Improving Incentives for Fairness in QBS:
A “need-adjusted” approach to coverage

The DIME ECM IE Team

20 May 2020

Key Messages

1. The Quality Bonus System (QBS) scheme has provided performance bonuses to Estonian GPs since 2006. To date, bonuses have been allocated based on a score related to the **proportion of a GP's eligible patient population who receive a specific set of primary care services**.

2. However, **health disparities across patient populations give rise to the concern that the current QBS scheme does not equally reward GP efforts**. It is significantly easier for GPs serving a relatively young, healthy population to receive QBS rewards than their colleagues serving populations with a higher proportion of complex cases.

3. To make the QBS scoring system fairer for all GPs, including those facing populations with high levels of need, we propose **three data-driven adjustments** to the QBS scoring system:
   a) *Re-weighting the indicator scores* based on the experience of the scheme to date
   b) Awarding *proportional credit* at the indicator level rather than using thresholds
   c) Adjusting coverage scores based on the *patient need* for each provider

4. **These adjustments can be calculated using only the data that is already used for QBS**, and this report includes an Excel spreadsheet with formulas for doing so.

5. The system extends easily to include **additional indicators** on a temporary or permanent basis, such as for work related to COVID-19.

6. We do not propose any specific system for assigning the **financial incentives**, but rather focus attention on easy-to-implement adjustments that could create a fairer scheme.

Introduction

Pay-for-performance (PFP) schemes have grown in popularity globally over the past 20 years, including for primary care. These programs have yielded mixed results in their various implementations across countries and health systems. Evidence from the longest-running programs (such as in the UK) suggests that they can contribute to improvements in the quality of primary care, but they are far from a "silver bullet" and that careful attention must be paid to program design, to avoid unintentionally misaligned incentives.

Estonia's pay-for-performance mechanism, the Quality Bonus System (QBS), has been in place since 2006. It was initially voluntary, and it accounts for a relatively small amount (2-4%) of total provider compensation (Merilind et al 2016). **The QBS system is not intended to motivate providers primarily through cash rewards, but rather highlight effective patterns of primary care.** As such, QBS should accurately reflect the different efforts GPs have to make to serve the different populations they face across the country. Otherwise, it may not succeed in its intended purpose as a form of recognition for merit and achievement, or as a tool to make issues of quality salient to family doctors. This understanding of the goals of QBS informs the analysis that follows.

Assessing the distribution of chronic disease burden[1] across clinics, two main issues stand out:

1. **The disease burden of chronic conditions is distributed highly unevenly across clinics.** Provider lists fall roughly into three categories with respect to chronic conditions: (A) lists with relatively younger and lower-risk patients on average; (B) lists with older, but low-risk, patients on average; and (C) lists with older, at-risk patients on average. See **Figure 1**.
2. **The QBS system as currently implemented makes it difficult to distinguish between these three groups and does not account for them in its scoring.** The current system awards points at "coverage thresholds" when a provider has delivered the required service to a fixed proportion of the targeted patients for each indicator.

[1] We use "chronic disease burden" in a non-technical sense in this report, and we refer broadly to all conditions that fall under either the ECM program or the QBS program.

This structure produces two "reverse incentives" that are common pitfalls in schemes of this type. First, the coverage basis rewards providers who have less need *overall* in their patient population, since they may only need to treat a small number of patients to reach the threshold. Second, the thresholds reward providers for pursuing the least need areas *within* their own practice, since a provider with a small number of patients in one disease category, but a large number of another, will automatically be incentivized to focus on reaching the threshold for the few patients in the first disease category.

Consider a family doctor with 100 patients served so far out of 200 with need in one domain and 2 out of 3 patients served in another. They will achieve the coverage threshold for the second domain by treating just one additional patient. They may not achieve the threshold for the first domain by treating even 50 additional patients. This dynamic will hold for every provider, and providers who have very small numbers of eligible patients overall will find it very easy to attain high scores in general.

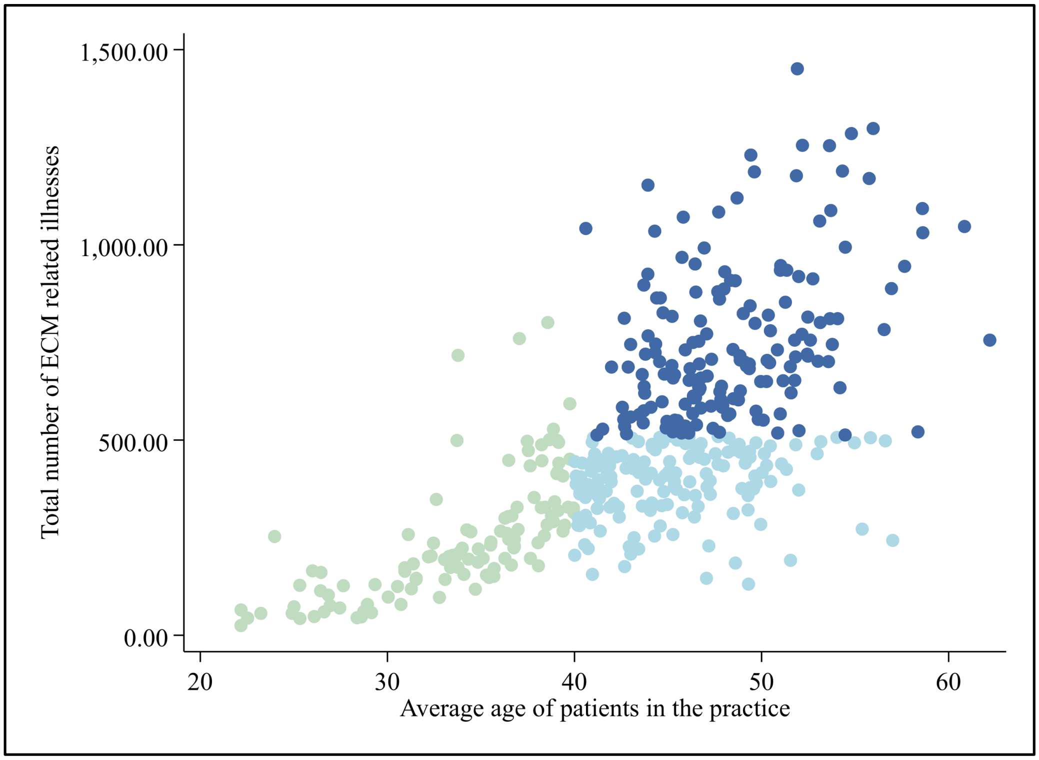
The QBS score as currently designed therefore does not efficiently incentivize a family physician who is motivated to "doing the most good for the most people". This document provides a statistical analysis of the underlying performance of providers in relation to QBS. It also provides a concrete proposal of how the current threshold-based system can be replaced by a score that is both *proportional* and *need-adjusted*. This score takes into account two dynamically adjusted parameters to score each provider. First, it accounts for the provider's overall patient population *relative to other providers*. Second, it adjusts for the relative need in different indicators within the provider's *own patient population*.

