Dr. Zaid Kimmie
1 Cruden Bay Road, Greenside, 2193
Johannesburg, South Africa
zaidkimmie@gmail.com

October 2020

To: **Council for Medical Schemes Section 59 Investigation Panel**

Final Report: Racial Bias in FWA Outcomes

1 Introduction

On 16 May 2019 the Council for Medical Schemes (CMS) announced that it would launch an investigation into allegations of racial profiling and bullying by medical aid schemes in terms of its regulatory mandate, under section 7(a)(b)(c)(d), 8(a) and (k) and 9(2) of the Medical Schemes Act, 131 of 1998. On 11 June, the CMS announced that the Section 59 Investigating Panel (hereafter referred to as "the Panel") would be chaired by Advocate Tembeka Ngcukaitobi (SC) together with Advocates Adila Hassim and Kerry Williams. The Panel approached the author (Dr Zaid Kimmie) to assist with the inspection and analysis of data relevant to their investigation.

After submitting a report on my findings the three parties covered in that report (Discovery Health, Medscheme and GEMS) submitted their responses to those findings. In this document I will briefly summarise the findings of the first report (Section 2, review the responses to those findings, and finally revisit my findings.

The review of the responses to my findings will be structured as follows:

- In Section 3 I deal with an issue raised in all three responses – that of the effect of non-differential misclassification.
- In Section 4 I discuss the GEMS responses.
- In Section 5 I discuss the Discovery Health responses.
- In Section 6 I discuss the Medscheme responses.

I will consider the following documents in the analysis that follows:

- **GEMS:**
  - *"A Review of the Expert Report Prepared for the Section 59 Investigation Panel"* by Insight Actuaries and Consultants [*Insight Report*]
- **Medscheme:**
  - *Response to the findings of Dr Kimmie – Racial Discrimination in identifying Fraud, Waste and Abuse: Medscheme Report* by Paul Midlane, General Manager Healthcare Forensics at Medscheme [*Medscheme Response*]
  - *"Adjustments to Medscheme's risk ratio using additional variables, and IFM score trends for relevant variables"* by Dr Mike Bergh, OLSPS Analytics [*Bergh Report*]
- **Discovery Health:**
  - "Analysis of fraud, waste and abuse outcomes by race", Discovery Health, 27 January 2020 [*DH1*]
  - *"Follow up submission to the Section 59 Investigation Panel"*, Discovery Health, 14 February 2020 [*DH2*]
  - *"Review of Dr Kimmie's report and Discovery Health analyses of forensic process"* by David Lubinsky [*Lubinsky Report*]

2 The Initial Report

The initial report considered two questions:

1. Was there an explicit racial bias in the algorithms and methods used by Discovery Health, GEMS and Medscheme to identify potential instances of FWA?
2. Were the outcomes of the FWA process racially biased? In particular, were Black practitioners identified as having committed FWA at a higher than expected rate.

This section briefly reviews the methodology used to answer these questions, summarises the main findings of the initial report, and concludes with some general comments on the manner in which the results were made available.

2.1 Methodology

The methodology consisted of:

- A review of the initial submissions made to the Panel, by Discovery, GEMS and Medscheme.
- A review of the methodological questions and data sources that would need to be addressed in order to satisfactorily address the question of whether the outcomes of FWA processes were racially biased.
- On-site interviews with the forensics teams of Medscheme (19 September 2019), Discovery Health (27 September 2019), and GEMS (8 October 2019). At these sessions the various teams demonstrated the implementation of their forensics systems.
- The acquisition of the primary data required to determine whether FWA outcomes were racially biased. This data consisted of the Practice Code Number System (PCNS) database (which contained the PCNS number, discipline, name and surname of all practitioners interacting with medical schemes), and the PCNS numbers and FWA status (whether or not they had been identified as committing FWA) of all practitioners delivering services to each of Discovery, Medscheme and GEMS.
- The classification of all practitioners on the PCNS list as Black or Non-Black based on their surname. This method has been widely used in other contexts but had not, to the best of my knowledge, been previously applied in South African. The surname-based race classification provided a useful proxy through which the racial discrimination in FWA outcomes could be examined.
- The final step consisted of a statistical analysis of FWA outcomes in order to determine whether or not a racial bias existed. The measure used to determine whether a bias existed was the risk ratio, which is defined as the proportion of Black practitioners who have been identified as having committed FWA divided by the proportion of Non-Black practitioners who have been identified as having committed FWA. A risk ratio of 1.5 can therefore be interpreted (in this instance) as Black practitioners being 50% more likely to be identified as having committed FWA than non-Black practitioners. Standard statistical techniques were used to determine whether the calculated risk ratios were likely to have occurred by chance or whether they reflected real differences in the population.

2.2 Results

The initial report found that:

1. There is no explicit use of race in the analytics process. This does not mean that the process is "race-blind". It may be the case that certain factors used, or areas prioritised are asymmetrically distributed by race, and that this may produce racially-biased outcomes.
2. There is a substantial difference in FWA outcomes between Black and non-Black practitioners over the period January 2012 to June 2019. Over this period Black practitioners were 1.4 times more likely to be classified as having committed FWA than those identified as not Black. The probability that this distribution occurred by chance (i.e. that there is no correlation between racial status and FWA outcomes) is for all practical purposes 0 (zero).

The findings with respect to racial bias were based on the racial classification method used in the analysis, and that the inference that this represents a meaningful actual difference was based on an analysis of the scale of the bias (measured by the risk ratio), the probability that the associations were the result of chance events, and an examination of the robustness of the result with respect to the racial classification method. Every assertion in the report was subjected to further tests to ensure that the results were not artifacts of the particular form of racial categorisation used.

