaningfully contribute to.

There are many risk-adjusted programs in existence. We give three examples:

- Jobactive and Disability Employment Service (DES): These are the primary employment programs in Australia to assist people with full or part-time work obligation. Employment assistance is delivered by third-party providers. A range of measures are recorded and used for both performance and payments (e.g. whether a person obtains a job and sustains it for 26 weeks). For each measure a risk model is fitted so that actual divided by expected performance can be calculated at a provider level. The measures are then standardised and combined to be converted into a star rating for each provider.

[2] See https://www.education.gov.au/performance-based-funding-commonwealth-grant-scheme

- Surgery in Europe and the United Kingdom: The analysis by Bridgewater et al. (2003) compares surgeon-specific mortality outcomes for doctors in the United Kingdom. While there are large differences between doctors, the proportion of high-risk patients varies significantly and needed to be controlled for. Looking at mortality for low risk outcomes all surgeons had confidence intervals that included the mean mortality, but there was a statistically meaningful relationship between volume and mortality (surgeons performing more surgeries saw lower mortality). The distinction between high and low risk patients uses the EuroSCORE measure, which itself is a well-researched means of understanding cardiac surgery risk, albeit with some limitations (see for example Gogbashian et al, 2004).
- Scaled examination results for calculating Australian university entrance rankings (Australian Tertiary Admission Rank, or ATAR) represents another type of risk-adjustment (here 'provider' represents students and 'cases' individual subjects). In NSW for instance, students take different subjects which must be scaled for relative difficulty before calculating average grades.
- The NSW Nominal Insurer's Workers Compensation portfolio had a range of incentive fees and performance measures across approved insurers for a number of years. It had some provider (insurer)-level risk adjustment to recognise differences in industry and employer size mix.

3 Model design

3.1 Choosing an outcome

Performance comparisons are usually made using one or more measures (or 'metrics'). The choice of measure is obviously important and ideally set by program experts who know what good activity looks like and understand the risks of perverse incentives due to imperfect measures. Bevan & Hood (2006) recount the experience in England where an election target that people would not wait longer than 48 hours to see a doctor immediately led to many practices refusing to take bookings more than 48 hours ahead. A startling improvement in the measure occurred despite more muted underlying change, and arguably increased patient frustration at not being able to plan their weeks.

BINARY VERSUS MORE DETAILED SCALES.

It is common to define and use clear binary measures. For example, Australian employment service programs' (jobactive and DES) performance models put significant weight on the '26-week outcome', which is achieved when people are placed in employment and remain continuously employed for 26 weeks. Binary measures are often significantly easier to collect (e.g. employment status need only be checked at one timepoint) and can be linked to concrete outcome payments (in employment services, providers receive a fee for 26-week outcome payments). However, we do note the potential drawbacks for both statistical analysis and program operation:

- A binary variable carries less information statistically. In the example above, recording the number of weeks in employment, capped at 26, would distinguish between providers that achieve longer spells for jobseekers who do not reach 26 weeks. Further, for binary outcomes with very high or very low rates of attainment, a large number of observations is required for each provider before meaningful comparison can be made. The manufactured example in the numerical section implies using a continuous measure is analogous to having 80% more data. In short, better performance comparisons will often result from continuous measures.
- Binary measures can encourage gaming, particularly for cases close to the threshold. A provider will generally have the incentive to do 'just enough' to achieve an outcome. Or alternatively, in cases where the case has a low chance of reaching the threshold, a provider may expend little energy at all. For injury schemes with incentives at particular stages of claim development, this will often generate increased activity in claims approaching those stages. This lumpiness in case management attention may not be optimal, being induced by program design.

ALIGNMENT WITH PAYMENT STRUCTURES

Many programs are seeing an increased emphasis on 'outcome-based' payments for providers, where payments rise and fall with performance rather than just service delivery. In general, measures that are good for measuring performance will also be good for outcome-based payments and will unambiguously define what 'good' performance looks like for providers. It can also provide a natural mechanism for data collection, since providers can see the value in collecting correct outcome data that is linked to payments.

Potential risks remain if the choice of measure is suboptimal. Aligning outcome payments and performance measures will also deemphasise any program elements that are not part of the measurement framework, which would then need to be encouraged in other ways.

3.2 Deriving performance scores

RESIDUAL VERSUS MODEL PARAMETER PERFORMANCE MEASUREMENT

Once one or more measures are selected then a formula is needed to convert it into a provider-level score. Without loss of generality, let higher scores indicate good performance. The two most straightforward families of approaches are:

1. **Residual models.** A model of expected performance is built (excluding provider as a predictor) and the actuals are compared to expected for each provider. Providers with a large positive residual (actuals compared to expected) are labelled good.
2. **Parameter models.** Provider is included as an explicit categorical effect in the model. Providers with a large positive parameter are labelled good.

Residual models appear more common and have an intuitive appeal (it is easy to explain 'we expected you to achieve X but you actually achieved Y'). However, there are some limitations:

- If providers are correlated with other variables and have a material fraction of particular subgroups, then the resulting model will be biased. For example, if a provider specialises in a particular injury type, and has a large fraction of cases with that injury, then good provider performance may be associated with the injury (an easier injury to manage), rather than the provider. The direction of bias will be towards over-normalisation (that is, a good provider appearing less good, or a poor provider less poor). This is explored further in the numerical results.
- The choice of residual formula may generate outliers that make performance measures more volatile. This is a particular risk if residuals are not naturally tied to the underlying model type.

The main risk of a parameter model approach is that if there are many providers (or a material number of providers with low numbers of cases) there is a risk of model overfitting or even nonconvergence. If this risk is controlled, we generally prefer starting with parameter-based approaches.

There are also situations where results will be consistent (or close to consistent) under both approaches.

CHOICE OF RESIDUAL FORMULA FOR RANKING

For residual approaches, it is worth ensuring that scores should match an intuitive feel for good performance. For example, if Provider A and B each get an 'easy' case and a 'hard' case, and they both achieve one outcome but only Provider A achieves it for the hard case, is it fair to say performance is the same? Or should Provider A be rewarded in their score for a harder outcome, even though they missed an easier one? This is effectively a question of symmetry in the residual formula.

To add some concreteness, denote actual and expected outcomes for case i as A_i and E_i respectively, with high values of A_i indicating more outcomes and higher E_i 'easier' observations for achieving high outcomes, then two common ways to compare providers are the average gap or the relative performance:

Average difference: $$\frac{1}{n_k} \sum_{provider\ k} (A_i - E_i)$$

Relative difference: $$\frac{\sum_{provider\ k}(A_i)}{\sum_{provider\ k}(E_i)}$$

Both measures give higher scores to the better performing providers, with the difference apparent when providers have different average E_i. If two providers have the same performance under an average difference approach (e.g. 0.1 more outcomes per case) the provider with the lower average expected outcomes will have a better score on the relative difference.

While these are both natural ways of thinking about performance, both rely on the total number of actuals $\sum_{provider\ k}(A_i)$ as sufficient information for actual performance. This means there is no distinction between outcomes achieved on 'hard' cases (high A_i when E_i is low) versus outcomes on easy cases (high A_i when E_i is low), so long as the total sum of actual outcomes is the same. This may not be appropriate if extra effort on hard cases deserves reward in a performance setting.

