PEER REVIEW HISTORY

BMJ Open publishes all reviews undertaken for accepted manuscripts. Reviewers are asked to complete a checklist review form (**http://bmjopen.bmj.com/site/about/resources/checklist.pdf**) and are provided with free text boxes to elaborate on their assessment. These free text comments are reproduced below.

ARTICLE DETAILS

TITLE (PROVISIONAL)	The impact of general practitioners' gender on process indicators in Hungarian primary health care: A nationwide cross-sectional study
AUTHORS	Kovács, Nóra; Varga, Orsolya; Nagy, Attila; Pálinkás, Anita; Sipos, Valéria; Kőrösi, László; Ádány, Róza; Sándor, János

VERSION 1 – REVIEW

REVIEWER	Grant Russell Monash University Melbourne Australia
REVIEW RETURNED	16-Nov-2018

GENERAL COMMENTS	Thanks for the opportunity to review the article The impact of general practitioners' gender on process indicators in primary health care: A nationwide cross- sectional study for BMJ Open. The study intended to illuminate the links between GP gender and primary care performance in GPs in Hungary. It used a cross sectional approach where the authors interrogated national health administrative data bases relating to general practices in Hungary. The aim was to (1) to investigate the association between GP gender and the quality ofprimary care with respect to various process indicators for practice performance and (2) to assess the size of the gender impact. While the artticle was carefully presented, I felt that there were a few examples where the quality of the written English detracted from the quality of the article. Terms such as "considerably high" (abstract conclusion) "patient centred care communication style is applied mainly by females" (background) GMP settlement type /rural or urban/ (in the methods). were among a number of relatively isolated examples where a review of the English would be beneficial. I had several suggestions concerning the paper, each relate to sections of the work. • The Background framed the problem iin terms of the evidence around gender and outcomes in terms of chronic disease management and prevention. The case for he research is reasonable and clear. • I liked the comprehensive approach in the presentation of the methods and how there appears to have been significant thought put into the unfolding analysis. It would have been helpful if the reader could have ascertained how several clinicial diagnostic measures were assessed. Particulary helpful would would have been to know how hypertension, COPD and T2 diabetes were

determined as cases in the data set.

• However, the analysis did have some important gaps - it was not clear as to whether the analysis was undertaken at the level of the practice or the level of the practitioners. The authors used the term GMP seemingly interchangably. This is compounded when the abstract had an error where it stated that there were 4575 GMPs instead of 4575 GPs

• The results are very difficult to interpret without having an understanding of the existence of group practice in Hungarian Primary care. Dahouge's Ontario sudy of physician gender needed to make considerable adjustments to incorporate the complexity of group practice size, and other characteristics. This is important in that I cant see how physician gender at the level of the practice was ascertained to account for the differng numbers of providers that may work there.

• There is no mention of cluster analysis - I would have thought that this needed to be accounted for

• In addition there seemed to be no controls for the number of hours worked by providers. For example one would have expected that a clinician with a panel size of 1500 who was working 2 days per week would have a very different pattern of practice than one working 5 days per week.

• The calculation of standardised prevention ratios could be better described - an example would be useful here.

• Finally in terms of presentation of the results I think that the authors should be careful about confusing clinical to statistical significance - especially with the very large sample. I would struggle to be very excited abou a 1% difference in preventive performance on influenza vaccination for example (p11 line 3)

• The discussion puts the results into the context of the broader literature and drew a few threads togwther concerning the "missed care" that followed as a consequence of male gender. Limitations were reasonable, however missed several that may be of some relevance, in particular the issue of the analysis not taking account of gender concordance, patient complexity (male GPs tend to see more complex patients than females) and clustering.

The article would benefit from more context on the Hungarian health care system - it was impossible to ascertain the degree to which the system required patients to enrol with a GP, the prevalence of group practice and the presence or absence of incentives for preventive activities.
There are some other errrors worth addressing -
• there seems a discrepancy in the p value between the text and the table as related to influenza vaccination
• I feel that it is important to explain the demographics in table 3 - I am a little surprised as to why the prevalence of individuals requiring influenza vaccination was less than the numbers presumably on anihypertensives and requiring a serum creatinine measure.
Finally, I am not a statistician, but was surprised to see that the authors analysed the data using a linear regression rather than has been done in other papers of this type, a logistic regression for

	the chronic disease indicators (see Dahrouge et al 2016 Medical Care) Minor additional suggestions are that the work lacked spaces or indents for paragraphs - this made it much more difficult to read, although may be a consequence of the process of preparing the manuscript for review. Thank you for the opportunity to review the article, I hope the review has been of assistance.

REVIEWER	B Stuart University of Southampton UK
REVIEW RETURNED	17-Nov-2018

GENERAL COMMENTS	This is an interesting paper which is clearly written. I was asked to review it as a statistician so I'll leave the subject area to those who are better able to comment. The authors describe the outcomes and the analysis very clearly. The tables are clear and support the results, with the exception of the supplementary table, which I found difficult to follow. However, I have a couple of concerns that require either clarification or perhaps further analysis. The first is that the data appears to be hierarchical - patients are nested in GPs who are in turn nested in GMPs. But the model does not seem to take this into account. It seemed to me that a multi-level model would be more appropriate here. The outcomes are likely to cluster at both GP and practice level and I fear that failing to account for this in the model could lead to incorrect conclusions. I was also puzzled by the decision to turn a number of continuous variables into categorical ones. Doing this leads to a loss of statistical power and I couldn't see a clear rationale for doing so. It would also be helpful to understand why the authors are reporting the standardised regression coefficients. Most readers will be more familiar with the understandardised coefficients so the authors should explain this choice and give some interpretation of the coefficients as it is not straightforward, particularly in the presence of multi-level data. Some of the differences that the authors comment upon in the Results section are statistically significant. However, I would disagree that the results shown in Table 2 are "remarkable differences". Many are small and their statistical significance may instead relate to the large sample. All are less than 5%. Even in the multivariate analysis, the coefficients and r-squared values are not large even if they are significant. It is unclear whether this is a clinically meaningful difference. The authors do mention the possibility of unmeasured confounding in their limitations section. i think this needs a bit more emphasis in the discussion. There are a limited number of variables in this model, particularly at the patient level, which might account for some or all of the observed association. Another issue may be the sample itself. The authors mention in the Background that 56.2% of physicians in Hungary are female. But in this sample, the proportion is 48.37%. It would be helpful to explain this discrepancy and to comment on whether this could be due to selection bias.