We show how these measures can be constructed using the exact same input data – the QBS indicator table – that EHIF currently uses to calculate the QBS scores. We provide the Excel sheet that makes these calculations as a supplement to this document. The measures that we use are somewhat technical, but we provide both with the intuition of correcting the imbalance and inequities between the family physicians described above, who – although equally hard-working and motivated by helping their patients – are treated very differently by the QBS.

The report proceeds as follows. **Part A** summarizes the current QBS scoring system, and highlights both EHIF and World Bank recommendations for improvement from past studies of the QBS system. **Part B** addresses the two identified reversed incentives in the design of QBS scores. It first examines patterns of indicator completion across all providers to suggest new scores for each. Then it demonstrates a method of adjusting each provider's score in each indicator so that rewards are maximized for serving the groups with the most need for that patient list. **Part C** discusses connections to ECM and extensions to other health system priorities.

Figures:

Figure 1. Disease burden and average patient age across practices



PART A: Analysis of current QBS score distribution

Section 1. Current patterns in overall QBS scores

EHIF has expressed their desire to carefully evaluate the calculation of QBS scores and the distribution of incentives in light of the issues raised in the previous section. This section provides an analysis of the issues that have been identified in past reviews of the QBS system and the enhancements that EHIF has actively pursued in response to their reviews. The EHIF team recognizes the need to take into account the specificities of every family physician list and is actively seeking ways to address that issue. In doing so, they hope to reduce the number of costly appeals GPs make on their QBS scores.

EHIF wants to determine the required coverage based on the specificities of the family physician list, including the size of the list and the number of persons targeted by the indicator. They consider that a change in the logic of the determination of coverage would also help in reducing the number of appeals. The World Bank's previous analysis has suggested a revision of QBS indicators, dropping those which have high achievement or low disease burden. They also suggested that QBS should reward both improvement and absolute level target achievement (World Bank, 2018).

QBS scores are calculated based on indicators in two domains. A total of 480 points are available in Domain II, and at least 80% of the 640 available between Domain I and Domain II are needed for a provider to receive QBS financial incentives (World Bank, 2018). This makes Domain II the most "valuable" domain, where the provider's effort makes the greatest difference to their achievement of the QBS payment. For reference, **Table 1** presents the indicators covered in Domain II, along with the score currently assigned to each of the indicators.

Providers in theory could achieve a wide range of scores, but we find that most providers either score very high or very low overall and that scores are consistent over time. There is almost no provider who receives a middling score. This implies that the current incentive thresholds are either easily achieved for a given provider or are essentially impossible. In such a system it is unlikely that these incentives are having strong impacts on provider effort.

Figure 2 presents the distributions of QBS scores across 2017 and 2018. **The left panel shows provider scores in these two years**, with an individual GP's 2017 score on the horizontal axis and their 2018 score on the vertical axis. Dots *on* the illustrated 45-degree line stayed the same in both years, dots *above* the line *improved* (green dots), and dots *below* (red dots) *declined*. **For a majority of the providers, their scores have stayed consistently low or high in both years.**