Other alternatives exist that attach greater reward to performance on hard cases, for example:

Cupped average difference (for δ>0): $$\frac{1}{n_k} \sum_{provider\ k} \max(A_i - E_i, -\delta)$$

Weighted relative performance (for $0 \leq \gamma \leq 1$): $$\frac{\sum_{provider\ k}\left({A_i}/{E_i^{\gamma}}\right)}{\left(\sum_{provider\ k}(E_i)\right)^{1-\gamma}}$$

The cupped average difference imposes a floor for a poor outcome, so that 'easy misses' do not count unduly relative to other outcomes.

The weighted average relative performance indexes a range of measures. When $\gamma = 0$ it recovers relative difference, whereas when $\gamma = 1$ it takes the average of the ratios, $\sum(A_i/E_i)$, which is much more heavily skewed towards providers who achieve outcomes when the expectation E_i are low. The results for $\gamma = 1$ can be quite volatile if there are large variations in the E_i, so moderating with an intermediate value of γ may be appropriate.

There is also the opportunity to use a residual formula that aligns with the underlying distribution; for a GLM-like model this means using deviance residuals, that will generate some favourable statistical properties. Ordinary least squares regression aligns with the average difference measure, for instance.

3.3 Combining measures

A complex program such as an injury scheme will have more than one measure of interest. It is common (but not always necessary) to combine these into a single score to assess overall performance. For cases where the measures are primarily used for incentive payments, such combination may not be necessary. Critics of composite indicators argue that the normative implications of such weights can be hard to justify (Stiglitz et al., 2017). On combining measures, we note:

- If program managers have a good feel for the relative importance different measures, weights can be selected by judgement. In other cases, statistically 'optimal' weights can be derived by other methods (e.g. The derivation of ATAR scores in NSW, where subject scaling takes the role of weights).

- Attaching appropriate uncertainty of a performance comparison is more complex for a combined measure, particularly if there are correlations between measures. Estimating and conveying uncertainty is still important though.
- Correlations between measures can also increase the relative importance of the correlated factors beyond the intended weights. Using the example of Paruolo (2013), if three measures are included in a composite index ($Z_i = Y_{i1} + Y_{i2} + Y_{i3}$), with $cor(Y_{i1}, Y_{i2}) = cor(Y_{i1}, Y_{i3}) = 0$ and $cor(Y_{i2}, Y_{i3}) = 0.7$, then Y_{i2} will have three times the influence on Z_i than Y_{i1} in terms of correlation between input and composite.
- Scaling is also an important consideration. When measures operate on different scales, the need to standardise before combining is clear. However, even when the scales are similar (e.g. all binary outcomes), some care is needed. If one measure is naturally more variable on the scale (e.g. one binary measure has an average of 50%, whereas another has an average of 95%, with corresponding standard error more than 50% lower), then the more variable measure will have outsize influence on the combined score.

3.4 Interpretability

One common complaint of risk-adjusted performance measurement is its opacity. A provider can understand 'percentage of clients who found a job', but struggle with 'actual divided by expected, where expected is a risk-weighted model of the caseload'.

We mention two broad approaches to improving the communication and transparency of a risk-adjusted performance model; unpacking provider-level risk-adjustments and focusing communication on the program outcomes that providers have control over.

UNPACKING PROVIDER-LEVEL RISK ADJUSTMENT

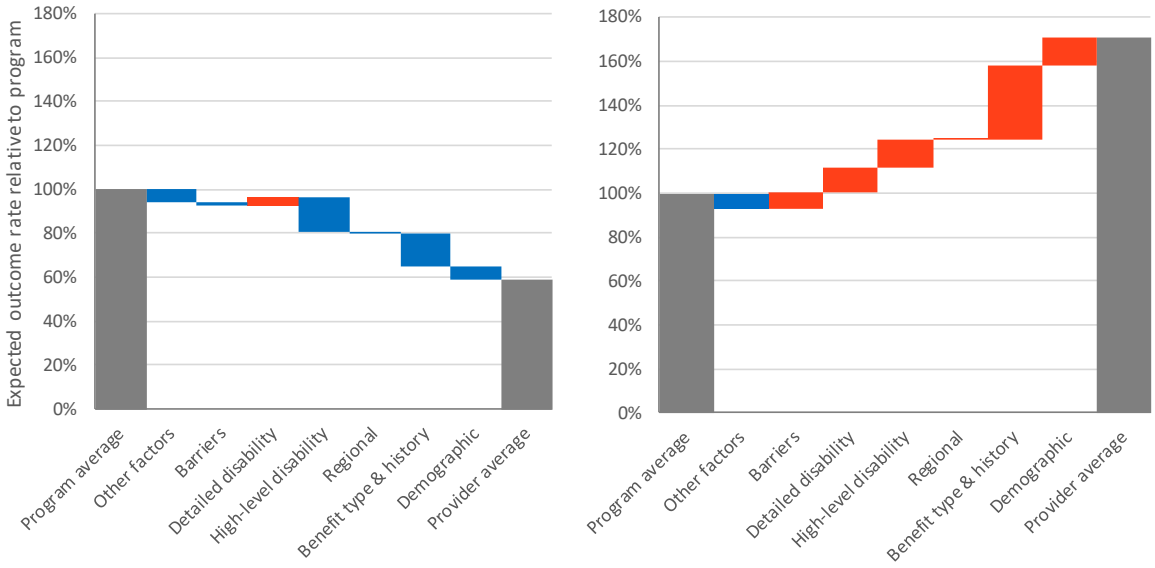
The 'actual' outcome achievement is usually easy to understand. The 'expected' portion can be unpacked using standard model interpretation techniques (for example see Molnar, 2019). Perhaps the most straightforward approach is a 'score-carding' approach This involves:

- Taking the modelling dataset and switching all predictor variables to some 'average' value. For instance, age = 35 for all observations. Predict expected outcomes rates by provider (in this case, all providers get the same score)
- Choose one (or a set) or predictors to toggle back to their original values. Re-score the model and calculated expected outcome rates
- Continue (cumulatively) toggling variables until all predictors are in their original form. The final scored values will give the actual provider-level expected outcome rates.

There are some subtleties to the approach (each step may need some bias correction, and the order of toggling can be important, particularly if there are lots of interactions in the model), but the result is an indicative split of which predictors are driving the expected outcome rate for each provider.

We show an example of such unpacking in Figure 1. This takes the 26-week employment outcome model from the DES program and presents illustrative (but not actual) scorecard data for two hypothetical providers. The dozens of variables are grouped for easy of reading. The left panel shows a provider with lower than average expected outcomes (a 'hard' caseload) primarily driven by disability type and the benefit type and history. The second provider has an easier caseload, expecting 70% more outcomes than the program average, spread across a range of factors, but benefit type & history making the greatest contribution of any variable group.

Figure 1 Unpacking predictor effects, two hypothetical providers from the DES program



PROVIDER COMMUNICATION

Some of the perceived opacity can be removed with appropriate communication also. For example, it is possible to automatically generate provider-level reports that explain:

- Their expected (predicted) and actual outcome rates
- Why their predicted outcome rate varies from the overall program average (unpacking the adjustment)
- What targets they would need to achieve to reach certain benchmarks (e.g. increase in outcome rates required to become a top-quintile provider)
- The change in their outcome rates over time, and the impact on their assessed performance.

Feedback to providers is seen as a key source of process improvement, so time invested in detailed feedback can be of great value.

3.5 Use of generalised linear models

RATIONALE FOR GENERALISED LINEAR MODELS

Most of the examples referenced above (NSW Workers compensation and NSW examination scaling being notable exceptions) make use of a regression model to calculate expected outcomes for each case.