The results are also conservative, in the sense that the effect estimate is likely an under-estimate of the true effect. This is a result of the fact that racial classification was conducted independently of FWA status (the technical term for this is non-differential classification) and that the classification method used was conservative – the default classification was non-Black and only names that were overwhelmingly likely to represent Black practitioners were classified as Black.

The racial bias simply represents a correlation between the race classifier and FWA status, and that it may be the case that the relationship is clarified by some intermediate confounding variable, and that the causal relationship is between that variable and the outcome.

The further detailed findings included the following:

- The scale of the deviation increased steadily from 2013 to 2017, at which point Black providers were almost twice as likely to have been identified as committing FWA than non-Black providers.
- Black general practitioners are 1.5 times more likely to be identified as FWA cases than their not Black counterparts;
- The rate at which Black physiotherapists are identified as FWA cases is almost double (1.87) that of their non-Black counterparts;
- Black psychologists are three times more likely to be identified as FWA cases.
- Black registered counsellors and social workers are also three times more likely to identified as FWA cases. More than 50 per cent of Black registered counsellors have been identified as FWA cases – this is the highest rate among the disciplines analysed.
- Black dieticians are 2.5 times more likely to be identified as FWA cases compared to their not Black counterparts.
- The FWA outcomes for each of the Administrators (Discovery Health, GEMS and Medscheme) exhibits clear racial bias, with Black providers significantly more likely to be identified as FWA cases.

- There are clear differences in the scale of racial discrimination between the three schemes – Discovery Health was 35% more likely to identify Black providers as having committed FWA, GEMS was 80% more likely, and Medscheme was 330% more likely to identify Black providers as guilty of FWA.
- For Discovery Health the pattern of racial bias first manifests in 2014 (with a risk ratio of 1.25) and then becomes steadily higher in subsequent years (rising to 1.61 in 2017). The bias in 2018, while still significant, reverted to the 2014 level. Although data for 2019 is limited to the first six months it appears the risk ratio this period is not significantly different from 1, i.e. no bias appears to exist for this period.
- For GEMS the pattern of racial bias is clear from 2013 onwards. The relative risk increases substantially over this period, from 1.5 in 2013 through to 2.5 in 2017.
- For Medscheme the data showed the most variation, largely as a result of the relatively small numbers of FWA cases identified before 2016. From 2016 onwards the risk ratios are high (approximately 4.0 for 2016 and 2017) implying that Black providers were being identified as FWA cases at four times the rate among Non-Black providers.
- Among GPs there was a clear racial bias in the identification of FWA cases, with the bias being significantly higher (risk ratios greater than 2 compared to 1.4 for Discovery) among GEMS and Medscheme.
- Among Pharmacies there was no evidence of racial discrimination within the Discovery and GEMS data. A large proportion of pharmacies were not been racially classified (since pharmacies tend to operate under a corporate identity) but even with these restrictions the Medscheme data showed significant racial bias (a risk ratio of close to 3).
- Among Optometrists the only evidence of racial bias exists within the GEMS data (a risk ratio of 2).
- Among Physiotherapists there is clear evidence of racial bias among all three entities. The risk ratios within the Medscheme ($\geq$ 12) and GEMS ($\geq$ 6) are particularly high. For example, within Medscheme the risk of physiotherapists being identified as FWA cases is 12 times higher than the risk for Non-Black physiotherapists.
- For social workers and dieticians: the pattern of racial bias exists across all three entities.

2.3 General Comments

The analysis was conducted in a completely transparent manner. All of the parties concerned were provided with complete sets of input data (including the PCNS database with two racial classification codes as well as the copies of any data they had supplied and that had been used in the analysis) and a documented copy of the analysis code used to produce the results (this code listed every step of the analysis conducted and allowed the recipient to reproduce any table in the final report). Any uncertainty or confusion with respect to the meaning of any term in the report, or questions about whether any analytical choices were valid could be checked by running the appropriate code or making modifications to that code.

My review of the responses by the various parties was conducted under very different circumstances, as will become clear in later sections. No code, and in many cases no formal definitions of terms were included with the documentation. Checking whether assumptions were valid, or disentangling results when multiple analytical steps were combined was not possible, and recreating the analysis from scratch

was, given the fact that I was working on my own and with significant other demands on my time, simply not feasible.

3 The effect of non-differential misclassification

Each of the parties makes reference to the possibility that errors in the classification process will possibly reduce the scale of racial bias detected. The excerpts below represent the concerns expressed.

"There are various concerns that may be raised about the method used by Dr Kimmie in identifying which practitioners are black and which are not black. There are indications that Dr Kimmie underestimated the total number of black practitioners in the Practice Code Numbering System ("PCNS") database 25 in an attempt to be "conservative". However, if the total number of black practitioners has been underestimated, then the percentage of black practitioners flagged by the FWA investigations could be lower than reflected in Dr Kimmie's report. Discovery's investigations reveal that the total number of black practitioners is likely have been understated by approximately 10%. There are also differences in the methodologies applied which resulted in 5.6% of practitioners being mis-classified. This is important because, as shown below, it is Discovery's submission that a small difference in the risk ratio 28 of practitioners of particular races flagged for further investigations is not legally significant for the purposes of analysis under PEPUDA." [DH2, para 33.1]

"Given that in this exercise, all practitioners were racially categorised (including the ones with less distinctive names), the margin of error is again likely to be higher and therefore caution must be exercised when interpreting the results." [DH1, page 5]

"The results represented are not necessarily conservative as claimed in the report." [DH1, page 12]

"It is the contention of Dr Kimmie that these organisations were defaulted to 'Non-Black' as a race classification as this was the most conservative approach to reaching a fair representative result, however Medscheme disagrees with this view as it dilutes the underlying baseline population against which the racial allocation of FWA cases is compared" [Medscheme Response, page 5]

"... brings into question the robustness of the approach used to infer the race of healthcare practitioners. The experts appointed by the Section 59 Investigation Panel acknowledged that certain practices may be mis-classified but suggested that the correction of any miscalculations would invariably strengthen their results. This is patently incorrect (as indicated in Section 4)" [Insight Report, Section 3.4]

The statement I made in the original report was very precisely formulated: *"As indicated in the methodology section the racial classification has been conservative, with the default classification being Not Black. This will, on the assumption that the classification is independent of the outcome (as is the case here), tend to increase the risk rate among the not Black group, thus reducing the risk ratio."*[1] In slightly different terms, non-differential classification will, in general, tend to bias the estimate towards the null, i.e. produce an estimate lower than the true effect.