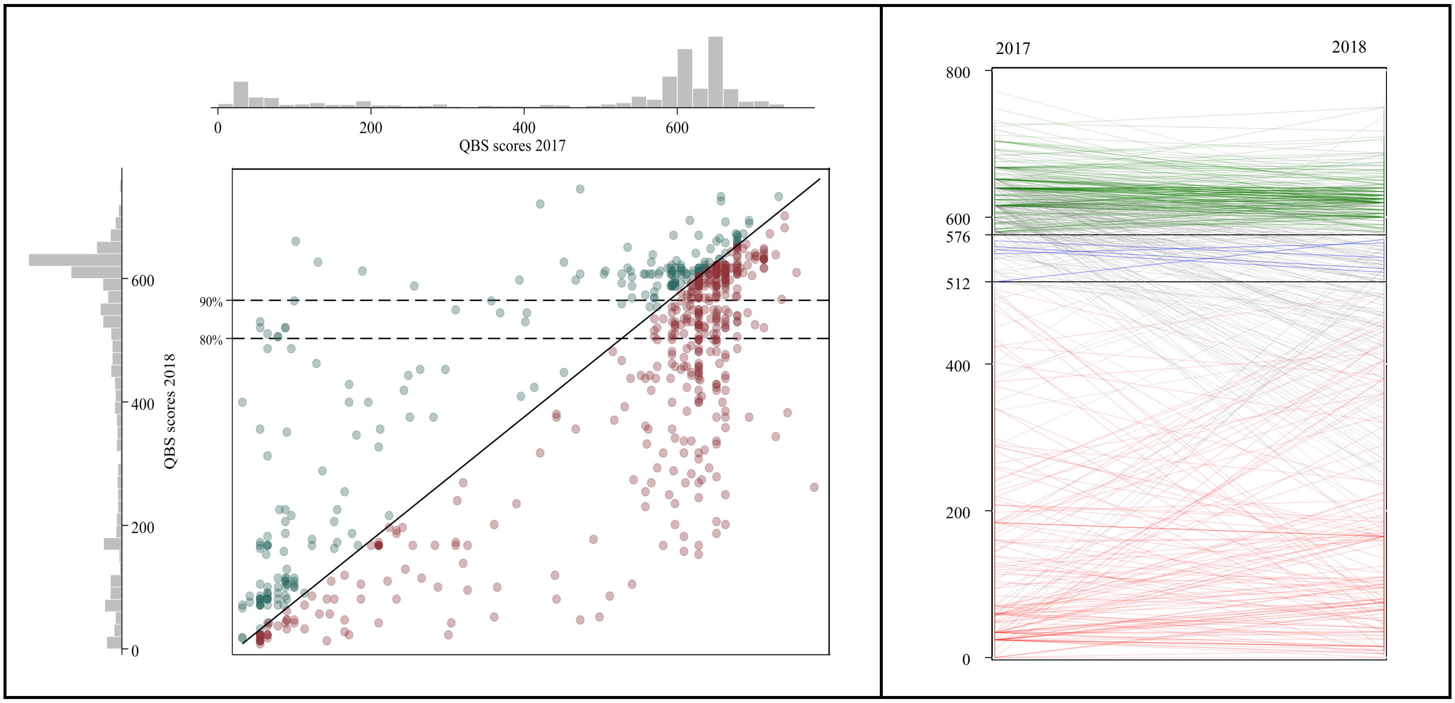
The right panel highlights improvement over time, with each line representing one provider. It shows that very few providers changed their measured performance very much from year to year. While there are a few providers whose scores have changed significantly, a large fraction of providers did not. Furthermore, for those who did change performance, they mostly jumped between extremely low and extremely high. The change is rarely gradual, suggesting the role of thresholds contributing to jumps in provider scores.

Figures and Tables:

Table 1. Current QBS design

Category	Indicator name	QBS current value	Required coverage threshold
Diabetes type II	Monitoring	65	76%
	Medication	10	70%
Hypertension	Monitoring, low risk	90	76%
	Monitoring, medium risk	175	70%
	Monitoring, high/very high Risk	40	73%
	Treatment, all risk levels	5	90%
	Treatment, medium/high Risk	20	83%
Myocardial infarction	Monitoring	20	90%
	Treatment, beta blockers	5	70%
	Treatment, statines	5	70%
Hypothyroidism	Monitoring	45	90%
	Sum of weights	**480**	

Figure 2. Distribution and trend of QBS scores in 2017 and 2018



Section 2: Assessment of the Domain II indicators

This section analyzes the determination of the QBS score at the indicator level. The overall score is simply the sum total of the indicator scores, and the indicator scores are constructed from raw observations in the billing data. We believe this is a good approach and we do not believe it is necessary to create anything more complex. Therefore, we restrict ourselves to working only from the target and coverage measures that EHIF already uses to analyze the current system and suggest alternatives.

Table 2 presented the Domain II indicators with the score assigned to each of them. We start by calculating the indicator-level coverage rates for all providers and analyzing the current point values (which can be understood as importance weights) for them based on the demonstrated ability of providers to achieve high coverage. This approach uses the method of Principal Components to calculate purely data-driven point values for each indicator.

The Principal Components method has two key features. First, it accounts for co-movement between the indicators. The reason why this is useful is because it is undesirable that a single action in effect gets multiple rewards. For example, if monitoring and treatment of the same condition may be very easy to combine, but both are QBS indicators, then each indicator should have relatively less weight to avoid double-counting effort. Second, the method places higher weight on indicators which have a wide distribution relative to the *average*, since this is taken as evidence that providers can, *through effort*, improve their performance. By contrast, when the average is low or there is less room for improvement relative to it, the method assigns a lower weight to the indicator, on the logic that there seems to be little providers can do to improve their outcomes.

Figure 3 shows the principal components analysis and weighting for each of the Domain II indicators. Note that the indicators are essentially split into two groups. The first is those awarded 22 points or less in the QBS system, which all have lower *average* performance in the general population. Furthermore, the low-scoring conditions are the treatment indicators and the high-scoring ones are the monitoring ones, with the exception of the diabetes indicators. The second group is those awarded 63 points or more, which all have higher *average* performance as well as a wide “tail” of providers who are below the average: these are areas where there are lots of real gains that might be achieved by additional provider effort. These valuations are compared with the scores under the current QBS system in **Figure 4**.

The analysis indicates that EHIF has done a good job in creating the score point values, as the chosen levels are remarkably close to the statistical optimum (as calculated via Principal Component Analysis). The current assignment of scores undoubtedly reflects additional information about the importance of these services in the Estonian clinical context. There may be some scope to reduce the weighting on the two hypertension indicators, but there is no general reason these cannot be further adjusted by EHIF to reflect their relative *importance* from a health perspective, as this is a purely statistical, not clinical, analysis. However, our analysis indicates that this is not a significant issue with the current QBS scheme.