Generalised linear models (GLMs) and its extensions (such as penalised GLMs or models that include random effects) are common in setting such expectations. While not the only option, GLMs offer several advantages. Compared to simpler approaches (such as segment-based adjustments, where a caseload is split into a small number of buckets, each with a single expected):

- GLMs estimate multivariate effects, so that more factors can be incorporated into a prediction
- GLMs can estimate continuous predictors, such as age, so that expected outcomes similarly vary continuously.
- GLMs simultaneously estimate the individual-level uncertainty (distribution around the mean), plus the parameter uncertainty attached to the estimates.

Relative to more complex machine learning approaches, GLMs also offer some benefits:

- GLMs enable formal hypothesis testing for variables. This is important since overfitting can be a concern for performance models (see section 4.4), and so excluding unnecessary variables is one of the best ways to protect against this.
- GLMs generate explicit estimates of partial effects (how a variable affects prediction, holding other factors constant). This is important for auditability and accountability of the model, particularly when gauging change over time.

3.6 Use of samples as representative performance

For some programs, outcome collection may be expensive (e.g. a compliance-related measure may relate to administrative correctness, only detectable through a detailed review of a case). In such cases a sample may be used rather than collecting data for every case, and the design of the sample becomes another statistical question in its own right with follow-on implications for any risk-adjusted model. Major considerations are:

- Sampling sometimes must balance competing objectives. If the key focus is correctly comparing different providers, then this would suggest similar numbers in the sample for each provider (so that the provider-level standard errors are comparable). In cases where the size of providers varies significantly this can lead to significant over-sampling of small providers, relative to a simple random sample. This is statistically inefficient if a tight estimate for the overall program is required.
- Risk-adjustment could conceivably be (wholly or partly) controlled for in the sample design itself. If cases span clear categories, with different outcomes rates in each category, then a sample might be stratified by category (reducing sample variability) and might also have the same ratio of cases across categories, controlling for the category effect.
- Regardless of the sample stratification, an estimate of total outcome rates is recoverable through correct use of sampling weights. See for example Lohr (2019).

3.7 Uncertainty and credibility

Expressing the uncertainty in a ranking of providers is an issue of fundamental importance; providers may appear quite different on a ranking but be statistically indistinguishable. This is particularly the case when some of the providers have smaller caseloads, and correspondingly larger errors associated with estimating their performance.

Placing confidence intervals around a measure or a ranking is the obvious first step, as it conveys the uncertainty.

For a parameter-based model of provider performance (and to a lesser extent the residual approach) there is the additional option of applying penalisation or a credibility approach to the parameters so that more volatile estimates are shrunk back further towards the average. Our numerical work in Section 5 suggests this is an effective way of improving the estimation of provider performance – although this will generally have a lesser impact on the ranking.

3.8 Checking fairness of the performance model

One challenge of a risk-adjusted performance model is ensuring providers have similar capacity to achieve 'good' or 'bad' scores. While this is obviously the whole point of risk-adjustment, there is often a need to test whether this is actually the case. For example, if using a simple residual approach (average distance or relative difference, as per Section 3.2) with a binary measure there are upper and lower bounds for each provider (all zeros or all ones). For a provider with a high expected score, their maximum performance is capped at a lower level than other providers, and this cap may ensure they sit below other providers even if they achieve a 'perfect' score. If a provider can only cannot achieve strong performance due to the performance formula this will be perceived as unfair.

3.9 Interactions and differential performance

The discussion throughout the paper focuses on a single estimate of overall provider performance on a particular measure. It may be that a provider performs well with particular types of cases and poorly on others. The parameter-based model approach allows this to be tested through interactions between provider parameters and other caseload-related variables in the risk model. Standard interaction detection techniques can automatically detect candidate interactions for providers that perform strongly on certain subgroups. This may not feature in the formal reporting of performance measurement but may give important information relevant to managing the program.

In a similar fashion, performance measures can be calculated on subgroups of the caseloads (e.g. particular injury types) to see how provider rankings differ on a cohort of interest.

4 Variable selection

4.1 Introduction

The choice of predictor variables in the risk-adjusted model will fundamentally affect the final performance measurement and any associated provider rankings. This section explores variable selection issues that are particularly relevant to these types of models.

4.2 Additional diagnostics

Standard model diagnostics are still relevant for a performance model used to set expected levels of outcomes. However, other diagnostics may also be warranted.

One key question is the sensitivity of the ranking to individual predictor variables in the risk-adjusted model. This is particularly important if the comparison of provider rankings is being circulated to a wide audience. Some variables might be of only moderate importance for individual-level prediction, but much more relevant for the rankings; traditional diagnostics would not recognise these effects. Two tests we propose are:

- **Removing the variable:** If the variable was not in the model (either by setting to a mean value and rescoring, or by refitting without the variable), how much does the ranking change?
- **Perturbing the variable:** If we perturb the parameter attached to a variable by one standard deviation (or test a range of deviations based on the observed standard error), how much does the ranking change?

These tests identify variables that are most important to the performance ranking. Often these will align with the most important variables, but not always; regional variables are a good example where, if providers are location based, the effect at a provider level will be more marked than other individual-level predictors that tend to average out.

With respect to a particular variable, the first test is a measure of importance, whereas the second is a measure of model uncertainty and robustness; these have different implications. If a variable is important, than it means that ensuring good data quality for that variable is of heightened performance. Conversely, if a variable has high model uncertainty, then consideration is needed whether the variable is appropriate at all – or alternatively one should consider how best to communicate the uncertainty. For providers that are limited to different regions, regional factors are a common example of variables that have low to medium importance and high uncertainty; see Section 4.3.

4.3 Correlations and regional effects

Section 3.2 introduced the concept that correlations between predictor variables and provider caseload can be problematic, particularly for residual-based approaches. At the very least, variables that are strongly predictive of provider *caseload*, as well as outcome measures, should be understood. The diagnostics in Section 4.2 should also highlight these.

A common correlation relates to regional effects. Providers will often have higher market shares in certain areas. As an extreme example, for employment services the regions of operation are contractually set. Since outcomes almost always vary by region irrespective of provider, this creates a natural correlation to be unpicked.

In some cases the issue is genuinely insoluble; if a provider works only in one region and has a large market share, there is no way to untangle the two effects (region and provider) without some judgement or assumption. More generally, there is a clear choice in modelling regionality between regional 'indicator variables' and economic variables that vary by region (e.g. unemployment rates or a socioeconomic score). The indicator variable approach effectively defines regions that become their own subdomain and providers are rated on their performance relative to the average for that subdomain. This may be attractive but has a large limitation in that it means that provider comparisons across regions are impossible. It is entirely plausible that providers in one region could be better, on average, than those in another, but this is not testable under an indicator approach.

The use of economic variables is more parsimonious, allows comparison across regions and provides some insight into why regions have particularly high or low rates of outcomes. Their main drawback is that it is unlikely that the variables will capture the full range of regional variation across a state or country; in this case underfitting is a risk that will disadvantage providers in a 'poor' regional environment. Routinely reviewing regional (risk-adjusted) performance can draw attention to this risk if it exists.

