MILLIMAN REPORT

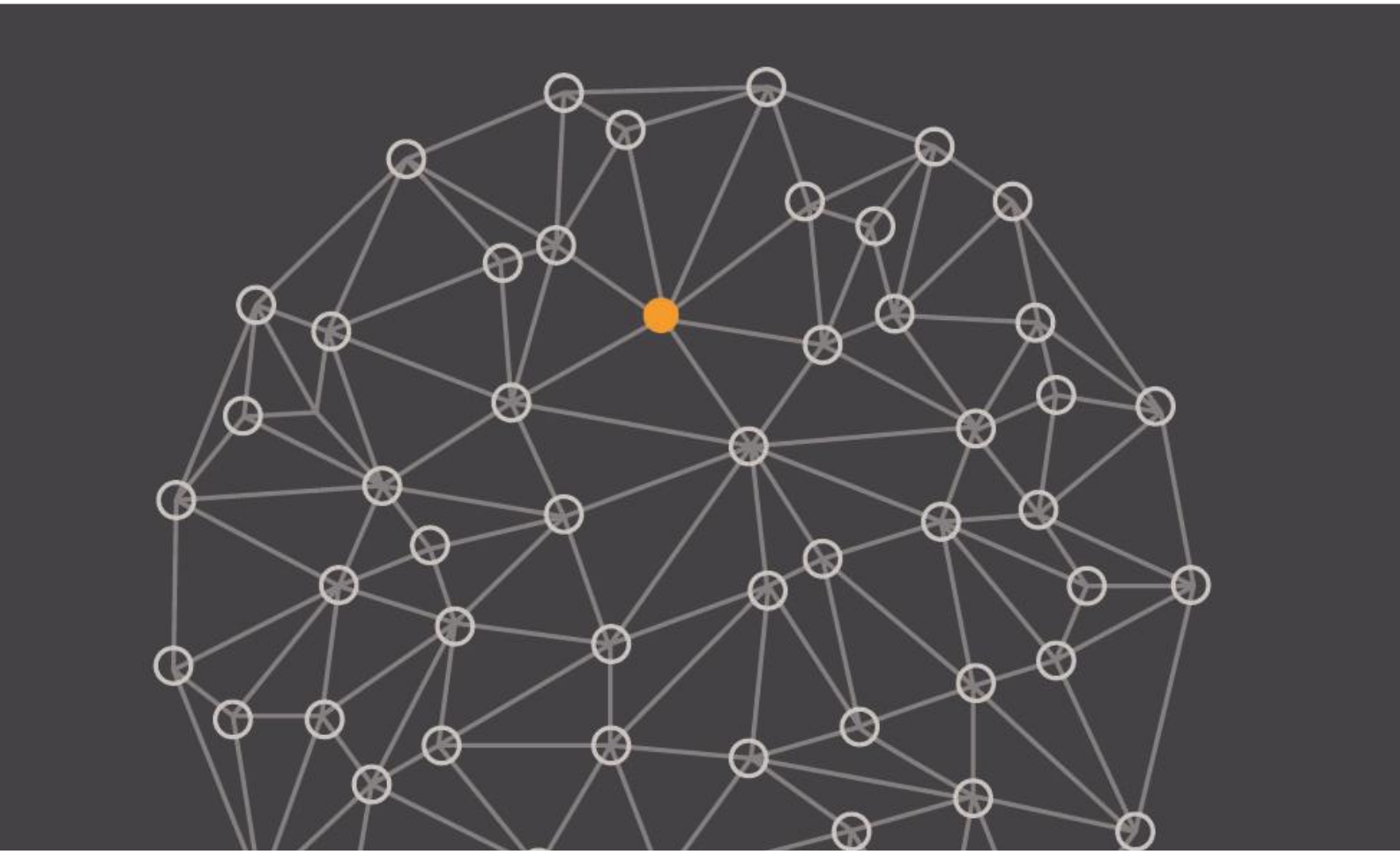
Vitality Health

Peer review: Performance metrics for premier specialist consultants relative to non-premier consultants

13 June 2021

Joanne Buckle, FIA
Tanya Hayward, FIA
Natasha Singhal, FIA

Commissioned by Vitality Health



Milliman


Milliman

Table of Contents

1. EXECUTIVE SUMMARY

1.1. Vitality Health ("Vitality") invites certain specialist physician consultants ("consultants") to become "Premier" consultants; other recognised consultants are by definition "non-Premier." Vitality has a methodology that it uses to stratify consultants into its Premier and non-Premier categories. We did not evaluate this methodology. We were asked by Vitality to peer review Vitality's approach to measure the relative performance of each group of consultants and provide our comments on the robustness of the methodology.