VERSION 1 – AUTHOR RESPONSE

Response to Reviewer 1:

1) WHILE THE ARTTICLE WAS CAREFULLY PRESENTED, I FELT THAT THERE WERE A FEW EXAMPLES WHERE THE QUALITY OF THE WRITTEN ENGLISH DETRACTED FROM THE QUALITY OF THE ARTICLE. TERMS SUCH AS "CONSIDERABLY HIGH" (ABSTRACT CONCLUSION) "PATIENT-FOCUSED CARE COMMUNICATION STYLE IS APPLIED MAINLY BY FEMALES" (BACKGROUND) GMP SETTLEMENT TYPE /RURAL OR URBAN/ (IN THE METHODS). WERE AMONG A NUMBER OF RELATIVELY ISOLATED EXAMPLES WHERE A REVIEW OF THE ENGLISH WOULD BE BENEFICIAL.
Response: The English of the original manuscript was corrected and proofread by English Language Editing Services of the American Journal Experts. The editorial certificate has been attached. However, revised version has been improved based on your comments.
The above mentioned terms have been corrected: “considerably high” changed to “notable”; "patient-focused care communication style is applied mainly by females" changed to “patient-centred communication style, which is more frequently applied by female physicians”;
“GMP settlement type /rural or urban/” changed to “types of settlement (rural, urban)”.

2) THE BACKGROUND FRAMED THE PROBLEM IIN TERMS OF THE EVIDENCE AROUND GENDER AND OUTCOMES IN TERMS OF CHRONIC DISEASE MANAGEMENT AND PREVENTION. THE CASE FOR THE RESEARCH IS REASONABLE AND CLEAR.
Response: Thank for this comment.

3) I LIKED THE COMPREHENSIVE APPROACH IN THE PRESENTATION OF THE METHODS AND HOW THERE APPEARS TO HAVE BEEN SIGNIFICANT THOUGHT PUT INTO THE UNFOLDING ANALYSIS. IT WOULD HAVE BEEN HELPFUL IF THE READER COULD HAVE ASCERTAINED HOW SEVERAL CLINICIAL DIAGNOSTIC MEASURES WERE ASSESSED. PARTICULARY HELPFUL WOULD WOULD HAVE BEEN TO KNOW HOW HYPERTENSION, COPD AND T2 DIABETES WERE DETERMINED AS CASES IN THE DATA SET.
Response: We completed the Table 1 with detailed description of the target group. Patients with hypertension or diabetes mellitus were defined on the basis of drug consumption (which is monitored by the National Institute of Health Insurance Fund Management). Patients with COPD were identified by drug consumption and by participation in pulmonary function test (which are monitored by the National Institute of Health Insurance Fund Management as well).

4) HOWEVER, THE ANALYSIS DID HAVE SOME IMPORTANT GAPS - IT WAS NOT CLEAR AS TO WHETHER THE ANALYSIS WAS UNDERTAKEN AT THE LEVEL OF THE PRACTICE OR THE LEVEL OF THE PRACTITIONERS. THE AUTHORS USED THE TERM GMP SEEMINGLY INTERCHANGABLY. THIS IS COMPOUNDED WHEN THE ABSTRACT HAD AN ERROR WHERE IT STATED THAT THERE WERE 4575 GMPS INSTEAD OF 4575 GPS
Response: The existence of group practices is absolutely not typical in Hungary. GPs are working in solo practices, since each GMP belongs to a GP who is responsible for care in his/her own practice. The analysis was undertaken at the level of the pactice (GMP), but one GMP is filled by only one GP, that is why the terms of GMP and GP can be used interchangeably in this setting. To explain this speciality a new section was added to the Methods summarizing some characteristics of the Hungarian primary health care system (“Setting”).

5) THE RESULTS ARE VERY DIFFICULT TO INTERPRET WITHOUT HAVING AN UNDERSTANDING OF THE EXISTENCE OF GROUP PRACTICE IN HUNGARIAN PRIMARY CARE. DAHOUGE'S ONTARIO SUDY OF PHYSICIAN GENDER NEEDED TO MAKE CONSIDERABLE ADJUSTMENTS TO INCORPORATE THE COMPLEXITY OF GROUP PRACTICE

SIZE, AND OTHER CHARACTERISTICS. THIS IS IMPORTANT IN THAT I CANT SEE HOW PHYSICIAN GENDER AT THE LEVEL OF THE PRACTICE WAS ASCERTAINED TO ACCOUNT FOR THE DIFFERNG NUMBERS OF PROVIDERS THAT MAY WORK THERE.
Response: The existence of group practices is absolutely not typical in Hungary. GPs are working in solo practices, since each GMP belongs to a GP who is responsible for care in his/her own practice. The analysis was undertaken at the level of the pactice (GMP), but one GMP is filled by only one GP, that is why the terms of GMP and GP can be used interchangeably in this setting. To explain this speciality a new section was added to the Methods summarizing some characteristics of the Hungarian primary health care system ("Setting").