This statement is not one over which good-faith scientists can agree to differ, it is a mathematical truth in much the same way as $1 + 1 = 2$. For example on p. 129 of Modern Epidemiology (2nd Edition) by Rothman and Greenland they state *"... the direction of bias introduced by an independent non-differential misclassification of a dichotomous variable is towards the null value"*. They note that this is not true for a non-dichotomous variable, but since we are dealing with a dichotomous variable (Black vs Non-Black) we are on safe ground. One could also refer to the paper "Non-differential misclassification and bias towards the null: a clarification". by S Wacholder, P Hartge, J H Lubin, and M Dosemeci in Occup Environ Med. 1995 Aug; 52(8): 557–558 in which they state *"the most important and the simplest point*

[1] Racial Discrimination in Identifying Fraud, Waste and Abuse, Section 5.2.1

is that non-differential misclassification of a binary exposure (exposed or not) and a perfectly classified binary outcome (diseased or not) does indeed produce a bias toward the null. Always."

No vague hand-waving about the size of the denominator (as in the Medscheme Response quoted above), or under-estimating the **number** of Black FWA cases carries any weight. That the statement is true is provable and can be easily demonstrated using basic algebra. The paper by Wacholder cited above contains a simple example of such a calculation.

The key point is that the racial classification was conducted independently of any knowledge of whether the case concerned had been identified as FWA or not. Discovery Health have reviewed my methods in detail, both internally and through a third party, and have largely agreed with the classification I produced. If there were a bias on my part this process would have revealed it, and I am therefore confident that the classification process has been independent and non-differential (which in this case means that the outcome variable has played no role in the classification process). Given that fact, the statement that the risk ratio produced by this process would tend to under-estimate the true risk ratio is undeniably true.

4 GEMS

The Insight Report raises a number of objections to the original findings. These are:

- That the original analysis does not take into account exposure
- That corporatised, state and group practices should be excluded
- That errors in the classification process increase the risk ratio
- That a similar racial bias exists with respect to the Vuvuzela hotline, over which GEMS has no control

4.1 The question of exposure

I agree that the effect of exposure should be investigated. There are, however, a number of factors to take into account before applying a simplistic measure of exposure.

Firstly, let me deal with the argument put forward by the authors in 3.1.1 of their report.

Assume that all black practitioners reside in a geographic region where only black healthcare practitioners are accessible. In this area, GEMS beneficiaries will only consult with black healthcare practitioners. By implication black healthcare practitioners will have an opportunity to perpetrate fraud, waste or abuse whilst non-black healthcare practitioner will have little opportunity to perpetrate fraud, waste or abuse.

The unavoidable consequence is that only black practitioners will be flagged as possibly guilty of fraud, waste or abuse. Without accounting for exposure and using the methodology employed by the experts appointed by the Section 59 Investigation Panel, one would then conclude that GEMS is more likely to flag black practitioners as possibly guilty of fraud, waste or abuse than non-black practitioners. By implication, the experts would incorrectly determine that GEMS is guilty of racial bias.

The example does not demonstrate what the authors hoped it would, but does, unfortunately, reveal flaws in their understanding of relative risk. Let us put some numbers to the example to see why this is so. Suppose we restricted our analysis to 1000 black doctors (let us just use doctors as a catch-all term for the moment) and 0 non-black doctors in the geographic region described above. Suppose further that 100 of the black doctors are found guilty of FWA. Obviously no non-black doctors are found guilty

of FWA. The relative risk cannot be calculated because there is no risk ratio for non-black doctors – applying the formula would involve dividing by zero:

$$(100/1000)/(0/0)$$

So instead of concluding that black doctors are more likely to be found guilty of FWA the analysis would, quite correctly, conclude that the question cannot be answered, and that no evidence of racial discrimination exists. We can take the example one step further – suppose again that we had 1,000 black doctors but this time that we had 100 non-black doctors. Further assume that 100 black doctors and 5 non-black doctors were guilty of FWA. The risk ratio is:

$$(100/1000)/(5/100) = 0.1/0.05 = 2$$

Would we conclude (as the Insight team would like us to believe) that this represented evidence of racial discrimination because there were so few non-black doctors and therefore that black doctors had "more opportunity" to commit FWA? The risk ratio is certainly high. Once again, however, this line of argument fails. The p-value associated with this risk ratio is 0.149 and we would conclude that this did not provide strong evidence of racial discrimination.

The authors then conclude that:

"This illustrative example is removed from reality. GEMS beneficiaries have access to both black and non- black healthcare practitioners. Nevertheless, GEMS beneficiaries exhibit a far greater propensity to interact with black healthcare practitioners than non-black healthcare practitioners. As such, black practitioners have a far greater opportunity to be identified as flagged as possibly guilty of fraud, waste or abuse. This is not considered by the experts appointed by the Section 59 Investigation Panel." Page 5, Insight Report

I believe that the error of this conclusion has been adequately demonstrated.

However, the question of whether to adjust for exposure (i.e. whether to take account, in some manner, of the number of consultations by each practitioner) is a valid one. There are three possible ways of achieving such an adjustment:

- Simply adjusting by the number of interactions (the method employed by the Insight team);
- Adjusting by some transformed variant of the number of interactions (to mitigate the effect of extreme observations, as detailed below);
- Categorising the exposure variable (into deciles, say) and conducting the analysis separately for each decile.