1.2. The methodology Vitality uses to measure the relative performance of Premier and non-Premier consultants considers four performance metrics for Premier consultants relative to non-Premier consultants:

- Frequency of change of original consultant within 90 days. This measure is intended to capture whether patients are seeing the right consultant the first time;
- Same-cause readmission rates within 28 days of the original discharge, where same-cause relates to admissions within the same episode type as the initial admission;
- Length of stay for inpatient and daycase admissions; and
- Total cost of episode.

1.3. Our review considered:

1) The appropriateness of the four selected performance metrics as a way to measure the relative performance of Premier and non-Premier consultants.

2) Whether the reported differences in performance between Premier and non-Premier consultants are statistically significant.

1.4. Three out of the four metrics used by Vitality to compare the performance of Premier and non-Premier consultants are commonly used by other organisations and health systems to measure quality or efficiency. These metrics are length of stay[1], readmission rates[2] and total episode cost[3]. The fourth metric, frequency of change of original consultant, is not a commonly used performance metric. However having multiple consultants within the same speciality involved in a single episode[4] could generate additional utilisation and subsequently additional costs. Therefore, this metric could be considered an efficiency metric. Vitality believes that a lower frequency of changing consultants could indicate that patients are more satisfied with their original choice of consultant.

1.5. For the base year used in the analysis (AY[5] 2019/20), aggregated results by speciality indicate that there is a statistically significant difference between Premier and non-Premier consultants for all performance metrics other than the 'total generated spend' metric. For the total generated spend, the result is significant at an 84.5% significance level (95% is typically the minimum significance threshold). This means that the probability of there being no difference between the performance of Premier and non-Premier consultants is at least 15.5% (or approximately 1 in 6, compared to 1 in 20 at a 95% significance threshold). Sensitivity-testing indicated that changing key parameters in the methodology would not result in materially different conclusions.

[1] https://www.england.nhs.uk/urgent-emergency-care/reducing-length-of-stay/
[2] 1] https://www.ncqa.org/hedis/measures/plan-all-cause-readmissions/; 2] 3b Emergency readmissions within 30 days of discharge from hospital - NHS Digital; 3] https://www.cms.gov/Medicare/MedicareFeeforServicePayment/PhysicianFeedbackProgram/Downloads/2015-ACR-MIF.pdf; 4] https://www.milliman.com/-/media/milliman/importedfiles/uploadedfiles/insight/2017/differences-in-details.ashx
[3] Cost of an episode is one of the major focus areas of the Quadruple Aim of Healthcare, which is a widely accepted compass to optimise health system performance. https://digital.ahrq.gov/acts/quadruple-aim
[4] Note that episodes and claims are interchangeable, as per Vitality's definition.
[5] Note analysis year 'AY' runs from October to September

1.6. Figure 1.1 below shows key conclusions resulting from our statistical significance analysis on the base case scenarios, at an aggregate level, for all four of Vitality's performance metrics, i.e. without varying any of the key parameters:

Figure 1.1 - Aggregate base case scenario results

Performance metric	Base Scenario	Conclusion
Frequency of change of consultant	Oct 19 - Sept 20; 90 day cut off for switch of consultant	There is a less than 5% chance that there is no difference in the mean of this performance metric for Premier consultants compared to non-Premier consultants. *p-value = 0.001*
Readmission rates	Oct 19 - Sept 20; 28 day readmission rate window	There is a less than 5% chance that there is no difference in the mean of this performance metric for Premier consultants compared to non-Premier consultants *p-value = 0.022*
Length of stay	Oct 19 - Sept 20	There is a less than 5% chance that there is no difference in the mean of this performance metric for Premier consultants compared to non-Premier consultants *p-value = 0.000*
Total generated episode spend	Oct 19 - Sept 20	There is a 15.5% chance that there is no difference in the mean of this performance metric for Premier consultants compared to non-Premier consultants. *p-value = 0.155*

1.7. We have reviewed Vitality Health's approach to measuring the performance of Premier consultants relative to non-Premier consultants, and concluded that Vitality Health's methodology is reasonable and likely to generate credible results at the aggregate level.