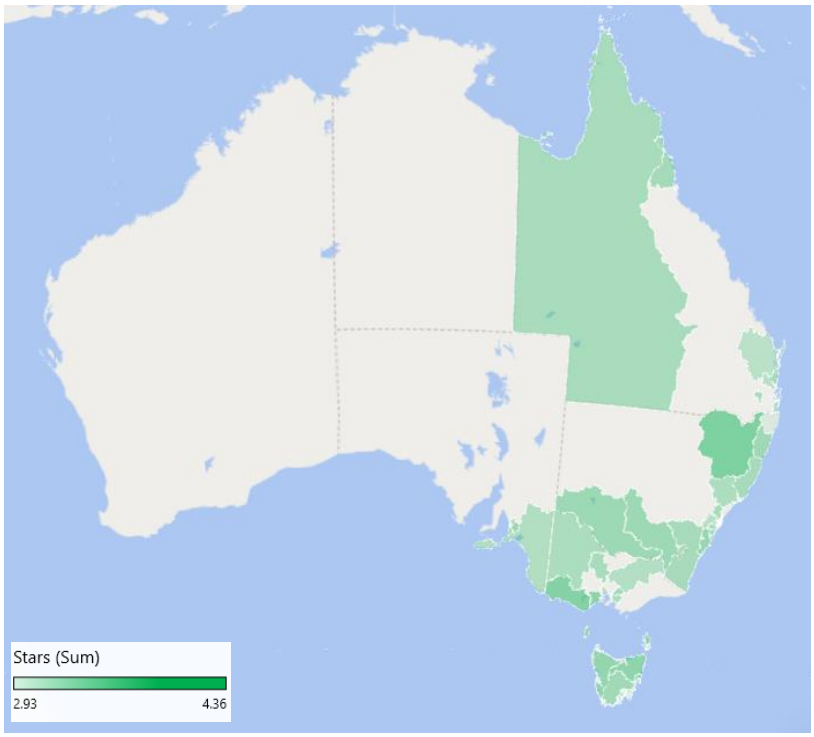
Figure 2 shows a heatmap at the (2011) Statistical Area 4 level for star ratings for the jobactive program as at June 2016. Providers within a given region are given a rating out of five, with 2.5 stars being the rough average. The heatmap shows only those regions with an average of 2.9 or higher. If shows that above average performance is seen:

- In all Tasmanian regions
- For the whole of Sydney and NSW coast.
- For regional VIC, parts of Melbourne and South-Eastern SA
- Nowhere in WA, NT or regional SA.

The regional patterns are striking and suggest that economic drivers may be affecting comparisons in a way not captured by the risk-adjustment process. Changes were subsequently made to the jobactive star ratings models to improve these observable effects.

The results are likely partly attributable to the mining downturn (particularly mining investment) that was evident at the time, which caused lower economic growth in WA and parts of Queensland particularly.

Figure 2 Average star ratings by statistical area 4, June 2016 jobactive star ratings. Regions with average star ratings above 2.9 are shown.



4.4 Impact of overfitting

The appropriate number of variables to include will depend on the context, and should also rely on standard statistical tests for inclusion. Unnecessary variables should be deleted to prevent overfitting.

However, special mention ofone extreme impact of overfitting is warranted; cases where the outcome is binary and a categorical variable has levels where all responses are zeros or ones. In this case the expected value will be driven to the extreme – if the responses are all ones then the corresponding E_i will be very close to one. This effectively deletes the residual (or signal) attached to those observations from the performance model – as actual versus expected are exactly matched. So in this extreme version of overfitting the impact is roughly analogous to assuming average performance for these observations. Providers affected will be shrunk towards the mean as a result.

Our testing suggests that provider level performance is fairly robust to overfitting, assuming the nuisance parameters are uncorrelated with provider. The impact on the model at a case (individual) level will tend to be more significant, so providers with small exposure or other applications that relay on small groups will suffer more in the presence of overfitting.

4.5 Change over time

Performance measures are often reported regularly to see if there are improvements in a program over time. As part of this, the relative rankings of providers will also change. This variation will arise from:

- Genuine improvement or deterioration in provider performance
- Compositional change
- Changes to the underlying risk-adjusted model, assuming it is refit (either a parameter update or a broader refit that reviews included variables)
- Statistical noise. This includes parameter error in the risk-adjusted model and process error attached to individual cases.

Too much volatility in the rankings over time will undermine confidence in the ranking (and rightly so!). Conversely, too much stability might mean that improvements by individual providers are not being appropriately recognised. The general approaches to controlling variability are:

- Selecting an appropriate time window, (or sample size for cases where a sample is drawn). Including multiple years of data will stabilise the results (both the risk-adjusted model and the provider rankings) but make it less responsive to recent trends. Or the risk model can be fit on a longer time window but applied to a shorted one for calculating a residual model.
- Choosing not to update the risk model. While this can ignore changes to the way outcomes are changing for different cohorts, it does reduce one source of variation in situations where the analyst believes the risk model is evolving slowly, if at all.
- Monitoring confidence intervals at a provider level gives a view of how changes compare to expected statistical noise.
- Incorporating more formal filters and time series elements into the statistical model.

As with other subsections, there are natural ways to monitor variation from each of the sources above. A variant of score-carding can unpack compositional change, while scoring the old and new risk models can show the impact of these changes.

5 Numerical examples

5.1 Primary measure of accuracy – correctness of provider rankings

Our numerical work focuses on the accuracy of a ranking. Suppose we have provider-level scores (or effect sizes) γ_k, which are estimated through some procedure $\hat{\gamma}_k$. The true effects induce a ranking R_k, where $R_k = 1$ for the k when γ_k is largest , $R_k = 2$ for the second-largest γ_k and so on. Similarly let $\hat{R}_k$ be the estimated ranking for the estimated γ_k.

Define the accuracy of the estimated ranking as the Kendall's Tau (rank correlation coefficient) score, which can be written as:

$$\mathcal{R} = \frac{\sum_{k \neq k} \mathbb{I}\left(R_k < R_{k'} \text{ and } \hat{R}_k < \hat{R}_{k'}\right) - \sum_{k \neq k} \mathbb{I}\left(R_k < R_{k'} \text{ and } \hat{R}_k > \hat{R}_{k'}\right)}{\sum_{k \neq k'} \mathbb{I}(R_k < R_{k'})}$$
$$= \frac{\#\ Pairs\ ranked\ correctly - \#\ Pairs\ ranked\ incorrectly}{\#\ Pairs\ of\ scores}.$$

This is a nonparametric measure of agreement that is robust against the choice of model and focuses on a key question of interest; whether one provider is correctly rated as better than another. A score of zero is random performance, 1 is perfect (and minus 1 is perfectly wrong). Its main disadvantage is that it ignores scale. It does not care that it correctly identifies big or small gaps between provider performance, so it can obscure some of the subtlety of a particular scenario.

For simulated data we can define the true ranking in advance. For real data we assume the ranking on the original dataset is authoritative and use the bootstrap to simulate draws from the population.

5.2 Setup

We start by defining a 'base' scenario and real dataset for results.

Base simulation: Let there be K providers sharing n cases with market shares allocated by performing $K - 1$ random uniform slices of the unit interval, subject to a minimum market share of ϕ. The providers are assumed to impact on a linear predictor related to the outcome, with effect size γ_k for provider k, and each γ_k standard normal but scaled by a global constant G_1. These can be expressed as a vector for the cases $\gamma = \gamma_{k(i)}$, where the value for each case is the corresponding value for that provider.

Additionally, p predictors X_j^* are generated as standard normal random variables for each case. We introduce a correlation between these and the provider effect $cor(X_j, \gamma) = \rho_j$ with ρ_j defined as $\rho_j = G_2(U_j - 0.5)$ for U_j random uniform and G_2 a constant. Define parameters β_j as standard normal scaled by constant B_1. The default response is binomial with probability defined as

$$p_i = logit^{-1}\left(\sum_j \beta_j X_{ij} + \gamma_{k(i)}\right),$$

This is consistent with a standard logistic regression setup and we sample response variable for the outcome Y_i as Bernoulli with mean p_i.