Calculating the risk ratio using the first method (this is the method employed by the Insight team) introduces the possibility that the risk ratio may fluctuate quite widely based on very small variations in the racial classification scheme. The details of the example will be included in a separate attachment, but in summary suppose that we have 1,010 practitioners distributed as follows:

- 1,000 practitioners who see 500 patients a year (low contact) and 10 practitioners who see 2,500 patients a year (high contact)
- 208 Black practitioners, 200 in the low contact group and 8 in the high contact group
- 802 non-Black practitioners, 800 in the low contact group and 2 in the high contact group

- Among Black practitioners 20 of the low contact group are found guilty of FWA, as are 1 of the high contact group
- Among non-Black practitioners 56 of the low contact group are found guilty of FWA, as are 1 of the high contact group.

Using this data we can derive a risk ratio of 1.42, i.e. that Black practitioners are 42% more likely to be convicted of FWA. If we redo the calculations using the using the exposure variable (low or high contact) we derive a risk ratio of 1.38.

Now we will re-classify **one** of the Black high contact practitioners who have been found guilty of FWA as non-Black. As expected this does not significantly affect the usual risk ratio – this now drops to 1.34, a change of about 5%. However, the calculation using the exposure variable changes dramatically – the risk ratio now drops to 1.05, a decline of 24%. This rather dramatic change (from 1.34 to 1.05) based on the reclassification of a **single** practitioner demonstrates the potential dangers of using the exposure variable in its raw form.

One of the alternatives is to transform the exposure variable to reduce the effect of outliers, in this case weighting observation by the natural log of the number of contacts. Using this method the risk ratio remains relatively constant (dropping from 1.41 to 1.31).

The final potential method is to stratify the population by exposure categories and to conduct the analysis separately for each exposure category. I will not deal with this possibility in any further detail but such a method should certainly be given due consideration.

The point of these remarks is that adjusting for exposure is not a simple matter, and that the advantages and deficiencies of various approaches need to be carefully considered. Presenting the results on only one such choice without considering (and reporting on) alternatives is not sound practice.

Finally, in their calculations (Annexure A, but also Annexure B, of the Insight report) the experts use a somewhat strange method to calculate the distribution of Black and non-Black practitioners. Rather than take the universe of all practitioners who have interacted with GEMS over this period (a total of 55,718) the authors appear to have summed the number of practitioners per year, arriving at a total of 266,648. Although this does not, at first sight, appear to have introduced any serious errors this is a strange choice. I am not able, given the absence of the raw data, to determine whether or not this choice does in fact affect the results of their analysis. In particular, given that they explicitly compare the results of their analysis with mine, and that they had full access to the data, code and methods that were used to generate my results, this choice is puzzling.

4.2 Exclusion of Corporatised, State and Group Practices

The Insight report (in sections 3.2 and 3.3) suggests that disciplines associated with state or corporatised practices, as well as group practices, be excluded. As detailed in my original report (see Sections 3.1.1 and 3.1.2) I took great care to ensure that the method of racial classification did not introduce any bias into the results, and that using a conservative classification method would tend to under-estimate the effect of any racial bias. There is therefore no reason to suspect that exclusion of these practices would significantly affect the results of the study.

The one area that may merit further investigation is the exclusion of pharmacies – although it is true that these may be corporatised practices this does not mean that the ownership structure cannot be determined. In general, where we have cases where there is little doubt about the racial classification of

the practitioner it would appear to be reasonable to make such a determination rather than eliminating an entire sub-group of practices from our analysis.

Unfortunately the authors have not conducted (or at least have given us no evidence of having conducted) an analysis of the immediate effect of these exclusions. I will expand on this point below, but for the moment only note that the authors conflate the results of multiple steps when they present their results – in their Table 5 they present the results of adjusting for exposure, removing corporatised and state practices, and removing group practices in a single step. A more transparent presentation would have allowed for the examination of the effect of each of these steps separately. This is not some minor quibble – in Table 5 of Section 4 the authors present my original estimate of the risk ratio (1.74 over the entire period) and contrast it to their own estimate of 1.54 after *"adjusting for exposure, removing corporatised and state disciplines and removing registered group practices"*. Even though the change is relatively small presenting the results of multiple effects in this way creates the impression that each effect individually affected the risk ratio in the same direction.

4.3 Classification Errors

In Section 3.4 the authors argue that *"This suggests that the classification prepared by the experts cannot be replied upon"*. They go on to conclude that the incorrect classification *"brings into question the robustness of the approach used to infer the race of healthcare practitioners. The experts appointed by the Section 59 Investigation Panel acknowledged that certain practices may be mis-classified but suggested that the correction of any miscalculations would invariably strengthen their results. This is patently incorrect (as indicated in Section 4)"*.

I have dealt with this misconception in Section 3. However, some additional points are worth making:

- In order to check the accuracy of the classification scheme the authors conducted an audit of the 800 practitioners with the **highest** number of interactions (see 3.4.1 of the Insight Report). This is not a random sample of all practitioners and hence the results cannot be generalised to the entire classification process. If this were not a sufficiently egregious mis-step the authors then use this non-random sample to effect changes to the classification scheme.

- The authors find that in this non-random sample 13% of the practices could have been classified as Black rather than non-Black. This is then the basis on which they conclude that the classification is not reliable. However, the finding that a significant proportion of Black practitioners are classified as non-Black should not come as much of a surprise – as indicated in Section 3.1.2 of my original report the classification is purposely conservative. The question is not, in fact, whether the classification under-counts the number of Black practitioners but what the effect of such an under-count will be.