This conclusion has been reached with the following caveats:

- We did not audit the performance metric calculations and did not conduct an independent analysis to evaluate the methodology Vitality Health used to identify its' Premier consultants.
- We do not guarantee that any specific metrics are optimal for measuring the relative performance of Premier vs non-Premier consultants, but acknowledges that length of stay, readmission rates and total episode cost are commonly used quality and efficiency metrics.
- We were engaged by Vitality to perform this review of its own analysis, and we relied on Vitality's data and presented results.

1.8. We note that statistical significance at the 95% confidence level was not demonstrated at the speciality level for any of the metrics apart from Length of Stay. This is very likely due to both small sample sizes at the speciality level and variability within episodes of care. The statistical tests have been conducted at an aggregate level, with additional scenario testing conducted at the specialty level.

1.9. We understand that a case mix adjustment is applied to all performance metrics. Our review has not validated the development or application of the case mix adjustment. We have run our statistical tests on the raw data, before case mix adjustment and note that the conclusions of our analysis remain largely consistent with and without the case mix adjustment; however at the speciality level there are some isolated differences.

1.10. Any reader of this Report must possess a certain level of expertise in areas relevant to this analysis to appreciate the significance of the assumptions and the impact of these assumptions on the illustrated results. Milliman recommends that third parties be aided by their own actuary or other qualified professional when reviewing this Report.

1.11. We understand that Vitality intends to provide public access to this Report through an internet link and therefore could be viewed by its prospective customers, competitors, potential investors, or other interested parties. We consent to this distribution if the work is distributed in its entirety. Milliman does not intend to benefit third parties from this work and assumes no duty or liability to other parties who review this work.

2. BACKGROUND, PURPOSE AND SCOPE

2.1. Vitality Health (“Vitality”) invites certain specialist physician consultants ("consultants") to become "Premier" consultants; other recognised consultants are by definition "non-Premier." Vitality has a methodology that it uses to stratify consultants into its Premier and non-Premier categories. We did not evaluate this methodology. We were asked by Vitality to peer review Vitality’s approach to measure the relative outcomes of each group of consultants and provide our comments on the robustness of this methodology.

2.2. It was not part of our scope to audit the performance metrics calculations and did not conduct an independent analysis to evaluate the methodology Vitality used to identify its’ Premier consultants.

2.3. This report is intended to provide documentation of the work we undertook to peer review the outcomes methodology to support the statement that Vitality intends to publish on its website referencing our peer review. It should not be used for any other purpose.

3. CAVEATS, LIMITATIONS, DISTRIBUTION AND USE

3.1. In performing our peer review, we relied on data and other information provided by Vitality. We have not audited or verified this data and other information. If the underlying data or information is inaccurate or incomplete, the results of our analysis may likewise be inaccurate or incomplete. We performed a limited review of the data used directly in our analysis for reasonableness and consistency and have not found material defects in the data. If there are material defects in the data, it is possible that they would be uncovered by a detailed, systematic review of the data to search for data values that are questionable or for relationships that are materially inconsistent. Such a review was beyond the scope of our assignment.

3.2. No reliance should be placed on any advice not given in writing, or on draft versions of our figures, reports or other forms of written communication. Milliman does not intend to benefit any third-party recipient of our work product.

3.3. We are only commenting on the generalised outcomes methodologies that were provided to us and evaluating the actuarial appropriateness of those methodologies overall. We are not commenting on results for any particular specialist consultant or on the outcomes that may be achieved by any specific corporate or individual client of Vitality's.

3.4. While we find the outcomes methodology appropriate, all methodologies, algorithms and formulas are by nature assumption-driven. While we have attempted to test the sensitivity of the results and conclusions to key assumptions to assess the robustness of the methodology, we are not providing an opinion on the appropriateness of any specific assumption.