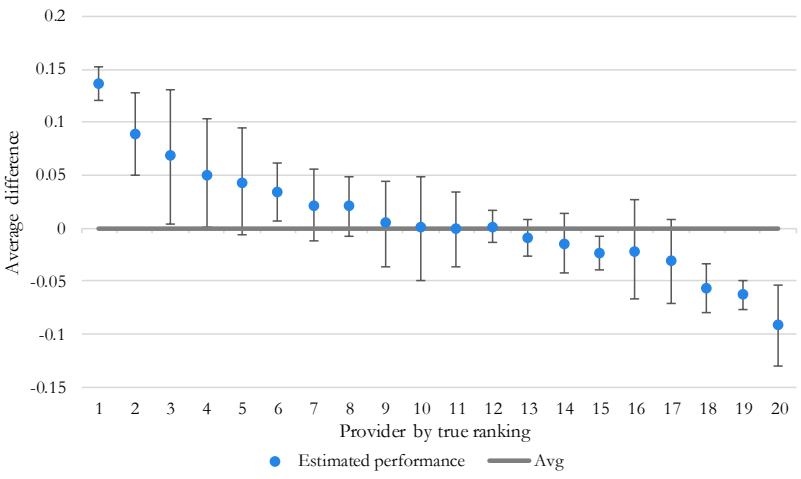
Base dataset: A dataset based on that used for the Disability Employment Service. The primary target is a binary outcome for 26-week full employment outcomes. There are about 300k observations related to individual person-quarters. The default model used in the results below contains about 150 predictors which span jobseeker characteristics, local labour market conditions and specific barrier and condition information. Data is at June 2017, and so does not reflect changes observed since the 2018 reforms to the DES program.

5.3 Simulation results - demonstration of key concepts for performance rating

BASE SIMULATION

We take the base simulation with $K = 20, \phi = 0.5\%, n = 10{,}000, p = 20, G_1 = 0.6, G_2 = 0.2, B_1 = 0.2$. We use a residual model with average difference. For a fixed set of parameters and linear predictors we simulate the Y_i 499 times and measure ranking performance.

Figure 3 Average provider difference, base simulation, averaged over 499 simulations, with 80% confidence intervals shown



In this particular simulation we can see:

- The value of confidence intervals; in this case more than half the providers are not statistically different from zero (average performance).
- The rankings are mostly recovered; the x-axis orders by true ranking and on average these are recovered. This is also reflected in a relatively strong average ranking score of $\mathcal{R} = 0.70$ (s.e. for the mean of 0.0002).

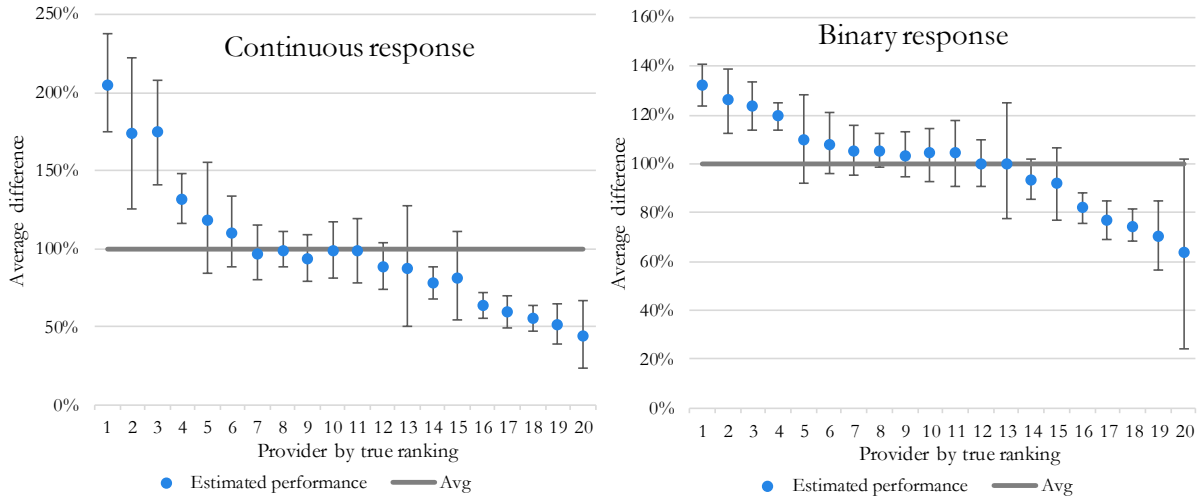
- There is reasonable variability across simulations. The 80% confidence interval for $\mathcal{R}$ is (0.6,0.79), and the figure visually shows plenty of overlap between pairs of providers. It is normal to expect changes in the ranking over time or samples as a result.

BINARY VERSUS CONTINUOUS MEASURES

This simulation shows the potential power of continuous measures over binary indicators (see Section 3.1). Keep the same setup from the base simulation except for:

- Change $n = 2{,}000$.
- Let Y_{1i} be exponentially distributed with mean $\exp\left(\sum_j \beta_j X_{ij} + \gamma_{k(i)}\right)$ and define $Y_{2i} = \mathbb{I}(Y_{1i} > 1)$. This is analogous to modelling claim duration Y_{1i} versus claims finalised at one year Y_{2i}, with the latter representing a binary variable carrying less information than the formula.
- Use Relative difference to rank providers.

Figure 4 Average provider difference, base simulation, averaged over 499 simulations, with 80% confidence intervals shown



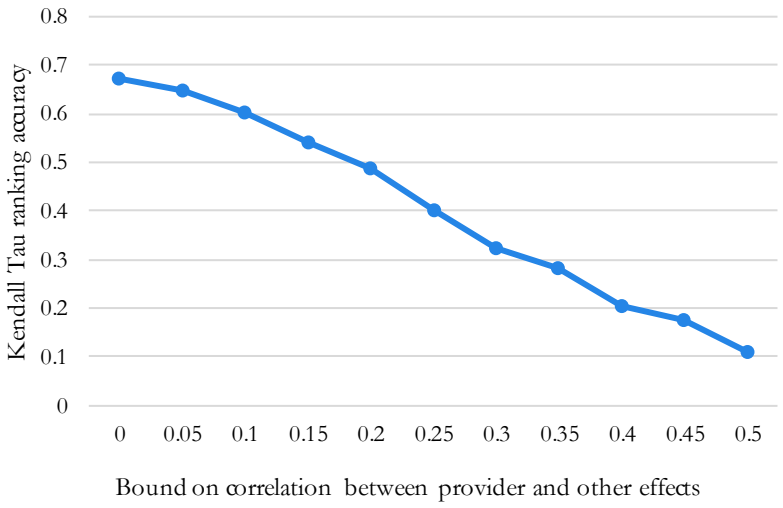
The average $\mathcal{R}$ for the continuous response was 0.750±0.0003, much higher than 0.698±0.0003 for the corresponding binary response model. This is a very large difference in ranking authority; the binary response would require about 80% more data to achieve the same level as authority.

Figure 4 shows the relative difference by provider. The first notable difference is the improved authority at the top end; high performing providers are much easy to spot with the continuous response, more than compensating for their larger confidence intervals. For the lowest ranked provider the binary response also leads to a much larger degree of uncertainty; the binary response obscures much of the signal.