  The authors base their finding that this mathematical truth is "patently incorrect" by taking a biased sample (the 800 practitioners with the highest number of interactions), re-classifying the Black/Non-Black status by correcting errors in this sample, and then comparing the corrected vs uncorrected risk ratios based on weighted usage (adjusting for exposure). This is the very antithesis of independent non-differential misclassification.

  I take particular exception to the authors misquoting my original statement by restating my claim as "any miscalculation would invariably strengthen (my) result". My statement was "As indicated in the methodology section the racial classification has been conservative, with the default classification being Not Black. This will, on the assumption that the classification is independent of the outcome (as is the case here), tend to increase the risk rate among the not Black group, thus reducing the risk ratio.".

4.4 Vuvuzela Hotline

In Section 5 of the Insight document the authors argue that since Black practitioners are 45% more likely to be flagged as possibly guilty of FWA via the Vuvuzela hotline their own calculation that Black practitioners are 47% more likely to be found guilty of FWA is due to factors that cannot be controlled by GEMS. Their claim is that "This implies that GEMS processes are not racially biased and that the results are indicative of other extenuating factors which differentiate between black and non-black practitioners (as reflected upon by members)." They offer no further insights into what those "extenuating factors" may be.

The logic upon which this argument is based is unfortunately flawed. The authors, whether by design or accident, compare two different outcomes – one based on accusations by members ("flagged as possibly guilty of FWA") to one based on an actual finding by the scheme ("found guilty"). Even if one overlooks this error the remainder of their argument reduces to the claim that GEMS is not guilty of any greater racial discrimination than the GEMS members who lay complaints via the Vuvuzela hotline.

4.5 Summary

After listing their view on the technical deficiencies in the original analysis the authors conclude that:

"These shortcomings materially distort results."

This is a somewhat strange conclusion to reach. On the best version of their analysis (i.e. assuming that all of the objections I have raised above are dismissed) they conclude that the risk ratio will drop from 1.78 to 1.47, a decline of 21%. At no point does do these findings rise to the challenge of determining that no racial discrimination with respect to the outcomes of FWA processes exist.

Finally the authors state that:

"Shortcomings relating to the interpretation of results pertain to:
The mistaking of a difference between black and non-black practitioners as racial bias as indicated by the fact that the GEMS results are consistent with that of a wholly independent process."

The clear implication of this line of argument is that black practitioners are in fact more likely to commit FWA.

4.6 Conclusion

The response to the expert report does not invalidate the original finding. The authors commit a number of mis-steps in their attempt to undermine the original findings. These include:

- The example they propose to demonstrate the effect of exposure is flawed, and reveals a distorted view of what a risk ratio is and how statistical significance is determined.
- The authors do not appear to understand the effect of non-differential misclassification errors, and persist in the false claim (backed up by a biased sample) that correcting such misclassifications will reduce the risk estimate.
- The authors raise a valid point about taking into account exposure, but fail to interrogate the possible side-effects of using their method to account for exposure. I note that in Section 5.2.1 of my original report I go into some detail about conducting sensitivity analyses (i.e. checking the extent to which the risk estimate is stable when changes are made to the classification scheme) of the calculated risk ratio.

- The authors use a total of 266,648 practitioners (or 212,315 after removing certain practices) in their calculations.[2] This is odd since there are only 51,486 practitioners who have interacted with GEMS over this period. It is not clear what effect this has on their calculations, but it is certainly not the basis on which the original risk ratios were calculated.

However, even if we accept all of their calculations the outcome is one which disputes only the scale of the racial bias.

Finally, their argument that this residual racial bias should be dismissed because it matches a racial bias in the whistle-blower process also does not stand. This comparison is invalid because it conflates two different quantities – allegations of FWA by members with findings of FWA by the scheme. The follow-on argument, which holds only if we accept this false comparison, is not a statistical argument but a socio-political one about the differential distribution of criminal tendencies.

5 Discovery Health

Discovery Health constructed their own racial classification system and had this classification scheme audited by an external service provider. They concluded that:

As evident from Table 9.4, using the above methodology, we were able to replicate Dr Kimmie's results almost exactly. Our analysis found that Black practitioners are 1.36 times more likely to be classified as having committed FWA than those identified as Non-Black practitioners. This is comparable to Dr Kimmie's risk ratio of 1.34.

However, in their detailed analysis the DH team produced a number of objections to the this finding:

- That race classification errors may lead to an overstatement of the risk ratio
- That the investigative process showed no evidence of racial discrimination
- That adjusting for possible confounding factors substantially reduces the risk ratio

Before dealing with these in detail I shall clarify some misconceptions about my findings as reflected in the DH documents. The DH team conclude that: *In addition to the lack of any evidence of explicit bias, neither Dr Kimmie nor our internal and external assessments have been able to identify any potential sources of implicit bias in any of the 30 the factors which are used in the RRT analysis.* [DH1, page 8]

and

This is a critical finding as it confirms that there is no deliberate profiling of Black practitioners in the way in which FWA cases are identified. [DH1, page 23]

Both of these statements misrepresent my findings. Both the internal assessment and my own work only looked at explicit biases. I clearly note that any investigation of explicit biases are beyond the scope of my current work. On page 19 of my original report I note:

In conclusion, there is no explicit use of race in the analytics process. It should be noted that this does not mean that the process is "race-blind". It may be the case that certain factors used, or areas prioritised are asymmetrically distributed by race, and that this may produce racially-biased outcomes.