3.5. This review incorporates Milliman's experience in working with similar programs that rely on claims data to determine outcomes. Future experience will differ from the outcomes we reviewed for many reasons, including, but not limited to: patient characteristics, changes to the underlying stratification methodology, benefit designs that influence utilisation, and consultant practice patterns, as well as other random and non-random factors. It is important that actual experience be monitored and that appropriate adjustments be made to the methodologies on a regular basis to ensure they remain appropriate. It is certain that actual experience will vary from expected, perhaps materially.

4. PROFESSIONAL AND TECHNICAL WORK STANDARDS

4.1. The three authors of this analysis are Fellows of the Institute and Faculty of Actuaries in the UK and therefore the work carried out and this Report fall within scope of the following professional guidance:

- Generic TAS 100: Principles for Technical Actuarial Work ("TAS 100"), as approved by the Financial Reporting Council ("FRC") with effect from July 2017.

4.2. We confirm that in undertaking this work and in preparing the final version of the Report we have complied with the above guidance, subject where appropriate to our judgements regarding materiality and proportionality.

5. PEER REVIEW METHODOLOGY

5.1. Our review considered:

1) The appropriateness of the four selected performance metrics as a way to measure the relative performance of Premier and non-Premier consultants.

2) Whether the reported differences in performance between Premier and non-Premier consultants are statistically significant.

5.2. Our review consisted of evaluating the Vitality methodology in two phases:

5.2.1. **A conceptual phase** to understand the theoretical basis on which the outcomes metrics of Premier and non-Premier consultants had been compared; and

5.2.2. **A testing phase** to run statistical tests and sensitivity tests to understand:

- Whether the reported differences in outcomes between Premier and non-Premier consultants were statistically significant; and
- To determine whether varying some specific parameters in the calculation methodology may affect the results materially and therefore change the conclusions.

5.3. In the conceptual phase, we carried out a desk review of the provided methodology statements and associated Excel output. Vitality discussed with us the end-to-end process and we provided Vitality with a checklist of questions and items for documentation, including, but not limited to: a) data quality limitations; b) reconciliations of data; c) time periods used for the analysis, d) data cleansing or truncation of data to remove outliers; e) methodological approach to risk-adjustment; f) implicit and explicit assumptions used in the consultant stratification methodology and subsequent performance metrics; g) case-mix adjustment; h) calculation of episode groups from individual claims; i) treatment of new consultants; and j) treatment of differences in policy terms and conditions or benefits.

5.4. Following discussions with Vitality, we supplied a testing plan designed to assess the statistical significance of the results of the performance metrics, and to test the robustness of the methodology to changes in critical assumptions. We supplied a summary data template to Vitality for them to populate.

5.5. We carried out a number of statistical tests on the performance metrics, using a hypothesis-testing framework. The following shows an example of this hypothesis testing for readmission rates:

Figure 5.1 - Example of hypothesis testing framework (readmission rates)

Study question	Are readmission rates lower for Premier consultants compared to non-Premier consultants?
Hypotheses:	
Ho (Null hypothesis)	There is no difference between readmission rates for Premier and non-Premier consultants *i.e. ($\mu_p = \mu_q$) where p = premier and q = non-premier*
Ha (Alternative hypothesis)	Readmission rates for Premier consultants are lower than for non-Premier consultants *i.e. ($\mu_p < \mu_q$ or $\mu_p - \mu_q < 0$)*

5.6. For the readmission rates and frequency of change of consultant tests, we performed one-tailed hypothesis tests for two proportions assuming a normal distribution.

5.7. For length of stay and total generated spend, we performed one-tailed hypothesis tests for the difference of two means assuming a normal distribution.

5.8. We carried out the hypothesis testing at both the speciality level and the aggregate level (i.e., aggregated over all specialties) separately for each performance metric.

5.9. The statistical testing yields a “p” value. This value determines the confidence we can place on the results – the lower the “p” value, the more confident we can be that the difference between Premier and non-Premier performance metrics does not simply arise by random chance. The p-value can be interpreted as the probability that there is no difference between the performance metrics considered for Premier and non-Premier consultants. The lower the p-value, the higher the probability that there is a statistically significant

difference between the performance metrics of Premier consultants compared to non-Premier consultants. A 95% significance level has been used when interpreting the results of the statistical tests.