6) THERE IS NO MENTION OF CLUSTER ANALYSIS - I WOULD HAVE THOUGHT THAT THIS NEEDED TO BE ACCOUNTED FOR
Response: The availability of NIHIFM's data are restricted to aggregated data, therefore we used traditional linear regression models. The relationship between GP gender and quality of care was examined using data aggregated at GMP level. GMP and GP characteristics are included in a multivariate model with the GMP as the unit of analysis. Conclusions are also drawn at the GMP level.
The Methods – Patient-level, physician-level and organizational characteristics of GMPs section has been corrected: "Patients data were available in aggregated form. GMP specific data were provided by NIHIFM aggregated for sex and age groups." has been inserted.
The following sentence was added to Methods – Performance indicators section: "Indicators reflect the proportion of patients who recieved the care in each GMPs." has been inserted.
The Methods – Statistical analysis section has been corrected: "Multivariate linear regression models were used to determine the association between the gender of GP and performance in GMP measured by SPRs, where the unit of analysis was the GMP." has been inserted.
The Limitation section was supplemented related to the issue of the analysis: "The major limitation of this study was that the GMPs were used as the unit of analysis, instead of patients. As a consequence of restricted availability of NIHIFM's data, the relationship between GP gender and quality of care could be exclusively examined using aggregated data at GMP level, which approach ignores clustering of patients."

7) IN ADDITION THERE SEEMED TO BE NO CONTROLS FOR THE NUMBER OF HOURS WORKED BY PROVIDERS. FOR EXAMPLE ONE WOULD HAVE EXPECTED THAT A CLINICIAN WITH A PANEL SIZE OF 1500 WHO WAS WORKING 2 DAYS PER WEEK WOULD HAVE A VERY DIFFERENT PATTERN OF PRACTICE THAN ONE WORKING 5 DAYS PER WEEK.
Response: Our study was restricted to take into account the panel size as a potential source of variation. The database does not contain any information about the working hours. Most of the GPs in Hungary work as self-employed workers organizing the provision on their own authority.
The Limitations has been completed with the sentence: "Since, most of the GPs in Hungary are self-employed workers organizing the provision on their own authority, there was not possible control for the length of their working hours."

7) THE CALCULATION OF STANDARDISED PREVENTION RATIOS COULD BE BETTER DESCRIBED - AN EXAMPLE WOULD BE USEFUL HERE.
The Methods section has been supplemented with the description of calculation of standardized prevalence ratios: "To control the effects of the demographic composition of the GMPs, PHC indicators were standardized for the age and gender of the patients. The national reference rates were calculated by gender and age groups, the expected number of cases was the multiplication of the national stratum-specific reference rates and the population of the GMP in the corresponding age and gender group. The sum of the expected numbers in each age and gender category gives the total number of cases that would be expected in the GMP if it had the same age- and gender specific prevalence rates as the country. The ratio of the observed (O) and expected (E) number of patients

was calculated to obtain the age- and gender standardized prevalence ratios (SPRs; O/E x 100) for all GMPs."

8) FINALLY IN TERMS OF PRESENTATION OF THE RESULTS I THINK THAT THE AUTHORS SHOULD BE CAREFUL ABOUT CONFUSING CLINICAL TO STATISTICAL SIGNIFICANCE - ESPECIALLY WITH THE VERY LARGE SAMPLE. I WOULD STRUGGLE TO BE VERY EXCITED ABOU A 1% DIFFERENCE IN PREVENTIVE PERFORMANCE ON INFLUENZA VACCINATION FOR EXAMPLE (P11 LINE 3)
The Results - Descriptive statistics in the main text has been completed by the following sentence: "Although there were statistically significant differences between genders related to vaccination against influenza (p=0.004), the size of difference was negligible (20.08% vs. 19.90%)."
In Result section, "remarkable differences" expression was changed to "statistically significant differences".
The following sentences have been added to the Discussion – Main finding section: "However, careful interpretation is required due to the large sample size. Gender effect on serum creatinine and lipid measurement seems to have a clinical relevance, as by these indicators higher effect of gender was observed compared with other indicators."
The Limitations has been completed with the further issues: "Additionally, the fact that the range of variables in the NIHIFM database which could be included in the analysis is limited, careful interpretation is required" ... "A further issue regarding to the clinical relevance is the large sample size, which can theoretically detect statistically significant but very small effects, with limited clinical/practical relevance."

9) THE DISCUSSION PUTS THE RESULTS INTO THE CONTEXT OF THE BROADER LITERATURE AND DREW A FEW THREADS TOGWTHER CONCERNING THE "MISSED CARE" THAT FOLLOWED AS A CONSEQUENCE OF MALE GENDER. LIMITATIONS WERE REASONABLE, HOWEVER MISSED SEVERAL THAT MAY BE OF SOME RELEVANCE, IN PARTICULAR THE ISSUE OF THE ANALYSIS NOT TAKING ACCOUNT OF GENDER CONCORDANCE, PATIENT COMPLEXITY (MALE GPS TEND TO SEE MORE COMPLEX PATIENTS THAN FEMALES) AND CLUSTERING.
Response: Neither complexity of patients, nor gender concordance cannot be assess in the absence of individual data. However, the association among GP gender and quality of care in male and female patiens was analysed using existing data, and there was no meaningful differences observed in performance related to the gender of patients (except for beta-blocker application, where contrary to our basic analysis significant relationship was observed only in male patients (analytical table is attached: Table to reviewer 1)). Considering the already quite expanded and detailed structure of the manuscript, we didn't want to address this issue in the present study. We would like to measure the complexity and gender concordance in a future study.
The Limitations has been completed with further issues:
"The major limitation of this study was that the GMPs were used as the unit of analysis, instead of patients. As a consequence of restricted availability of NIHIFM's data, the relationship between GP gender and quality of care could be exclusively examined using aggregated data at GMP level, which approach ignores clustering of patients."
"Selection of the GP may be affected by patients preferences and expectations, as patients have free choice of health care provider in Hungary."
The issue of patients complexity was added to the Discussion – Further research need section:
"The onset/duration of a chronic disease and accordingly the complex needs of patients may modify the gender effect.[44] Therefore, it would be worthwhile to investigate the potential influences on performance indicators that are differentiated according to the preferences[45-47] and type of health problem.[48]"