[2] See Table 8 in Appendix A of the Insight Report

5.1 Classification Errors

I have dealt with the general issue of classification errors above. However, I will spend some time on the example provided by DH on page 22 of DH1. This example claims to demonstrate that the premise of my analysis (that a conservative classification scheme will under-estimate the level of discrimination) is false. I do not have access to the underlying data that produced their table[3] so am unable to verify the underlying assumptions behind the table. In essence DH wish to demonstrate that the risk ratio will potentially increase if a less conservative racial classification scheme is used. The example is flawed on a number of counts:

- Table 1.5 on page 22 uses data from 2015 to June 2019 for GPs and derives a risk ratio of 1.17 using the DH racial classification. This is then compared to the risk ratio of 1.38 calculated in Table 5.3 of the report prepared for DH. However, the table it purports to compare itself to uses data from **2012** to June 2019. It is not clear why a different set of dates was used, but that choice invalidates the intended comparison.
- Note that DH and my own estimates of the total number of Black practitioners over the period 2012 to June 2019 differs by less than 1% (17,691 to 17,506) and that our estimates of the number of Black FWA cases differs by about 0.3% (5,061 to 5,075). It is therefore somewhat strange that over the period 2015 to June 2019 DH produce an estimate for the number of Black GPs (5,981) that is substantially higher than my estimate for the total number of Black GPs over the full period (5,228). This is a difference of about 12%. Finally, DH estimate that there were 1,148 FWA cases among this group compared to my estimate of 1,725, a difference of -577 or 50%. It is difficult to reconcile these numbers – the number of Black GPs increases and the number of FWA cases among this group drops precipitously. After these differences are taken into account it is not surprising that one will find a lower risk ratio. It is somewhat remarkable, given the coherence among the global figures, that such a difference would exist for this sub-group.

5.2 Investigative Process

Table 2 shows that the total number of investigations conducted on Black practitioners is disproportionate to the underlying race distribution of the population of billing practices submitting claims to DH. Table 2 also shows that there is no difference in the proportion of investigations that result in a valid FWA finding between practitioners classified as Black and Non-Black. Further, there is no difference in the proportion of cases where a recovery was made between practitioners who are Black or Non-Black. These results confirm that any disproportion in FWA outcomes arises in the initiation of investigations, and that once an investigation is initiated, there is no difference in the way in which forensic investigations are conducted between Black and Non-Black practitioners. This is consistent with the findings of the review conducted by Harris Nupen Molebatsi (HNM) (included in the DH submission to the Panel of 19 July 2019). These results demonstrate that there is consistency and uniformity across race groups in the investigative process undertaken by DH employees once a case has been initiated. More importantly, it stresses that there is no racial discrimination in the investigative processes applied by DH. [DH1, page 7]

There is a significant flaw in this analysis. Let us start from the default assumption that the rate of FWA does not differ by race (Black/Not-Black in our case). Using the figures referenced in the report we can assume that the rate of FWA is 7,493/41,715, which is about 18%. Suppose we now over-sample, for whatever reason, the Black group when identifying potential cases (as is the case in table 2 in the report). We would then expect that that the number of true positives (accepting for the moment that

[3] As noted above when my analysis was presented to DH it included all of the data and coding routines files needed to reproduce the tables.

the results of the investigation represent the true outcome) would differ between the two groups. In particular, since we over-sampled the Black group (by a factor of almost 2) we would expect to have collected significantly fewer real positives. However, no such difference is recorded. This does not prove that the investigative processes are in fact unbiased since it is just as likely that these processes are biased against Black practitioners.

The DH report continues:

Dr Kimmie's report and his testimony have confirmed that there is no explicit racial bias in the DH analytics systems used to identify potential FWA cases. This is a critical finding as it confirms that there is no deliberate profiling of Black practitioners in the way we identify FWA cases. [DH1, page 8]

This statement mis-represents my findings by conflating the meaning of explicit and deliberate. When we say there is no "explicit" racial bias this means that we can find no component of the measures that directly includes race in the criteria by which potential cases are identified. This does not exclude the use of "implicit" criteria which may bias the selection, the use of discretionary criteria by people at various stages of the system, nor indeed the fact that such application may have been deliberate.

5.3 Confounding

The DH report introduces a set of possible confounding factors and then show that controlling for these possible confounding factors reduces the risk ratio from 1.36 to 1.09. The DH team then claim that an examination of further confounding factors may reduce the risk ratio even further.

No formal definition of a confounder is ever provided by the Discovery team or their expert, and that this oversight materially undermines their analysis. In particular the Discovery analysis appears to rely on the erroneous assumption that confounding can be determined simply by examining the effect of including such a variable on the risk ratio (see for example the comments by Dr Broomberg on line 10, page 129 of the transcript). While this may be the version of confounding undergraduates are initially exposed to it is not in fact the version used by practising statisticians.

In particular the formal definition (see for example a standard text such as "Modern Epidemiology" by Rothman and Greenland) explicitly disallows consideration of a variable as a confounder if it is affected by the exposure variable (race in this case) or the outcome variable (FWA status in this case). In particular a variable cannot be a confounder if it lies on the causal path between exposure and the outcome. In such a case "adjusting" for the confounder would artificially lower the size of the effect.

In particular, if the potential confounder lies on a causal pathway (acting as a mediator say) then controlling for this variable would reduce the risk ratio **because** it attenuates the causal effect. A simple example would be the relationship between race and employment in South Africa. One could control for the "confounding" effect of education and find that the relationship between race and employment is reduced. This would be an error, since education lies on the causal path between race and employment (in simple terms education is affected by race). There is no evidence that any such considerations were taken into account by the DH team.

The DH analysis suggests that direct payments (whether or not the practitioner is on direct payment, i.e. the practitioner is paid directly by the scheme rather than by the patient) is potentially a confounder. However, direct payments is clearly a consequence of race. Black practitioners are, given the socio-economic circumstances of their patients, more likely to be paid directly by the medical scheme. This was explicitly conceded by Dr Broomberg in his verbal evidence (see page 134 of the transcript).