Figures and Tables:

Table 2: Current weight and principal component weight for domain II indicators

Indicator Name	**PC weight**	**Current weight**
Diabetes Type II: Monitoring	68	65
Diabetes Type II: Treatment	12	10
Hypertension Low Risk: Monitoring	66	90
Hypertension Medium Risk: Monitoring	68	175
Hypertension High Risk: Monitoring	66	40
Hypertension All Risk: Treatment	17	5
Hypertension Med - High Risk: Treatment	22	20
Myocardial Infarction: Monitoring	68	20
Myocardial Infarction Beta Blockers: Treatment	14	5
Myocardial Infarction Statines: Treatment	19	5
Hypothyroidism: Monitoring	63	45
Sum of Weights	**480**	**480**

Figure 3: Indicator coverage and PC weight

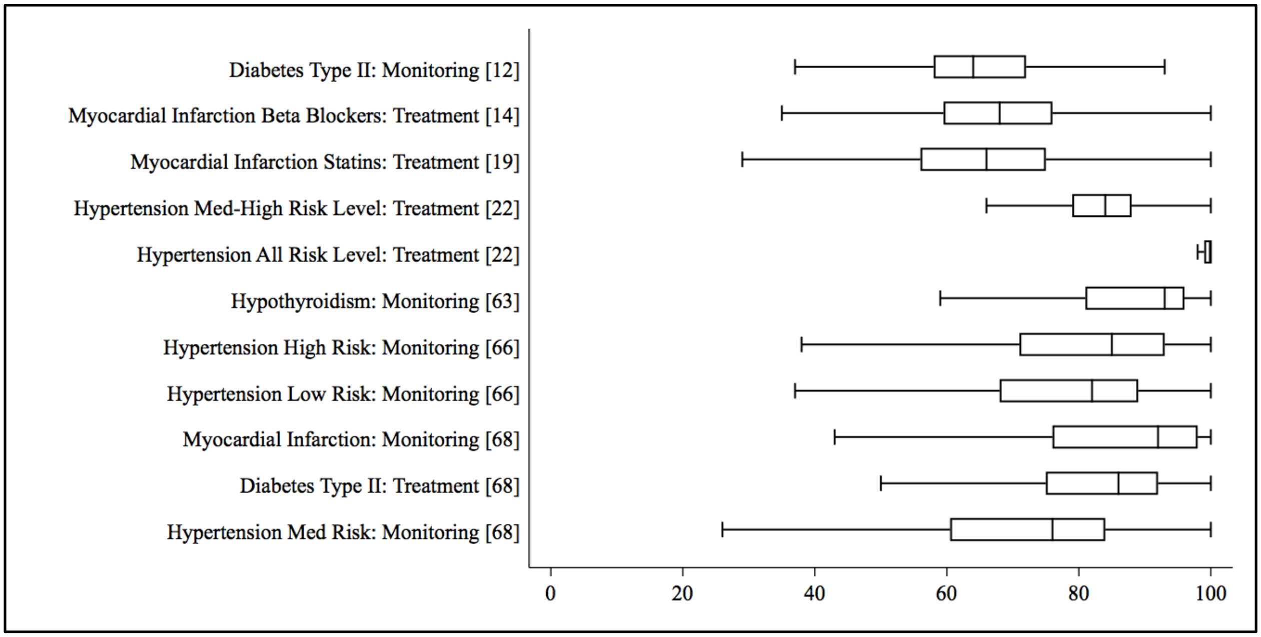
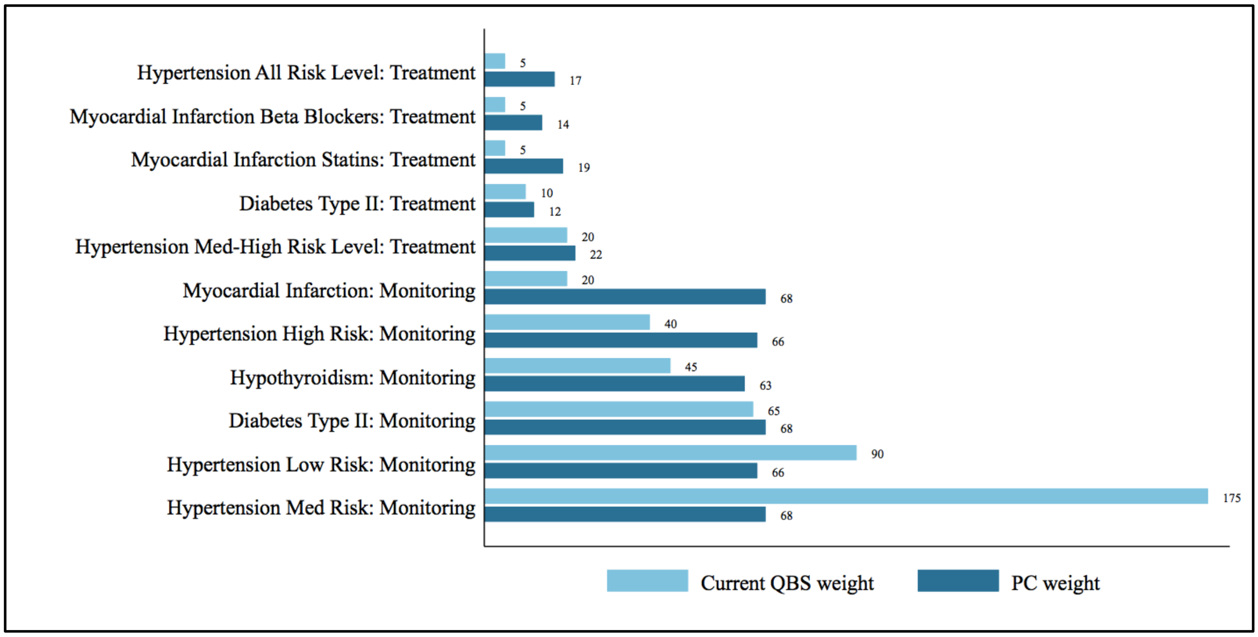


Figure 4: Current weight and PC weight for each indicator



PART B: Improving the QBS system

Section 1: Thresholds without adjustments create bad incentives

This section proposes a "need-adjusted" scoring system that avoids the issues that we have noted above. Specifically, the adjusted system dynamically accounts for the provider's eligible population for each indicator as well as the observed difficulty of service delivery for specific indicators across the provider population. We then demonstrate (in Section 3) how this approach would reward real improvements in service delivery by examining how the distribution of QBS bonuses would change as we simulate behavior changes across the provider population.