10) THE ARTICLE WOULD BENEFIT FROM MORE CONTEXT ON THE HUNGARIAN HEALTH

CARE SYSTEM - IT WAS IMPOSSIBLE TO ASCERTAIN THE DEGREE TO WHICH THE SYSTEM REQUIRED PATIENTS TO ENROL WITH A GP, THE PREVALENCE OF GROUP PRACTICE AND THE PRESENCE OR ABSENCE OF INCENTIVES FOR PREVENTIVE ACTIVITIES.
A new section was added to the Methods including some characteristics of the Hungarian health care system (“Setting”)
"The health care system in Hungary is based on compulsory health insurance with universal coverage. Primary care services are provided by general practitioners (GPs) working in solo practices, therefore one GMP is owned and operated by one GP. In vacant GMPs the services are provided by temporary GPs with restricted availability in time and place. The GPs are contracted with the NIHIFM, and they have territorial supply obligation (municipalities are responsible for the provision of primary care for the local population within their terrority), but patients can choose and change their primary care provider without any restriction.[37] "
The Background was supplemented: ”... and the provision of financial incentives for GPs are based on reaching desired target values for these indicators.”

THERE ARE SOME OTHER ERRRORS WORTH ADDRESSING -
11) THERE SEEMS A DISCREPANCY IN THE P VALUE BETWEEN THE TEXT AND THE TABLE AS RELATED TO INFLUENZA VACCINATION
Response: The p value in the text related to influenza vaccination was corrected (p=0.004)

12) I FEEL THAT IT IS IMPORTANT TO EXPLAIN THE DEMOGRAPHICS IN TABLE 3 - I AM A LITTLE SURPRISED AS TO WHY THE PREVALENCE OF INDIVIDUALS REQUIRING INFLUENZA VACCINATION WAS LESS THAN THE NUMBERS PRESUMABLY ON ANIHYPERTENSIVES AND REQUIRING A SERUM CREATININE MEASURE.
Response: The Table 3 shows the achievement of GPs in 2016. The target group in case of influenza vaccination is patients over 65 years. In case of influenza vaccination 19.9% means that almost 20% of patients aged 65 or over of male GPs was vaccinated against influenza.
We completed the Table 1 with detailed description of the target groups.
Furthermore, the title of the Table 3 has changed as “Achievement of GPs in 2016 for the whole country by gender of GPs with 95% confidence intervals and P-value”

13) FINALLY, I AM NOT A STATISTICIAN, BUT WAS SURPRISED TO SEE THAT THE AUTHORS ANALYSED THE DATA USING A LINEAR REGRESSION RATHER THAN HAS BEEN DONE IN OTHER PAPERS OF THIS TYPE, A LOGISTIC REGRESSION FOR THE CHRONIC DISEASE INDICATORS (SEE DAHROUGE ET AL 2016 MEDICAL CARE)
Response: The availability of data at individual level is restricted by the NIHIFM, therefore we could not assess the relationship at individual level. Indicators of chronic diseases reflect the proportion of patients who received the care in each GMP (e.g. the number of patients who had had lipid measurement were aggregated at the GMP level and an overall lipid measurement rate was calculated for each GMPs, resulting continuous outcome variable).

14) MINOR ADDITIONAL SUGGESTIONS ARE THAT THE WORK LACKED SPACES OR INDENTS FOR PARAGRAPHS - THIS MADE IT MUCH MORE DIFFICULT TO READ, ALTHOUGH MAY BE A CONSEQUENCE OF THE PROCESS OF PREPARING THE MANUSCRIPT FOR REVIEW.
Response: Lack spaces were added for paragraphs.

Thank you for the careful review and for the suggestions to improve the manuscript!

Response to Reviewer 2:

1) THE AUTHORS DESCRIBE THE OUTCOMES AND THE ANALYSIS VERY CLEARLY.
Response: Thanks for this comment!

2) THE TABLES ARE CLEAR AND SUPPORT THE RESULTS, WITH THE EXCEPTION OF THE SUPPLEMENTARY TABLE, WHICH I FOUND DIFFICULT TO FOLLOW.
Response: Our intention was to present the supplementary data for readers who are interested in details of analyses. We supposed that these readers are able to follow the rather complex content.

HOWEVER, I HAVE A COUPLE OF CONCERNS THAT REQUIRE EITHER CLARIFICATION OR PERHAPS FURTHER ANALYSIS.

3) THE FIRST IS THAT THE DATA APPEARS TO BE HIERARCHICAL - PATIENTS ARE NESTED IN GPS WHO ARE IN TURN NESTED IN GMPS. BUT THE MODEL DOES NOT SEEM TO TAKE THIS INTO ACCOUNT. IT SEEMED TO ME THAT A MULTI-LEVEL MODEL WOULD BE MORE APPROPRIATE HERE. THE OUTCOMES ARE LIKELY TO CLUSTER AT BOTH GP AND PRACTICE LEVEL AND I FEAR THAT FAILING TO ACCOUNT FOR THIS IN THE MODEL COULD LEAD TO INCORRECT CONCLUSIONS.
Response: Patient-level data were not available for us. NIHIFM provided aggregated data for analysis, therefore we used linear regression models with aggregated data. In the absence of individual patient data, the relationship between GP gender and quality of care were examined using data aggregated at the GMP level. GP characteristics are then included in a multivariate model with the GMP as the unit of analysis. Conclusions are also drawn at the GMP level.
The existence of group practices is absolutely not typical in Hungary. GPs are working in solo practices, since each GMP belongs to a GP who is responsible for care in his/her own practice. The relationship between GP gender and quality of care could be exclusively examined using aggregated data at GMP level, which approach ignores clustering of patients.
Chronic disease indicators reflect the proportion of patients who received the care in each GMP. For instance, number of patients who had lipid measurement were aggregated at the GMP level and an overall lipid measurement rate were calculated for each GMP.
The Methods – Patient-level, physician-level and organizational characteristics of GMPs section have been corrected: “Patients data were available in aggregated form. GMP specific data were provided by NIHIFM aggregated for sex and age groups.” has been inserted.
The following sentences were added to the Methods – Performance indicators section: “Indicators reflect the proportion of patients who recieved the care in each GMPs.” has been inserted.
The Methods – Statistical analysis section has been corrected: “Multivariate linear regression models were used to determine the association between the gender of GP and performance in GMP measured by SPRs, where the unit of analysis was the GMP.” has been inserted.
The following sentences were added to the Limitation section: “The major limitation of this study was that the GMPs were used as the unit of analysis, instead of patients. As a consequence of restricted availability of NIHIFM’s data, the relationship between GP gender and quality of care could be exclusively examined using aggregated data at GMP level, which approach ignores clustering of patients.”