RISKS OF CORRELATION BETWEEN PROVIDER AND VARIABLE FOR RESIDUAL APPROACHES

Let $n = 2{,}000$, but otherwise keep the setup of the base simulation. Vary the correlation scalar G_2 from 0 to 0.5 in increments of 0.05. At higher correlations stronger providers tend to have easier cases, meaning that there is potentially greater risk of mistaking caseload effects as provider performance.

Figure 5 Average ranking accuracy, using residual average difference, under different levels of correlation between provider effects and other variables. 20 simulations



The figure shows a remarkable decrease in ranking accuracy as correlations grow. Note that the average correlation is about half the bound, on average, so even at the right of the plot the correlations are mostly moderate.

The conclusion is that if there are correlations (some providers have easier or harder caseloads), residual methods cannot consistently recover provider effects, with a corresponding reduction in ranking accuracy. Such correlations will not always be an issue but can be.

RESIDUAL VERSUS PARAMETER APPROACHES

The most straightforward solution to the previous simulation is to incorporate provider effects into the model, where parameter estimates are once again consistent; this is the 'parameter approach' of Section 3.2.

Figure 6 Average ranking accuracy, using residual average difference, under different levels of correlation between provider effects and other variables. Comparison of residual and parameter approaches.

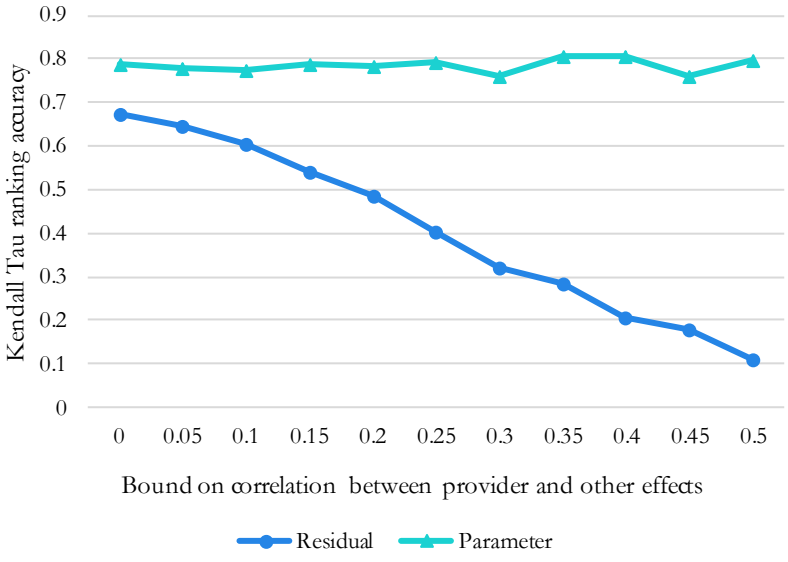


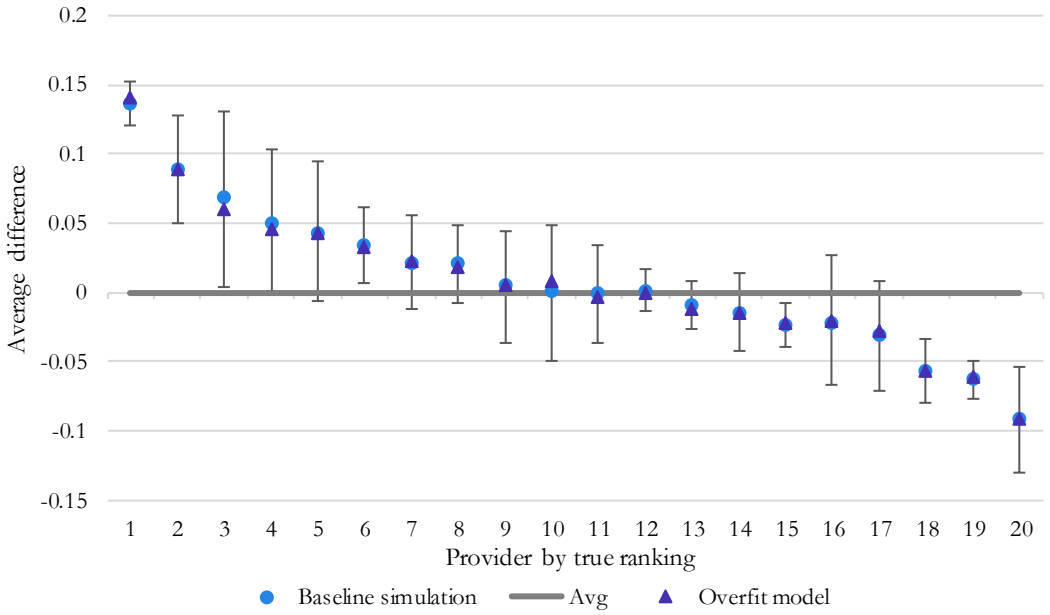
Figure 6 repeats the preceding chart but adds a line for ranking accuracy when the ranking uses parameter effects for when provider is included as a categorical variable in the GLM. The difference is stark:

- Results are materially better even at zero correlation. This is partly due to better alignment between the simulation data and estimation (the simulation is set up as an effect on the linear prediction).
- Results are stable under the parameter approach in the presence of correlations; the dataset is large enough to disentangle the correlated effects. At $G_2 = 0.5$, the residual approach is virtually useless while the parameter approach is just as strong.

OVERFITTING

To demonstrate the some of the implications of overfitting (see Section 4.4), albeit in rather extreme fashion, we take the base simulation and add 200 nuisance parameters Z_j with each a Bernoulli with mean sampled from a uniform distribution. Each Z_j is uncorrelated with all other variables (and the provider allocations).

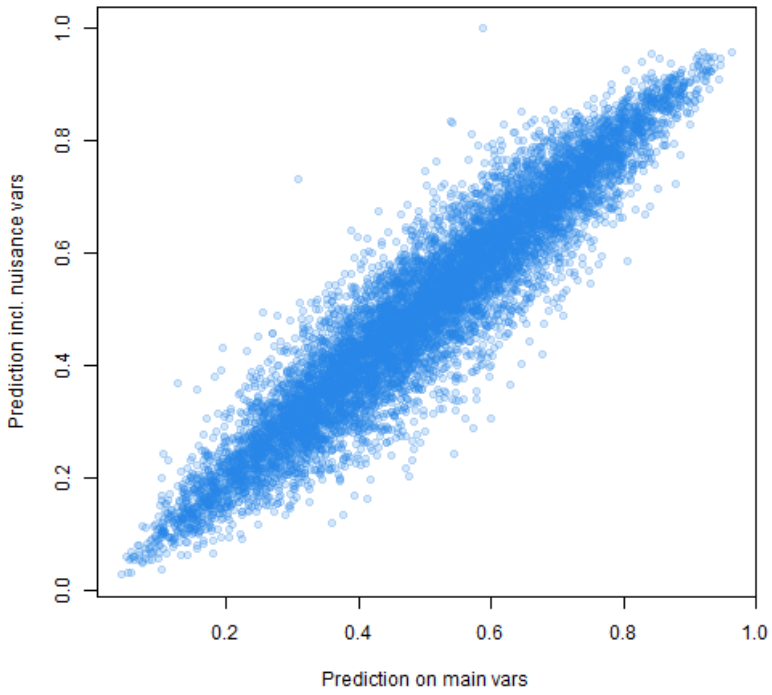
Figure 7 Average provider difference, averaged over 199 simulations, in the presence of 200 additional nuisance parameters with 80% confidence intervals shown. Results overlaid on Figure 3, the base simulation



The results suggest the average differences at a provider level are virtually unchanged. The ranking accuracy is slightly lower 0.685 rather than 0.703, which is still statistically meaningful. The result suggests that as long as the noise variables are uncorrelated with provider, the impact is muted for reasonable sample sizes.