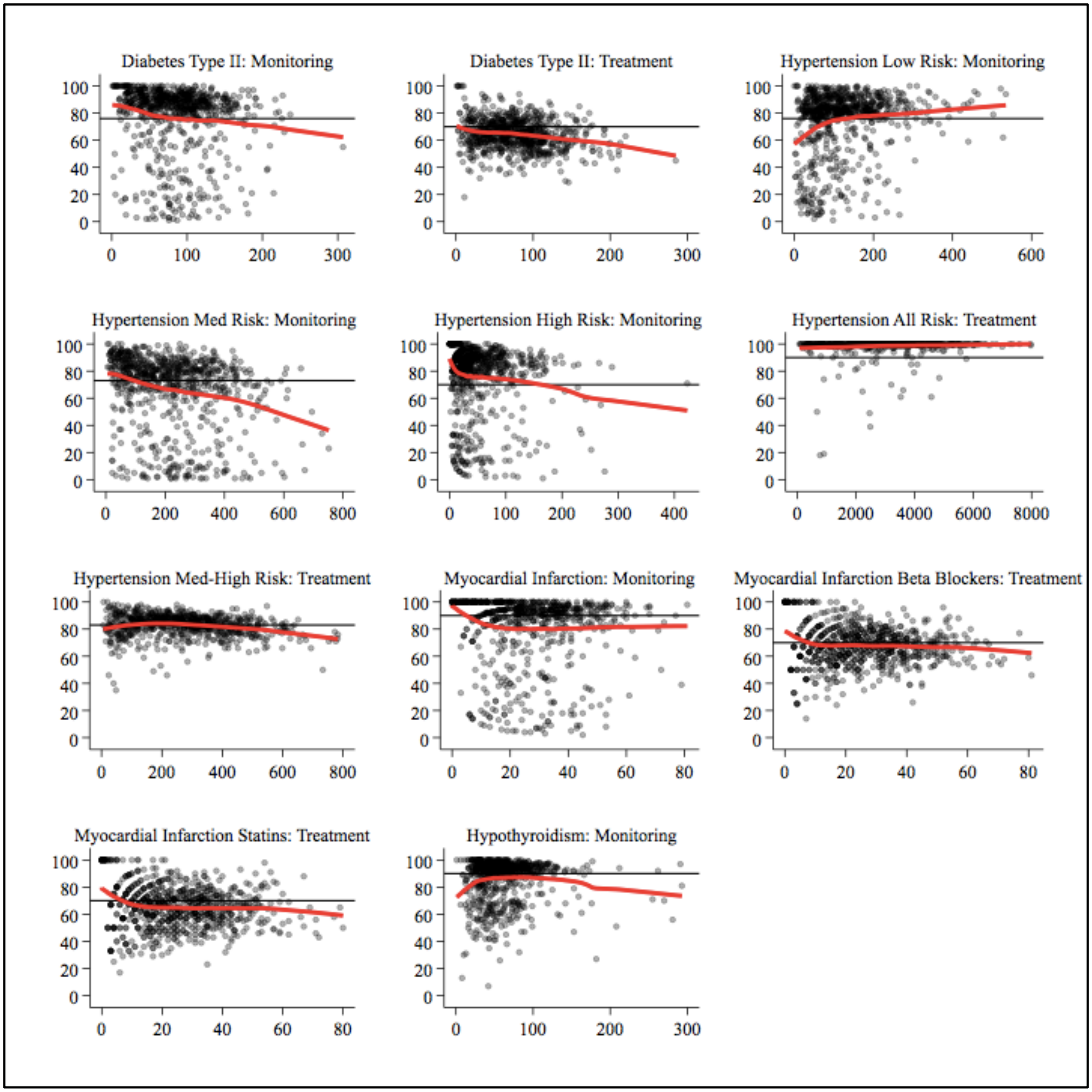
Figure 5 highlights the basic tension in the construction of QBS scores using coverage of indicators. For almost all the Domain II indicators, this analysis shows that *as the number of eligible patients increases, coverage decreases*. This is well-known by EHIF, who have written in a previous report that they would like to change the logic of determining coverage so that the specificities of each family physician list are considered. **In the current system, providers who are doing the most work (in terms of the number of patients treated) are not receiving recognition for it in their QBS scores.** Our analysis bears out that concern. As providers cannot control the distribution of their patients relative to the weights assigned to the indicators, it may put some providers in a situation where scoring satisfactorily in the dimensions with high point values seems unattainable.

The graph also indicates the coverage thresholds for each indicator. **In the present system the points are awarded on an all-or-nothing basis for each indicator: either the threshold is achieved or it is not.** This means that achieving the threshold means more work for providers who have more patients. A doctor with 100 patients with diabetes would have to provide service for 70 of them to achieve 70% coverage, and otherwise they would get 0 points for the indicator. However, another doctor with 10 diabetic patients in their eligible population would only have to provide service for 7 of them to achieve complete coverage for the same indicator. The graph shows clearly that as the number of patients increases, the increasing need causes clinics to fall short of the coverage threshold, even though they are serving more patients on a numerical basis.