5.10. We also tested the sensitivity of the results to key parameters and these are shown in Appendix A. These additional tests include:

5.10.1. **Choice of year of data.** The performance metrics were originally run by Vitality on Oct 2019 to Sept 2020 incurred claims data, with a run-out period for paid claims of three months. We also requested that Vitality provide the same metrics using data for two earlier complete years.

5.10.2. **Cut off days for the Less Frequent Change of Consultant metric.** The performance metric specifies that the cut-off time to measure whether or not a patient has changed consultant is 90 days. We also requested results based on 30 days, 60 days and 120 days.

5.10.3. **Window for readmissions metric.** The performance metric specifies that same-cause readmissions are measured within 28 days of original discharge. We also requested results assuming that a 14, 60 and 90 day cut off is used.

5.10.4. **Beddays.** We were provided results for length of stay, and using the data shared with us we performed our statistical analysis on an additional metric of beddays. Beddays considers the total number of days a patient is hospitalised within an episode rather than the average length of stay across admissions within an episode.

6. RESULTS AND CONCLUSIONS

6.1. Figure 6.1 below shows key conclusions resulting from our statistical significance analysis for all four of Vitality's performance metrics for the base case scenarios, i.e. without varying any of the key parameters, at an aggregate level:

Figure 6.1 - Aggregate base case scenario results

Performance metric	Base Scenario	Conclusion
Frequency of change of consultant	AY Oct 19 - Sept 20; 90 day cut off for switch of consultant	There is a less than 5% chance that there is no difference in the mean of this performance metric for Premier consultants compared to non-Premier consultants *p-value = 0.001*
Readmission rates	AY Oct 19 - Sept 20; 28 day readmission rate window	There is a less than 5% chance that there is no difference in the mean of this performance metric for Premier consultants compared to non-Premier consultants *p-value = 0.022*
Length of stay	AY Oct 19 - Sept 20	There is a less than 5% chance that there is no difference in the mean of this performance metric for Premier consultants compared to non-Premier consultants *p-value = 0.000*
Total generated episode spend	AY Oct 19 - Sept 20	There is a 15.5% chance that there is no difference in the mean of this performance metric for Premier consultants compared to non-Premier consultants *p-value = 0.155*

6.2. The above results indicate that when looking at all specialities combined, there is a statistically significant difference at the 95% level between the performance metrics of Premier and non-Premier consultants for all metrics other than total generated episode spend. For the total generated episode spend, the difference is significant at the 84.5% significance level which is lower than the typical 95% threshold. This means that the probability of there being no difference between the performance of Premier and non-Premier consultants is at least 15.5% (or approximately 1 in 6, compared to 1 in 20 at a 95% significance threshold).

6.3. Three out of the four metrics used by Vitality to compare the performance of Premier and non-Premier consultants are commonly used by other organisations and health systems to measure quality or efficiency. These metrics are length of stay[6], readmission rates[7] and total episode cost[8]. The fourth metric, frequency of change of original consultant, is not a commonly used performance metric. However having multiple consultants within the same speciality involved in a single episode could generate additional utilisation and subsequently additional costs. Therefore, this metric could be considered an efficiency metric. Vitality believes that a lower frequency of changing consultants could indicate that patients are more satisfied with their original choice of consultant.

6.4. The results for each of our tested scenarios yielded results that lead to the same conclusions as the base case for each performance metric, with some isolated differences.

6.5. Vitality has shared the following summarised data tables with us. We have reviewed these to study the distribution of cost and utilisation and investigate potential factors which may impact the quality of our analysis. Overall, we did not identify any major issues with the summaries provided and note any considerations in Section 7 of this Report.

6.5.1. Speciality and Clinical Treatment Group (CTG)[9] level summary of key fields, including claimant count, claims count, total cost per claim, average cost per claimant per year and standard deviation of cost per claimant.

6.5.2. Proportion of admissions and episodes with packaged pricing by speciality and hospital group.

6.5.3. For each performance metric, a review of proportion of cost per claim and proportion of episodes that are the responsibility of the primary consultant.