4) I WAS ALSO PUZZLED BY THE DECISION TO TURN A NUMBER OF CONTINUOUS VARIABLES INTO CATEGORICAL ONES. DOING THIS LEADS TO A LOSS OF STATISTICAL POWER AND I COULDN'T SEE A CLEAR RATIONALE FOR DOING SO.
Basically, age of the patients and GMP size were available as a categorical variable in the dataset. Age of GP was originally continuous variable. The reason of the dichotomization was that a significant part of GPs works over 65 years, which is the retirement age in Hungary. The retirement age could be identified as a risk factor for lower quality of care.
The following sentence has been added to the Methods section: “The created age groups were 44 years or younger, 45-64 years and 65 years and above (as a significant part of GPs works over 65

years, which is the retirement age in Hungary and could be identified as a potential risk factor for lower quality of care).”

5) IT WOULD ALSO BE HELPFUL TO UNDERSTAND WHY THE AUTHORS ARE REPORTING THE STANDARDISED REGRESSION COEFFICIENTS. MOST READERS WILL BE MORE FAMILIAR WITH THE UNDERSTANDARDISED COEFFICIENTS SO THE AUTHORS SHOULD EXPLAIN THIS CHOICE AND GIVE SOME INTERPRETATION OF THE COEFFICIENTS AS IT IS NOT STRAIGHTFORWARD, PARTICULARLY IN THE PRESENCE OF MULTI-LEVEL DATA.
The following sentence has been added to the Methods section: “Standardized linear regression coefficients (beta) were calculated to present the GP-gender effect in a way that ensures its comparability to other explanatory variables’ effects.”
The following sentence was added to the Results - Multivariate analyses section: “Gender had one of the strongest effects among all studied variables on hypertension and diabetes care indicators (serum creatinine, lipid and HbA1c measurement, as well as eye examination).”
In addition, the Table 4 was supplemented with the understandardised linear regression coefficients (b).

6) SOME OF THE DIFFERENCES THAT THE AUTHORS COMMENT UPON IN THE RESULTS SECTION ARE STATISTICALLY SIGNIFICANT. HOWEVER, I WOULD DISAGREE THAT THE RESULTS SHOWN IN TABLE 2 ARE "REMARKABLE DIFFERENCES". MANY ARE SMALL AND THEIR STATISTICAL SIGNIFICANCE MAY INSTEAD RELATE TO THE LARGE SAMPLE. ALL ARE LESS THAN 5%. EVEN IN THE MULTIVARIATE ANALYSIS, THE COEFFICIENTS AND R-SQUARED VALUES ARE NOT LARGE EVEN IF THEY ARE SIGNIFICANT. IT IS UNCLEAR WHETHER THIS IS A CLINICALLY MEANINGFUL DIFFERENCE.
In Result section, “remarkable differences” expression was changed to “statistically significant differences”.
The following sentences have been added to the Discussion – Main finding section: “However, careful interpretation is required due to the large sample size. Gender effect on serum creatinine and lipid measurement seems to have a clinical relevance, as by these indicators higher effect of gender was observed compared with other indicators.”
The Limitations has been completed with further issues: “A further issue regarding to clinical relevance is the large sample size, which can theoretically detect statistically significant but very small effects, with limited clinical/practical relevance.”

7) THE AUTHORS DO MENTION THE POSSIBILITY OF UNMEASURED CONFOUNDING IN THEIR LIMITATIONS SECTION. I THINK THIS NEEDS A BIT MORE EMPHASIS IN THE DISCUSSION. THERE ARE A LIMITED NUMBER OF VARIABLES IN THIS MODEL, PARTICULARLY AT THE PATIENT LEVEL, WHICH MIGHT ACCOUNT FOR SOME OR ALL OF THE OBSERVED ASSOCIATION.
Response: As the reviewer-2 wrote the model we tested is rather simple considering the complexity of the indicators we studied. Further, it is mentioned in the text that further research is needed to explore the mechanisms responsible for the GP-gender effect. Actually, the paper was to demonstrate that GMP-indicators show dependency on GP-gender which is not neglectable and worth to investigate in details.
The Limitation section has been completed with further issues:
“Additionally, the fact that the range of variables in the NIHIFM database which could be included in the analysis is limited, careful interpretation is required.”

8) ANOTHER ISSUE MAY BE THE SAMPLE ITSELF. THE AUTHORS MENTION IN THE BACKGROUND THAT 56.2% OF PHYSICIANS IN HUNGARY ARE FEMALE. BUT IN THIS SAMPLE, THE PROPORTION IS 48.37%. IT WOULD BE HELPFUL TO EXPLAIN THIS DISCREPANCY AND TO COMMENT ON WHETHER THIS COULD BE DUE TO SELECTION BIAS.

Response: The proportion of female physicians in the Background (56.2%) reflects the total share of female physicians (including pediatricians). Our study is exclusively focusing on the GPs providing care for adults (N=4575).
The Background was corrected as follows: The country has one of the highest shares of female physicians of general medicine and paediatrics, with a total share of 55.9% in 2015[1] and 56.2% in 2017[35].
The Results – Descriptive statistics was corrected as follows: The studied 4575 GPs (providing care for adults) consisted of 2213 (48.37%) female, and 2362 (51.63%) male (p<0.001).