Combined, these two features create a "reverse incentive" scheme for these providers that intuitively feels at odds with the principle of rewarding effort and quality care. First, providers are incentivized to prioritize their efforts towards achieving coverage cutoffs in the areas where they have the fewest patients. Second, providers are incentivized to invest in areas with the small gap between their current performance and the scoring cutoff – rather than invest in the areas where there is the largest gap and therefore the *most need* remaining among their patient base, as these investments are not likely to pay rewards under the current scoring schema. Consider a doctor who has with 80 diabetic patients and 10 who have had myocardial infarction and require monitoring. If the coverage threshold is the same for both, providing the required services to each diabetic patient is worth 1/8 the value of the services for each MI patient. In this type of situation, the provider is always rewarded for focusing effort on the category with the least need remaining.

Figures:

Figure 5. Patient eligibility and indicator coverage



Section 2: The "need-adjusted" proposal for QBS improvement

We next tackle the problem of uneven disease distribution across providers. For this we borrow a technique from the human resources assessment literature known as Empirical Bayesian Estimation (similar to that described in Guarino et al. 2015 for the analysis of teacher performance). This method quantifies the amount of information provided in the data and adjusts scores appropriately. Each provider is assumed to be an average performer in each indicator, until they have enough opportunities (patients) to demonstrate otherwise, either positively or negatively away from the average.

Providers who have very few chances to treat a given condition are neither penalized nor rewarded. This is in practice achieved by adjusting both the numerator and denominator of the coverage ratio for each provider by a fixed amount, that, when the number of patients is 0, produces an exactly average coverage ratio for the indicator. When the provider's denominator is large, the fixed adjustment will be very small relative to the real number of patients and their earned score will approach their true (unadjusted) performance. **Table 3** below demonstrates how this process affects scores for some hypothetical providers under the current and need-adjusted systems.

Tables:

Table 3. Simulation of need based coverage

INDICATOR - DIABETES TYPE II MONITORING								
QBS raw data 2018			**Unadjusted coverage**			**Need based coverage**		
Provider	**Target population**	**Actual work done**	**Unadjusted coverage**	**Current weight**	**Current QBS score**	**Adjusted coverage**	**PC weight**	**PC adjusted score**
A	2	2	100%	65	**65**	61%	68	**41**
B	98	30	31%	65	**0**	40%	68	**27**
C	153	100	65%	65	**0**	64%	68	**43**
D	122	90	74%	65	**0**	70%	68	**47**

Section 3: Assessing the new proposal for QBS

Based on a combination of Principal Components for reweighting the indicators and Bayesian re-estimation for adjusting indicator coverage, we can now examine the complete "need-adjusted" QBS scores. Here, we use the real performance data from QBS 2018 to compare and contrast the performance of providers under the current systems and our proposed scheme.

Figure 6 shows the overall distribution of the new scores. **Unlike the current scores, which demonstrated two large clusters and a "missing middle", the need-adjusted scores have a normal distribution around the average, and a long tail of underperformers.** Note that these underperformers are *not* the same ones who scored near zero in the current scheme. To show this we focus on the lowest performing providers, who score under 10 points in the current system. They have an average of about 3,600 eligible patients per provider, and an average coverage rate of 40%. Though they are covering more than a third of their population on average they did not meet the thresholds of any indicators, because of their large patient population and high disease burdens. **Figure 7** shows **the adjusted scores of such providers who scored under 10 with the application of a need-adjusted approach**. We now see that providers who are achieving high coverage over their population in need are getting substantial recognition for it. **Their new scores range from 160 points to 360 based on their adjusted coverage and partial credit, and the need-adjusted score shows a strong relationship between services delivered and scores**.

In **Figure 8**, we compare the two scores directly, highlighting the role of "met and unmet need" in the two approaches. The current QBS score appears on the left panel; in the middle is the score purely proportional to coverage rates with the current weight system and; the need-adjusted score appears on the right. The vertical axis is the achieved score for Domain II, and the horizontal axis is the total number of patients targeted across all indicators. For reference, we add cutoff lines and color banding for 90% and 80% score (out of 480) overall, in line with the current QBS approach to financial incentives (although these may need to be revised, we do not take a stance on that).

There are several notable patterns. First, the adjusted scores are lower. This is intentional: no provider is currently meeting *all* the recorded needs in their population, so each one has room to improve in the future. By contrast, in the current system, we see that providers are clustered near the maximum possible performance, meaning there is little recognition for improved performance. In some cases, because of the "coefficient" adjustment, providers are allowed to exceed the theoretical maximum. This stands out very clearly in the upper region which is "unattainable" in the adjusted system – and many of those same providers remain those performing at the top of the adjusted distribution.

To illustrate the altered incentive system, **Table 4** uses four hypothetical providers to illustrate how providers might maximize their score *improvement* in the next year's score. In particular, we highlight that, because of the Bayesian approach, each provider gets the most additional points by investing in the area where they have the most patients *and* the least coverage. To bring the two parts together, this requires accounting for the Principal Components point value of the indicator; which shows where *other* providers have been able to achieve high coverage.

Such a system rewards providers that make significant improvements from one year to the next. We note that it is not possible to compare these scores to current ones for the purposes of distribution of financial incentives as the alteration of the scores will move a large number of providers in and out of the current incentive cutoffs in unexpected ways and the scores are not intended to be comparable on the same scale. This should be carefully considered by EHIF in adopting any of the proposed scoring changes.

At the same time, this system is built on a principle of fairness across providers. Embedded in our approach is the idea that scoring should automatically take into account both the unique situation faced by the individual provider, but also hold them to the standards set by the others around the country. Performance on a given indicator can and should always be compared over time using the principle of *adjusted coverage*. Similarly, providers should be compared to each other and over time based on their *relative rank*, not their *total score*. We do not necessarily recommend directly comparing the points awarded over time, because the weighting of the indicators and the total score outcome in a given year is a highly flexible policy tool for EHIF to bring to bear on emerging priorities.