The analysis by Discovery (as reflected in the submission of 27 January 2020) contains no indication that the authors are aware of the formal definition of a confounder or of the dangers of incorrectly adjusting

for a variable that is not a confounder. The commentary by their independent expert (who confirms that the analysis was done correctly) uses a simplistic definition of a confounder extracted from a website that attempts to "explain" confounding in 7 paragraphs without any formal definitions. The verbal evidence presented by Dr Broomberg similarly reveals a flawed understanding of confounding. When questioned by the Panel about the link between race and direct payment (which would therefore exclude the use of this factor as a confounder) Dr Broomberg incorrectly continues to insist that the aforementioned linkage is of no concern (page 136 of the transcript).

The technical details of causal analysis require some advanced study, but I will quote extensively from a recent article by 47 editors of Respiratory, Sleep and Critical Care Journals which provides some guidance on the control and reporting of confounding.[4] This article does not break any new ground but synthesizes the current state of our knowledge about how best to handle potential confounding.

In a section headed "Variable Selection Methods That Do No Adequately Control for Confounding" the authors note that:

"P value–based and model-based variable selection methods (including forward, backward, and stepwise selection) should not be used for causal inference."

Such an approach would be tantamount to attempting, seeing what "works" (in the sense that the risk ratio changes sufficiently) and then post-facto finding a justification of why the factor is a confounder. This type of approach is not statistically sound – it would tend to encourage continual searching for "confounders" until one finds the answer one wants.

The authors continue:

"These approaches ignore the causal structure underlying the hypothesis and therefore do not adequately control for confounding. Confounders and colliders are treated similarly. . . . Selection of variables that, when included in a model, change the magnitude of the effect estimate of the exposure of interest should not be used to identify confounders, for the reasons discussed above. Identification of multiple "independent predictors" ("winners") through purposeful or automated variable selection is an unacceptable approach for testing causal associations. If the authors have hypotheses about each variable, then a separate model for each variable should be generated using one of the above preferred approaches."

The DH approach of hunting for variables that change the outcome measure (in this case the risk ratio) and then post facto declaring that they are confounders amounts to nothing more than a fishing expedition. The guidance by the editors referred to above is to identify potential confounder based on the state of existing knowledge and causal relationships. The danger of fishing for variables that affect the outcome measure is that once such a variable is found and identified as a "confounder" the search for a post-hoc explanation then commences. The lack of a theoretical framework in which valid causal relationships can be investigated leads inevitably to incoherent causal explanations. If, for example, one were to find that the hair colour confounded the relationship between smoking and lung cancer (because controlling for hair colour changed the risk ratio) one would be tempted to concoct all sorts of superficially reasonable sounding causal explanations for a phenomenon with no physical manifestation.

This sort of difficulty is adequately demonstrated by Dr Broomberg's attempts to explain why the year in which FWA occurred is a potential confounder ("Beyond that I am afraid to say I can't think of a reason why year would be a confounding factor save to say that it is a confounding factor", page 129 of the transcript).

A further factor posited as a potential confounder by DH is whether or not the practitioner was the

[4] Control of Confounding and Reporting of Results in Causal Inference Studies, Guidance for Authors from Editors of Respiratory, Sleep, and Critical Care Journals, Lederer et al, Annals ATS Volume 16 Number 1 January 2019

subject of a tip-off. The question of what the causal mechanism by which the confounding occurs is not addressed in the written submissions and was also not covered in the verbal evidence.

5.4 Conclusion

The DH review has confirmed my finding of a racial bias in proportion of Black vs non-Black practitioners found to have committed FWA. They have further confirmed that the methods employed with respect to racial classification were not biased.

The DH team demonstrate only a very basic understanding of confounding, and of how to control for confounding. The approach of sequentially introducing variables into the analysis, assessing whether the risk ratio has changed, and if so labelling the variable a confounder may pass muster in an introductory analysis course but is not acceptable in any serious analysis.

The question of whether there are confounding factors has not been settled, and the DH attempts to conduct such an analysis are not credible.

6 Medscheme

The response by Medscheme and their expert (as reflected in the Bergh Report) takes three approaches to contest the finding of racial bias in FWA outcomes:

1. That the analysis should be restricted to only certain categories of practices, identified either by discipline or by origin of the complaint
2. That the analysis should take into account usage factors, and in particular the number of claim lines processed per practitioner.
3. That correcting misclassification errors would reduce the scale of the bias.

6.1 Removing Juristic Entities

The Bergh report argues (page 4) that "juristic entities have no racial identity and should be excluded from the determination of risk ratios". This may indeed be the case for many of the entities which are listed in Table 1 of that report – however, the claim that none of the pharmacies can be assigned a racial identity is not true. The classification I completed may have been conservative but it did manage to classify approximately 10% of pharmacies. This is an important step since, even with this conservative classification the risk ratio for Black vs Non-Black within this group was significant.

No analysis is presented on what the effect of this exclusion is on the risk ratio, nor how many practices are affected. The step is never reconsidered and all further analysis proceeds with excluding these cases.

6.2 Removing Whistleblower Cases

From Page 5 of Bergh Report:

"As detailed in Medscheme's Main Submission, the Whistle-blower tip-offs and referrals we receive are entirely independent and Medscheme Forensics have no influence or control over the racial demographic of the healthcare practice against whom the allegation is made. From a governance perspective, a medical scheme (or the Administrator in their capacity as their agent) has a fiduciary duty to investigate every tip-off received as part of good fraud risk management and assurance controls. Failure to investigate and act upon a tip-off, irrespective of the source, would render scheme Trustees and Management in breach of these fiduciary duties and in contravention of the Medical Schemes Act. It is factually accurate to

describe these externally independent sources of forensic investigation as obligatory / compulsory.Taking the above into account, there is no possibility of implicit racial profiling by Medscheme in those specific investigations. They are external, independent and compulsory. We do not consciously or unconsciously select those cases. Even if there may potentially exist statistically biased racial outcomes in those cases (which is denied for reasons set out elsewhere in this response), Medscheme Forensics cannot be determined to be the causal factor for those outcomes."