VERSION 2 – REVIEW

REVIEWER	Grant Russell Monash University Victoria Australia
REVIEW RETURNED	22-Jan-2019

GENERAL COMMENTS	I think that the authors have done an excellent job in addressing the issues raised by the reviewers - it is a complex area, we asked targeted questions and they were all answered thoughtfully and respectfully. I really have nothing to add other than congratulations. Thank you and good luck

REVIEWER	Dr Kathryn Taylor University of Oxford UK
REVIEW RETURNED	24-Apr-2019

GENERAL COMMENTS	Figure 1 is missing. General point It's confusing reading about GPs and the performance of GMPs. As the practices only involve a single GP, the performance of a particular GMP is the performance of its GP. Abstract There are too many abbreviations in the abstract (two are only used once). Introduction Introduction, 5th paragraph refers to “territorial supply obligation”. This phrase is defined later in Methods (1st paragraph) – the definition should be moved to the introduction. Methods Study design – “Vacant GMPs without [insert "a"] permanent GP…” Setting – have already defined GP abbreviation in the introduction. The description of age is confusing. The age categories of the data provided by NIHIFM and the categories used in the analysis

should be more clearly stated. Does Methods (6th paragraph) refer to NIHFM categories being merged into fewer categories (<44, 45-64, 65+) for the analysis. This refers to the ages of patients but Supplementary Table 1 refers to ages of GPs using the same categories??

Methods needs to clearly state all the reference categories and it would also be useful to clarify the reference categories again in the footnote to Supplementary Table 1. Budapest is the reference category for county of GMP but the Methods does not make this clear.

I don’t find relative education convincing as a proxy for socioeconomic status given that this is usually based on education, income and employment. Relative education seems be simply an indicator of educational attainment.

Attributable risk should be relabelled attributable proportion – there is no risk of a disease or medical condition.

“The differences between the summarized observed and expected number of care events were calculated for GMPs with male GPs in order to quantify the excess number of care events”. “Excess” has negative connotations. It would be better to say something like “to quantify the differences in the number of care events compared to female GPs”.

“P-levels below 0.05” should be “p<0.05”.

The definition given for the adjusted-R squared is the definition for R-squared. The adjusted R-squared quantifies the variation in the dependent variable that is explained by only the explanatory variables that affect the dependent variable, and penalises for the inclusion of poor predictors in the model.

Need to describe what was done to test that the assumptions of regression are met.

“Multivariate linear regression models were used to determine the association between the gender of [insert "the"] GP and [insert "their"] performance [delete "in GMP"] [insert "as"] measured by SPRs

Results
Patients of female physicians were more likely [insert "to be"] female..”

“According to the crude values[should be plural] of the PHC indicators..”

Descriptive statistics and Table 2 – percentages should only be given to 1 decimal place.

Remove spaces in numbers e.g. 4 567 should be 4567.

Discussion
“careful interpretation is required due to the large sample size” Why? Large sample sizes provide more accurate estimates than small sample sizes and provide the ability to detect small effects

	It would be helpful to have a subheading “Comparison with other studies” “The size of this gender effect proved to be unneglectable” – Do you mean that the gender effect was sizable/notable? “[insert "A" strength of this study… ….. gender effect, which has rarely been investigated [delete "due to" insert "as reflected by"] a scarcity of relevant publications.” (Scarcity of evidence is a reason to find evidence) “As a consequence of restricted availability of NIHIFM’s data, the relationship between GP gender and quality of care could be exclusively examined using aggregated data at GMP level, which approach ignores clustering of patients.” It’s not clear what this means. The NIHIFM could only provide aggregate data at GMP level? Can you suggest examples where clustering might be an issue. “A further issue regarding to clinical relevance is the large sample size, which can theoretically detect statistically significant but very small effects, with limited clinical/practical relevance.” This is another criticism of the large sample size. See my comments above. Clinical significance and statistical significance are different issues. A statistically significant difference simply means that it was unlikely to have occurred by chance. It doesn’t necessarily measure the clinical significance which is about the clinical importance of this difference for patient care. The authors should be considering the ”statistically significant vs clinically significant” issue when interpreting their findings by highlighting that some gender differences that they found were statistically significant but not clinically significant. Need to acknowledge the limitation in that the analysis was based on data on solo GP practices, which are specific to the Hungarian primary health care system and other solo-provider primary health care systems. Further research need The gender differences were only statistically and clinically significant for serum creatinine and lipid measurement. In referring to the “strong association” it would be better to be more specific. Implications The scarcity of resources in Hungary is not relevant to the discussion about changes to the medical school curriculum and changes to the attitudes of GPs unless these require greater investment, and in that case, the resource implications of the changes need to be mentioned afterwards. Conclusion “Provision of guideline-recommended care, which is expressed in the higher proportion of care events, was observed more often in patients of female GPs.” Higher proportion of what? The sentence would read more clearly as “Provision of guideline-recommended care was observed more often in patients of female GPs.” Table 2 “Number of GMPs’ patients”

	Age of GP "x-44 years" should be "less than 44 years" or "≤ 44 years" Remove spaces in numbers e.g. 4 567 should be 4567. Table 3 Need to state in the title what is meant by achievement. Don't need to state the p-value in the title. Supplementary Table 1 There is no need to list the counties in the footnote.