Figures and Tables:

Figure 6. Distribution of Need Adjusted Scores

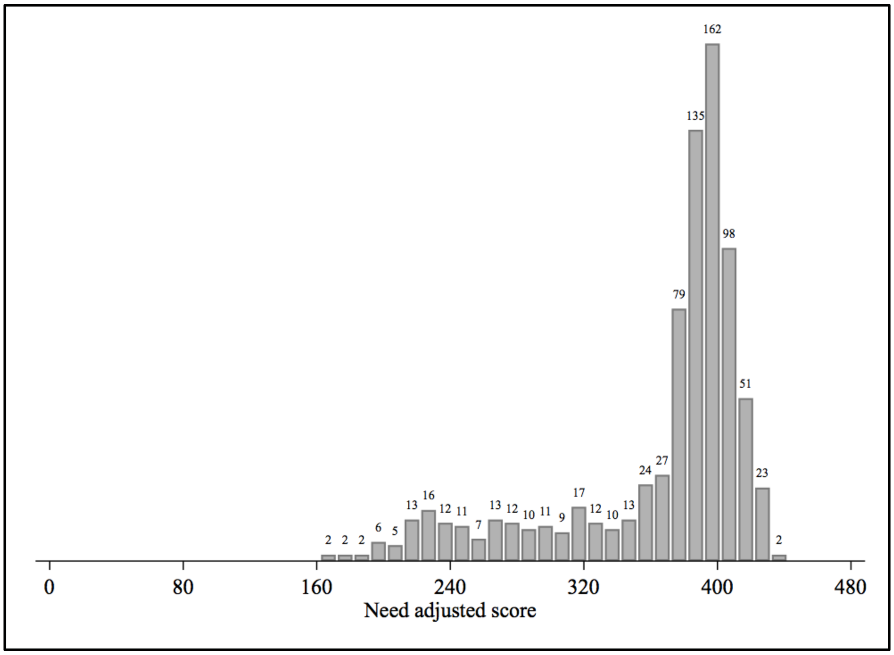


Figure 7. Need-adjusted scores for providers with current QBS scores below 10

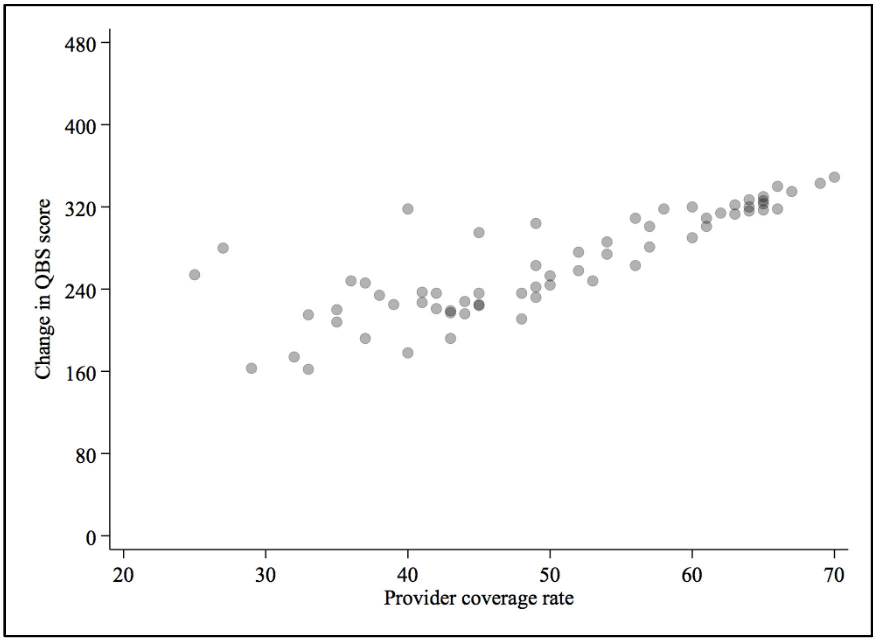


Figure 8. QBS Scores and needs of population

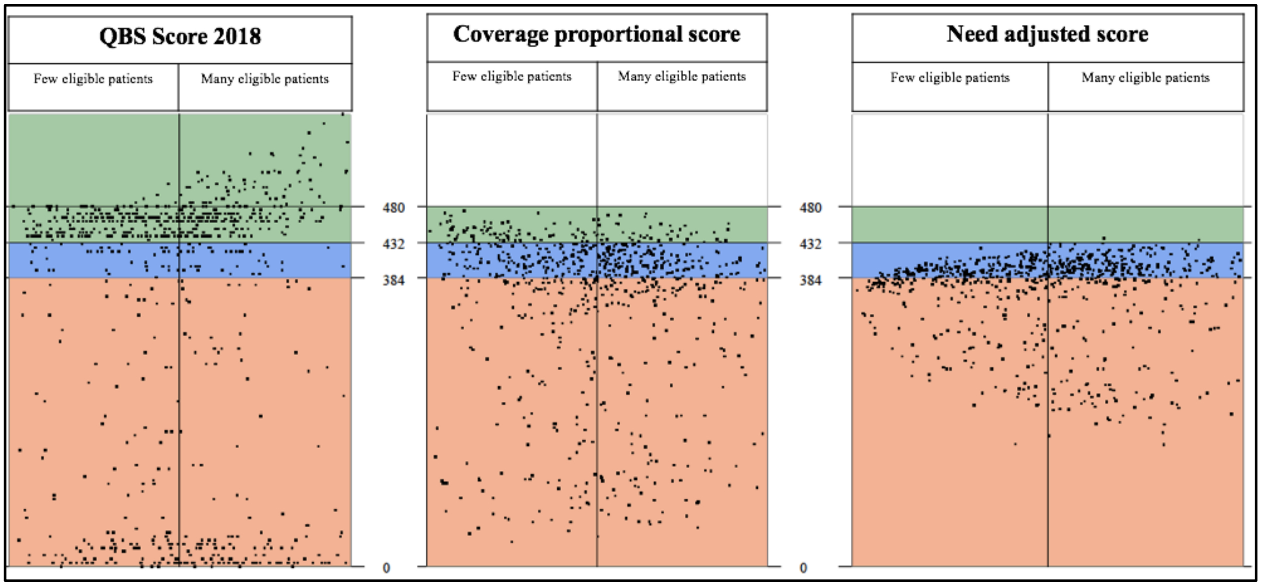


Table 4. Improvement of scores under "need-based coverage"

INDICATOR - DIABETES TYPE II MONITORING								
Raw 2018 QBS data			Unadjusted coverage			Need based coverage		
Provider	**Target population**	**Actual work done**	**Unadjusted coverage**	**QBS score weight**	**Current QBS score**	**Adjusted coverage**	**PC weight**	**PC adjusted score**
A	2	2	100%	65	**65**	61%	68	**41**
B	98	30	31%	65	**0**	40%	68	**27**
C	153	100	65%	65	**0**	64%	68	**43**
D	122	90	74%	65	**0**	70%	68	**47**

INDICATOR - DIABETES TYPE II MONITORING								
Hypothetical 2019 QBS data			Unadjusted coverage			Need based coverage		
Provider	**Target population**	**Actual work done**	**Unadjusted coverage**	**QBS score weight**	**Current QBS score**	**Adjusted coverage**	**PC weight**	**PC adjusted score**
A	2	2	100%	65	**65**	70%	68	**47**
B	98	**40**	41%	65	**0**	50%	68	**34**
C	153	**115**	75%	65	**0**	74%	68	**50**
D	122	**100**	82%	65	**65**	78%	68	**53**

PART C: Conclusion and next steps

Section 1: Relationship to future ECM implementation

At the moment ECM is a lump-sum payment per patient. This ensures that there is a baseline resource for all the needs of chronic patients. However, it does not provide an incentive at the margin for providers to ensure patients receive the treatments they may require. **Whatever the rationale for making QBS an incentive-based treatment model, implicitly that reasoning has not been applied to ECM or it has and there is a good reason that ECM is a fixed payment.** This analysis emphasizes that EHIF must be careful when assigning either financial or merit value to *additional* services provided. Threshold systems or complex conditionality can, as demonstrated, combine to create unexpected incentives that do not fairly reward additional effort. This may be particularly poor at reflecting provider improvement over time when the patient population is fixed.

However, ECM must certainly include accommodation or design elements for the widely-varying distribution of eligible patients across different providers. If the QBS analysis is any indication, there will be a segment of providers for whom ECM will be a relatively light burden given their underlying populations, and some for whom it is a very difficult task. **Analyzing these needs in advance and appropriately working to accommodate providers for the work that is being asked of them,** and supporting and managing participation in the program carefully, will be of paramount importance to ensure that the system is received as fair and equitable for all.