"It seems therefore that Medscheme has no control or influence over cases that are reported via the whistle blower or tip-off mechanism. As a result the methods followed here remove all FWA case resulting from whistle-blower tip-offs from further analyses and from the computation of risk ratios."

This is an unfortunate misinterpretation of the question under consideration. The question of interest is one of outcomes – are **findings** of FWA disproportionately being made against a certain group of practitioners – not whether the set of investigations exhibits such a bias, nor whether or not any explicit intent was involved. For example, let us assume that the set of all complaints received via tip-offs are unbiased. It is, however, entirely conceivable that the results of the investigations, i.e. the finding that a practitioner is guilty of FWA, may exhibit such a bias. To take a simple example: A scheme receives 1,000 complaints of possible fraud, with 500 against Black and 500 against Non-Black practitioners, and after investigation concludes that 400 of the Black and 120 of the Non-Black practitioners are in fact guilty of FWA. This would, absent further information, be suggestive of a racially biased outcome. The risk ratio in this instance would be:

$$\frac{400}{500} / \frac{120}{500} = 0.8/0.24 = 3.33$$

To use a contemporary example – if a police force were accused of differential treatment of Black and Non-Black perpetrators it would not be credible to claim that all instances where members of the public had reported crimes should be excluded on the basis that the police force has no control over such incidents, and are in fact legally obliged to conduct an investigation. Whether or not their conduct is unfair, or discriminatory, would then not be conveniently ignored.

Unfortunately this choice by the analysts taints the remainder of their analysis – once this group is removed they can then proceed with further (potentially erroneous) components of their analysis based on a false assumption.

6.3 Claim Lines Weighting

I agree that the question of whether to adjust for exposure (i.e. whether to take account, in some manner, the number of consultations by each practitioner) is a valid one. The analysis presented in the Bergh report chooses claim lines rather than contact visits as their method of adjustment – it is not clear why alternatives were not examined, but we will proceed with this choice. There are then three possible ways of achieving such an adjustment:

- Simply adjusting by the number of claim lines (the method employed in the Bergh Report);
- Adjusting by some transformed variant of the number of interactions (to mitigate the effect of extreme observations, as detailed below);
- Categorising the exposure variable (into deciles, say) and conducting the analysis separately for each decile.

Calculating the risk ratio using the first method introduces the possibility that the risk ratio may fluctuate quite widely based on very small variations in the racial classification scheme. This is in fact recognised in

the Bergh report when the analyst plots (in Figure 2 and Figure 3) the histograms of the distribution of claim lines and the natural log of claim lines. The details of the example will be included in an appendix, but in summary suppose that we have 1,010 practitioners distributed as follows:

- 1,000 practitioners who see 500 patients a year (low contact) and 10 practitioners who see 2,500 patients a year (high contact)
- 208 Black practitioners, 200 in the low contact group and 8 in the high contact group
- 802 non-Black practitioners, 800 in the low contact group and 2 in the high contact group
- Among Black practitioners 20 of the low contact group are found guilty of FWA, as are 1 of the high contact group
- Among non-Black practitioners 56 of the low contact group are found guilty of FWA, as are 1 of the high contact group.

Using this data we can derive a risk ratio of 1.42, i.e. that Black practitioners are 42% more likely to be convicted of FWA. If we redo the calculations using the using the exposure variable (low or high contact) we derive a risk ratio of 1.38.

Now we will re-classify **one** of the Black high contact practitioners who have been found guilty of FWA as non-Black. As expected this does not significantly affect the usual risk ratio – this now drops to 1.34, a change of about 5%. However, the calculation using the exposure variable changes dramatically – the risk ratio now drops to 1.05, a decline of 24%. This rather dramatic change (from 1.34 to 1.05) based on the reclassification of a **single** practitioner demonstrates the potential dangers of using the exposure variable in its raw form.

One of the alternatives is to transform the exposure variable to reduce the effect of outliers, in this case weighting observation by the natural log of the number of contacts. Using this method the risk ratio remains relatively constant (dropping from 1.41 to 1.31).

This method has not been considered (or was perhaps not even tested) by the analyst. The point of these remarks is that adjusting for exposure is not a simple matter, and that the advantages and deficiencies of various approaches need to be carefully considered.

6.4 Removing Auxiliary Providers

The analyst notes that there is a much higher risk of FWA among auxiliary providers and that the IFM system (at least in 2019) tended to flag Black providers in this category. It is worth noting that the risk ratio (for Medscheme) was significantly higher in these categories.

The analysis then concluded that, since the IFM scores were higher there is no concern about racially biased outcomes and we can thus safely ignore this group of practices. The new risk ratio is now 1.02.

6.5 On or Off Network

Once again, the argument is that off-network providers have higher IFM scores and that since the IFM system is race-blind we can assume that there is no problem. The risk ratio when off-network providers are excluded now falls to 0.88.

6.6 Conclusion

This analysis presented by Medscheme and their expert, is essentially that once we exclude all the problematic areas, and apply an (possibly flawed) adjustment based on claim lines, it appears that Medscheme may in fact be biased against Non-Black practitioners!

This is not a credible analysis. It is fatally flawed by its initial assumptions (permanently excluding whistle-blower cases and choosing the raw number of claim lines as an adjustment mechanism), and then proceeds to remove areas of concern until an acceptable answer is returned.

7 Conclusions

There are certainly areas in which the original analysis could be improved. I should note that none of the additional factors suggested were available when the first round of analysis was completed.

However, I have seen no convincing argument that the original findings are substantially incorrect. The base facts remain – there is firm evidence that there is a racial bias in FWA outcomes, in the sense that Black practitioners are more likely to be found guilty of FWA then their non-Black counterparts. There may, as I stated in the original report, be external factors that may help to explain this disparity, but given the existence of centuries of systemic racial discrimination we cannot blithely dismiss the notion that racial prejudice has played some part.