REVIEWER	Nicholas Moloci Dow Division of Health Services Research, Department of Urology, University of Michigan Medical School, Ann Arbor, Michigan, United States of America
REVIEW RETURNED	01-May-2019

GENERAL COMMENTS	This paper looks to investigate the effects of a primary care physician gender and the quality of care being provided. The authors use one year of data from a national database, examining Hungarian primary care physicians. While the question that the authors are trying to assess is an important one, there are a few ways that the authors can help improve the analysis they've run. The authors need to add in a fixed effect for each primary care physician within their study. The care that is being provided by each physician is going to be different from that of their peers. This needs to be accounted for. Are there other GP practice effects that you can account for? Things like the number and quality of staff? How are GPs who practice in multiple locations being accounted for? As far goes the age categories, I would advise against further categorization for the patients ages and run the regression using the data as given. I understand that the GP age is a variable of interest to you, but you may also wish to run the analysis with the GP age as a continuous variable. Due to the clustering of your data and the heteroscedasticity of the data, I would advise the authors to run the analysis using robust standard errors. This will help the authors obtain unbiased standard errors. In either an appendix or additional tables, please present regression results. It made it challenging to understand what variables were in each model being run. Please justify how the physician practice sizes were being categorized. Are you able to establish how long the provider and patient have been seeing each other? Things like physician patient relationship have an important impact on how the provider provides care to the patient.

	Please correct multivariate to multivariable regression. How are missed care events being defined? Is this just that the patient didn't receive the care or how do you know that the patient wasn't advised to do so, and then refused? Or are the missed care events just not meeting the number of expected events from the regression? I'm hesitant to state that provider gender alone can explain some of the differences of care, especially when you're only able to explain 5-27% of the variability for each condition. Further, this study only looks at one year of care being provided. Many patients may not seek the care being advised to them right away, thus not showing up in the single year of data. However, your findings do show that this may be an area of interesting research to be investigated further.

REVIEWER	Pietro Manuel Ferraro Fondazione Policlinico Universitario A. Gemelli IRCCS, Roma, Italia
REVIEW RETURNED	09-May-2019

GENERAL COMMENTS	- The source of performance indicators is not clear. Were the data obtained directly from the GPs? - Are the indicators validated/established parameters within the Hungarian health system or were they devised ad hoc for this study? - The baseline characteristics of female-led GMPs are systematically and consistently different compared with male-led GMPs. For all, male-led GMPs are consistently larger, which is likely a major determinant of whether adequate care is given to the individual patient. Although the statistical analysis is adjusted for such differences, it is still indicative of the possibility that residual confounding or systematic bias not otherwise measurable could explain the differences in sex - The outcome variable is a proportion (number of patients receiving a given intervention divided by the number of patients eligible for that intervention). The use of a different regression model such as negative binomial or Poisson would be preferable compared with linear regression. On top of a potentially better fit, the use of such models would allow the authors to obtain more understandable estimates in this context rather than betas

VERSION 2 – AUTHOR RESPONSE

RESPONSE TO REVIEWER 1:

1) I THINK THAT THE AUTHORS HAVE DONE AN EXCELLENT JOB IN ADDRESSING THE ISSUES RAISED BY THE REVIEWERS - IT IS A COMPLEX AREA, WE ASKED TARGETED QUESTIONS AND THEY WERE ALL ANSWERED THOUGHTFULLY AND RESPECTFULLY.
I REALLY HAVE NOTHING TO ADD OTHER THAN CONGRATULATIONS. THANK YOU AND GOOD LUCK

Thank you!

RESPONSE TO REVIEWER 3:

1) FIGURE 1 IS MISSING.
Response: We are sorry for any inconvenience. Although Figure 1 was uploaded as part of the previous version, we have uploaded the Figure 1 again (as a separate file).

2) GENERAL POINT
IT'S CONFUSING READING ABOUT GPS AND THE PERFORMANCE OF GMPS. AS THE PRACTICES ONLY INVOLVE A SINGLE GP, THE PERFORMANCE OF A PARTICULAR GMP IS THE PERFORMANCE OF ITS GP.
Response: The confusing ”performance of GMPs” expression has been changed to „performance of GPs”.

3) ABSTRACT
THERE ARE TOO MANY ABBREVIATIONS IN THE ABSTRACT (TWO ARE ONLY USED ONCE).
Response: Thank you for your comment, the abbreviations have been expanded.

4) INTRODUCTION
INTRODUCTION, 5TH PARAGRAPH REFERS TO “TERRITORIAL SUPPLY OBLIGATION”. THIS PHRASE IS DEFINED LATER IN METHODS (1ST PARAGRAPH) – THE DEFINITION SHOULD BE MOVED TO THE INTRODUCTION.
Response: The definition has been moved to the Introduction.

5) METHODS
STUDY DESIGN – “VACANT GMPS WITHOUT [INSERT "A"] PERMANENT GP…”
Response: This has been corrected.

6) SETTING – HAVE ALREADY DEFINED GP ABBREVIATION IN THE INTRODUCTION.
Response: The GP abbreviation has been removed.

7) THE DESCRIPTION OF AGE IS CONFUSING. THE AGE CATEGORIES OF THE DATA PROVIDED BY NIHIFM AND THE CATEGORIES USED IN THE ANALYSIS SHOULD BE MORE CLEARLY STATED. DOES METHODS (6TH PARAGRAPH) REFER TO NIHFM CATEGORIES BEING MERGED INTO FEWER CATEGORIES (<44, 45-64, 65+) FOR THE ANALYSIS. THIS REFERS TO THE AGES OF PATIENTS BUT SUPPLEMENTARY TABLE 1 REFERS TO AGES OF GPS USING THE SAME CATEGORIES??
Response: Due to reviewers' requests the entire data analysis has been revised.
Accordingly, the following age categories were used:
The data of patients were provided by NIHIFM in age groups of 18-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59, 60-64, 65-69, 70-74, 75-79, 80-84, 85-89, and 90 years and above. The age categories of patients which were used in the analysis was the same as provided by NIHIFM.
The age of the GP provided by NIHIFM considered continuous variable. In the current analysis we used years of age instead of age groups. To clarify this issue, the following sentence has been added to the Methods (Patient-level, physician-level and organizational characteristics of GMPs) section:
“The age of the GP was continuous variable.”
The 4th paragraph in the Methods section has been modified: “Patients and GMP specific data were provided by NIHIFM. The number of adults registered in each GMP was determined by gender and age groups of 18-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59, 60-64, 65-69, 70-74, 75-79, 80-84, 85-89, and 90 years and above in the database. The age group of 65-69 (and 60-64 where 65 years and above patients were not available) was used as reference. The age of the GP was continuous variable. The male gender of GPs and patients were used as reference in the analyses.”