The results belie some quite significant differences at an individual level. A scatterplot comparison is shown in Figure 8. In fact, the standard error of the change (due to introducing the nuisance variables) is about 35% of the total standard deviation in the original fitted values in this example, so a very large portion of the fitted values can be described as noise. In such situations any correlations between the noise and provider, or any reliance on small sample sizes for inference, can lead to adverse results.

Figure 8 Individual level predictions for the base model and overfit models from one of the simulations



CREDIBILITY APPROACHES TO STABILISE RANKING

In situations with lower (and variable) numbers of cases per provider, credibility or shrinkage can improve results (see Section 3.7).

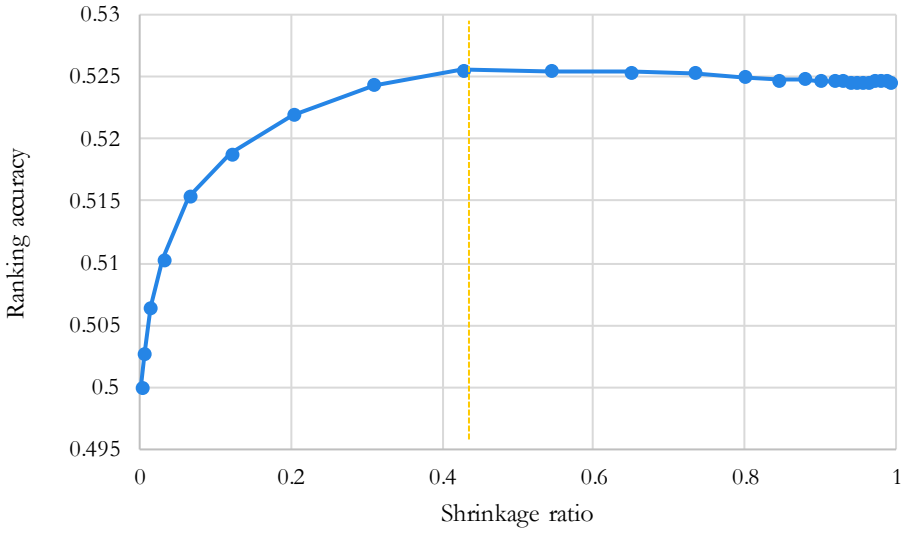
We use the same setup as the base simulation except that we use $K = 40$ providers and randomly select a caseload between 5 and 40 cases each, with n defined as the some of this. We estimate provider effect as a model parameter and apply a ridge regression penalty on the provider terms to apply credibility or shrinkage to the model.

The figure below shows the results. The x-axis is the shrinkage ratio which is defined as:

$$\frac{\sum(\gamma_k^*)^2}{\sum(\gamma_k)^2},$$

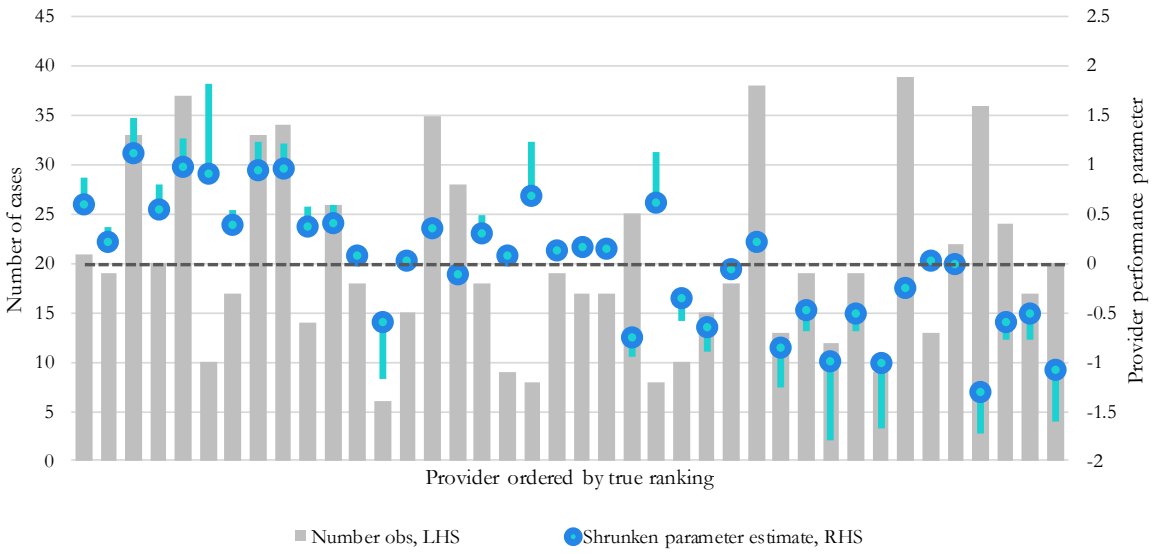
where γ_k^* is the shrunken estimate for provider k and γ_k is the estimate when no shrinkage is applied.

Figure 9 Average ranking performance under various shrinkage estimates, 199 simulations



The optimal degree of shrinkage ratio based on ranking performance is about 0.45, although the improvement in ranking performance compared to the non-shrunken estimate is only relatively slight. More significant is the impact on the provider effects themselves, with the results from one of the simulations shown in Figure 10. The shrinkage is greatest for smaller providers with large parameters. Some cases of improved rankings are visible, but the reduction in extreme provider effect sizes is the more prominent feature visible.

Figure 10 Impact of ridge regression effects on provider parameters. The lines indicate the movement of parameter compared to the model without credibility applied.

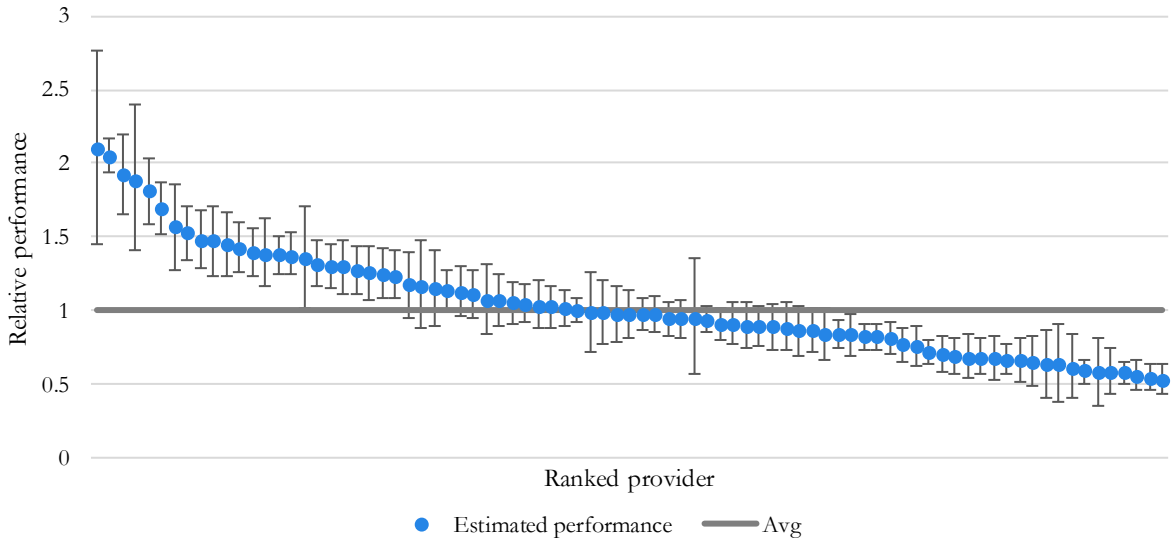


5.4 Real data example

MAIN RESULTS

Applying the GLM and bootstrapping results gives estimated performance and confidence intervals for roughly 400 providers. We show a 20% sample of these in the figure below.