There may be an argument for shifting ECM payments to a treatment or performance-basis, if it requires significant incentives to have providers take up treatments that may be costly to them in terms of time and effort. This should of course be considered if it would make ECM a more effective system. If such an approach were taken, it would be important to ensure that any performance-basis for ECM is designed as complementary to the incentive scheme outlined in this document. Similarly, it would be important to determine to what extent any scoring or reward system is designed as a merit or recognition system, and to what extent it is intended to be a direct financial incentive or compensation for particular practices.

Section 2: Extensions to COVID and other health system priorities

The proposed system as laid out is simple to extend to other priority areas through the addition or reweighting of indicators. Since we additionally do not suggest any particular allocation of the financial resources, EHIF can exercise a great deal of judgment and flexibility in doing so, and use that system to enforce additional priorities, so long as the conditions do not become overly complex and re-introduce the original issues of fairness, threshold effects, and conditional unattainability. But there is no reason that careful combinations of requirements cannot be imposed. For example, one incentive condition that combines this work with a new policy priority might be: "attain 70% of the overall Domain I and II scores *and* at least 90% in the COVID indicator". Depending on the chosen COVID suppression strategy of the Ministry, such an indicator could, for example:

- Require specific individuals to be tested as determined by a central plan
- Require individuals with certain characteristics or pre-existing conditions be tested
- Require a certain proportion of individuals overall to be tested
- Require reporting and contact tracing to be completed for positive testing individuals

These approaches are easy to extend to other emerging or temporary priorities that may emerge for EHIF in the future. We reiterate that this requires that absolute scores not necessarily be comparable from year to year. Performance should always be compared on an indicator basis using the adjusted coverage, and providers should be compared over time using their relative rank. This will of course require active effort from EHIF to add and remove measures without adding too much complexity to the system or seeming to overburden providers. However, the overarching structure of this report has focused on the need to balance simplicity, fairness, and information; and these will always remain the key challenges of any performance incentive system.

References

Estonia Health Insurance Fund. 2020. *Taustinfo ja probleemkohad perearstide kvaliteedisüsteemi kohta*

Merilind, E. 2016. *Primary health care performance: impact of payment and practice-based characteristics.*

Guarino, C. M., Reckase, M. D., & Wooldridge, J. M. 2015. *Can value-added measures of teacher performance be trusted?* Education Finance and Policy, 10(1), 117-156.

World Bank Group. 2018. *Revising Estonia's Quality Bonus Scheme in Primary Care*