6.5.4. By speciality, a summary of proportion of episodes with a switch of consultant.

[6] https://www.england.nhs.uk/urgent-emergency-care/reducing-length-of-stay/
[7] 1] https://www.ncqa.org/hedis/measures/plan-all-cause-readmissions/; 2] 3b Emergency readmissions within 30 days of discharge from hospital - NHS Digital; 3] https://www.cms.gov/Medicare/MedicareFeeforServicePayment/PhysicianFeedbackProgram/Downloads/2015-ACR-MIF.pdf; 4] https://www.milliman.com/-/media/milliman/importedfiles/uploadedfiles/insight/2017/differences-in-details.ashx
[8] Cost of an episode is one of the major focus areas of the Quadruple Aim of Healthcare, which is a widely accepted compass to optimise health system performance. https://digital.ahrq.gov/acts/quadruple-aim
[9] The CTGs are the classifications that Vitality has developed to group claims into episodes of care.

6.5.5. By speciality, a summary of the proportion of episodes with more than one admission per claim.

6.5.6. By speciality, a summary of the proportion of consultants based in London in both the Premier and non-Premier groups.

7. CONSIDERATIONS

7.1. Although we determined that Vitality's methodologies measure the relative performance of Premier and non-Premier consultants using the four performance metrics are reasonable and consistent with typical actuarial practices, we identified several possible limitations that should be considered by any user relying on results generated using Vitality's methods. These limitations apply to the base cases and scenarios for each performance metric. They include but are not limited to the following:

7.1.1. There was low statistical significance between the performance metrics for Premier vs non-Premier consultants when measured at the speciality level for all metrics except Length of Stay. This may be simply due to small sample sizes when measured at the speciality level as well as high variability within episodes.

7.1.1.1. For Length of Stay, statistical significance of this performance metric was high for all specialties ($p = < 0.039$) apart from Dermatology and Cardiology, where p values were higher than 0.5. This conclusion remained broadly stable while running our sensitivity tests, other than when statistical significance was tested using Oct-17 to Sept-18 claims experience, the p value was higher than 0.8 for Gastroenterology. Dermatology has a low number of admissions, which results in small sample size issues. There is variability in the Cardiology claims due to the large number of sub-specialities, which leads to heterogeneity between the Premier and non-Premier groups, as well as the necessity to have a wide geographical spread of consultants within the Premier specialist consultant panel. It is not clear the cause of the high p-value in Gastroenterology, however it may be related to the type of claims experienced in that particular year.

7.1.1.2. For other metrics, $p > 0.1$ for all specialties and all metrics, indicating there was a reasonable probability that the difference in performance metrics observed was by chance. This conclusion remained broadly stable while running our sensitivity tests, but the p values by speciality and metric exhibited a high level of volatility.

7.1.2. We note that the data provided included both inpatient and day case admissions. The use of day case admissions may distort length of stay and readmission results for inpatient admissions, given the natural short-term nature of these procedures. Our analysis has not tested Premier consultants against non-Premier consultants using inpatient admissions only; however in all tests inpatient and day case admissions are included for both Premier and non-Premier, ensuring consistency between the two groups.

7.1.3. We note the potential impact of COVID-19 on the analysis, particularly on the base period of Oct-19 to Sept-20. Vitality has indicated it believes that the impact of COVID-19 would affect both cohorts in a similar way, therefore in comparing the two cohorts it should not distort the analysis unduly. The sensitivity testing reviewed the impact of using pre-pandemic periods of data and demonstrated similar results to the results using the base period at an aggregate level.

7.1.4. Some re-admissions for insured patients are likely to take place in NHS hospitals and hence the re-admissions are likely understated for all specialties. It is not possible to determine whether this affects some specialties more than others, nor is it possible to determine if this impacts the Premier consultants differently than the non-Premier consultants.

7.1.5. We note that the case mix adjustment applied at a speciality level is applied to all performance metrics. Our review has not validated whether the application of the case mix adjustment to all performance metrics in the same way is mathematically correct and valid. We have run our statistical tests on the raw data, before case mix adjustment, and note that the conclusions of our analysis remain largely consistent, although at the speciality level there are some isolated differences.