Figure 11 Sample of 20% of providers, relative performance with 80% confidence interval shown



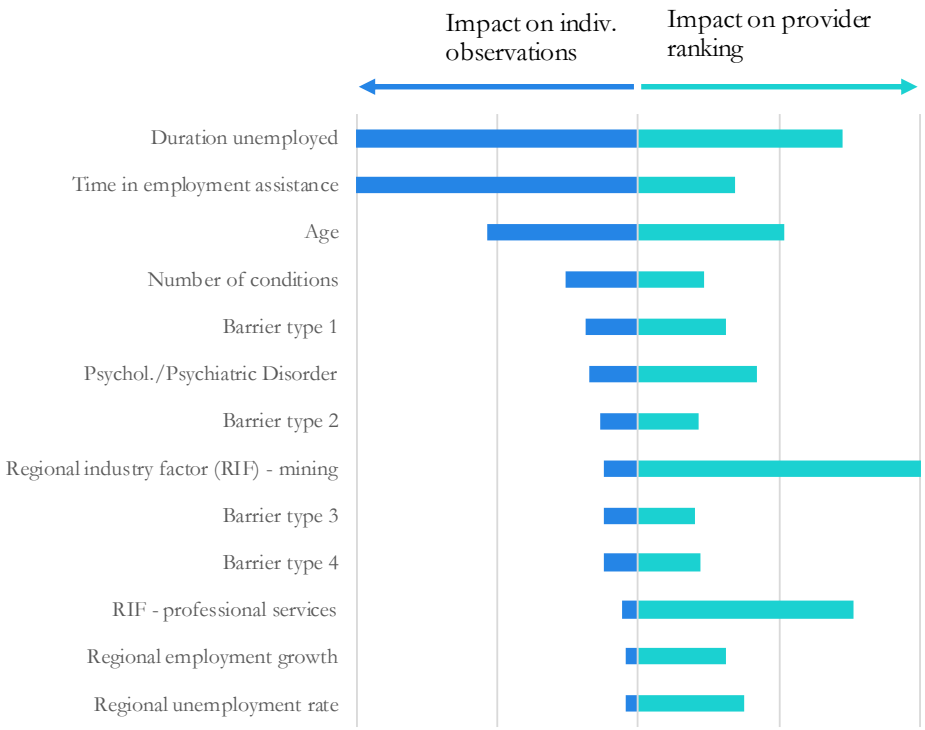
There is relatively good separation of providers under the model; a higher proportion of them are significantly (in the statistical sense) above or below the average, so inferences can be drawn on performance. Interestingly the highest performing provider has one of the largest confidence intervals, suggesting that some of this performance could be attributable to noise – a credibility based approach would certainly pull it back behind others.

The ranking accuracy of the bootstrap is $\mathcal{R} = 0.75$, with a relatively tight 80% confidence interval of (0.74,0.77). This represents reasonably good recovery of ranks, in light of the results shown in Figure 11.

VARIABLE IMPORTANCE

We have measured variable importance in two ways, in line with the discussion in Section 4.2. We have estimated the impact on model deviance (how much the deviance would decay if the variable was zeroed out of the model), which is a measure of individual-level importance. We have also then looked at the influence on ranking (how much the Kendall Tau measure decays if the variable is zeroed out). Results for the top ten variables on either measure are shown in the figure below.

Figure 12 Relative variable importance using a case (individual)-level measure and a provider-level measure. Variable appears if it is in the top ten list under either test.



We observe:

- There is significant overlaps in the lists; only 3 of the top 10 for the provider ranking impact are not in the top 10 of the individual-level list.
- The top two individual level variables relate to duration (unemployed and in assistance) and are very strong in the model. However, they are only ranked 3 and 7 for the provider impact, suggesting that most providers have a spread of different durations.
- Conversely, regional factors, which are more closely correlated to providers, are very prominent in the impact on providers. All three terms not in the individual top 10 (but in the provider top 10) are regional terms. The top two relate to regional industry factors (the proportion of a region employed in a certain industry). These variables help capture the state of an industry – if mining is experiencing a downturn then regions highly exposed to a downturn can be appropriately recognised in the risk-adjusted model.

The differences visible in the figure suggest the importance of considering the provider-level impact of variables, in addition to traditional diagnostics.

6 Acknowledgements

I am grateful to the Department of Social Services for using a subset of their data for some of the results in Section 5, and for numerous people at Taylor Fry for their helpful comments in the construction of this paper.

7 References

Bevan, G., & Hood, C. (2006). What's measured is what matters: targets and gaming in the English public health care system. Public administration, 84(3), 517-538.

Bird, S. M., Cox, D., Farewell, V. T., Tim, H., & Peter C, S. (2005). Performance indicators: good, bad, and ugly. Journal of the Royal Statistical Society: Series A (Statistics in Society), 168(1), 1-27.

Bridgewater, B., Grayson, A. D., Jackson, M., Brooks, N., Grotte, G. J., Keenan, D. J., ... & Mark, J. (2003). Surgeon specific mortality in adult cardiac surgery: comparison between crude and risk stratified data. *Bmj*, *327*(7405), 13-17.

Dranove, D., Kessler, D., McClellan, M., & Satterthwaite, M. (2003). Is more information better? The effects of "report cards" on health care providers. Journal of political Economy, 111(3), 555-588.

Gogbashian, A., Sedrakyan, A., & Treasure, T. (2004). EuroSCORE: a systematic review of international performance. *European Journal of Cardio-thoracic surgery*, *25*(5), 695-700.

Goldstein, H., & Spiegelhalter, D. J. (1996). League tables and their limitations: statistical issues in comparisons of institutional performance. Journal of the Royal Statistical Society: Series A (Statistics in Society), 159(3), 385-409.

Hall, P., & Miller, H. (2009). Using the bootstrap to quantify the authority of an empirical ranking. The Annals of Statistics, 37(6B), 3929-3959.

Hall, P., & Miller, H. (2010). Modeling the variability of rankings. The Annals of Statistics, 38(5), 2652-2677.

House of Commons Public Administration Select Committee (2002a) Examination of Witnesses (Professor Tim Brighouse, Mr David Butler, Director, National Confederation of Parent Teacher Associations, Dr Gill Morgan, Chief Executive, NHS Confederation and Mr Mike Newell, President, Prison Governors' Association), Dec. 5th. Questions 362–451.

Lohr, S. L. (2019). Sampling: Design and Analysis: Design and Analysis. Chapman and Hall/CRC.

Molnar, C. (2018). *Interpretable machine learning: A guide for making black box models explainable*. E-book at https://christophm. github. io/interpretable-ml-book/

Paruolo, P., Saisana, M., & Saltelli, A. (2013). Ratings and rankings: voodoo or science?. Journal of the Royal Statistical Society: Series A (Statistics in Society), 176(3), 609-634.

Stiglitz, J. E., Sen, A., & Fitoussi, J. P. (2017). Report by the commission on the measurement of economic performance and social progress.