8) METHODS NEEDS TO CLEARLY STATE ALL THE REFERENCE CATEGORIES AND IT WOULD ALSO BE USEFUL TO CLARIFY THE REFERENCE CATEGORIES AGAIN IN THE FOOTNOTE TO SUPPLEMENTARY TABLE 1. BUDAPEST IS THE REFERENCE CATEGORY FOR COUNTY OF GMP BUT THE METHODS DOES NOT MAKE THIS CLEAR.
Response: The Methods section and the Supplementary tables have been completed with the reference categories used in the analysis. The following sentences have been added to the Methods: „The male gender of GPs and patients were used as reference.", "The age groups of 65-69 (and 60-64 where 65 years and above patients were not available) was used as reference.", "The practice size of 1201-1600, the rural settlement and Budapest were used as reference categories."

9) I DON'T FIND RELATIVE EDUCATION CONVINCING AS A PROXY FOR SOCIOECONOMIC STATUS GIVEN THAT THIS IS USUALLY BASED ON EDUCATION, INCOME AND EMPLOYMENT. RELATIVE EDUCATION SEEMS BE SIMPLY AN INDICATOR OF EDUCATIONAL ATTAINMENT.
Response: The socioeconomic status was replaced by educational attainment in the Methods section: „The relative education was used to indicate the educational attainment of adults registered in each GMP."

10) ATTRIBUTABLE RISK SHOULD BE RELABELLED ATTRIBUTABLE PROPORTION – THERE IS NO RISK OF A DISEASE OR MEDICAL CONDITION.
Response: The attributable risk has been relabelled as attributable proportion.

11) "THE DIFFERENCES BETWEEN THE SUMMARIZED OBSERVED AND EXPECTED NUMBER OF CARE EVENTS WERE CALCULATED FOR GMPS WITH MALE GPS IN ORDER TO QUANTIFY THE EXCESS NUMBER OF CARE EVENTS". "EXCESS" HAS NEGATIVE CONNOTATIONS. IT WOULD BE BETTER TO SAY SOMETHING LIKE "TO QUANTIFY THE DIFFERENCES IN THE NUMBER OF CARE EVENTS COMPARED TO FEMALE GPS".
Response: The original sentence has been rephrased to „The differences between the summarized observed and expected number of care events were calculated for GMPs with male GPs in order to quantify the differences in the number of care events compared to female GPs."

12) "P-LEVELS BELOW 0.05" SHOULD BE "P<0.05".
Response: Done.

13) THE DEFINITION GIVEN FOR THE ADJUSTED-R SQUARED IS THE DEFINITION FOR R-SQUARED. THE ADJUSTED R-SQUARED QUANTIFIES THE VARIATION IN THE DEPENDENT VARIABLE THAT IS EXPLAINED BY ONLY THE EXPLANATORY VARIABLES THAT AFFECT THE DEPENDENT VARIABLE, AND PENALISES FOR THE INCLUSION OF POOR PREDICTORS IN THE MODEL.
Response: The statistical analysis has been revised. We used multilevel logistic regression models, and the proportion of total variance explained by physician groups were presented using the intraclass correlation coefficient (ICC) instead of adjusted-R squares. The following sentence has been added to the Methods section: „We presented intra-class correlation coefficient (ICC), which shows the proportion of total variance explained by the physician as grouping factor."
The Discussion (Comparison with other studies) has been completed with the results related to ICC: „Besides, our findings are consistent with other studies found that relatively low percentage of the variance attributable to physicians on process measures after adjusting for characteristics of the physician, patient and practice. [44,45]"

14) NEED TO DESCRIBE WHAT WAS DONE TO TEST THAT THE ASSUMPTIONS OF REGRESSION ARE MET.
Response: The statistical analysis has been revised. The following sentence has been added to the Methods section: "We used multilevel logistic regression models because of the binary outcomes and

in order to account for the clustering effect of patients within physicians.”
Furthermore, the Hosmer-Lemeshow test was used to determine the goodness of fit of the logistic regression model. Due to the extreme sensitivity of the H-L test in very large samples, the model calibration was assessed graphically by comparing predicted and observed outcomes in deciles of predicted risk. The results of test in each indicator are attached as ”Additional data to Reviewer 3.” Accordingly, the Methods has been completed with the following sentence: „The Hosmer-Lemeshow test was used to determine the goodness of fit of the model (by comparing predicted and observed outcomes in deciles of predicted risk), which verified the adequacy of the model.”

15) “MULTIVARIATE LINEAR REGRESSION MODELS WERE USED TO DETERMINE THE ASSOCIATION BETWEEN THE GENDER OF [INSERT "THE"] GP AND [INSERT "THEIR"] PERFORMANCE [DELETE "IN GMP"] [INSERT "AS"] MEASURED BY SPRS
Response: The sentence has been replaced by „We used multilevel logistic regression models because of the binary outcomes and in order to account for the clustering effect of patients within physicians. The models were adjusted for characteristics of patients (age), GPs (gender, age) and GMPs (practice size, types of settlement, geographical location and relative education) to assess the effect of the gender of the GP on their performance in case of each indicators.”

16) RESULTS
PATIENTS OF FEMALE PHYSICIANS WERE MORE LIKELY [INSERT "TO BE"] FEMALE..”
Response: Corrected.

17) “ACCORDING TO THE CRUDE VALUES[SHOULD BE PLURAL] OF THE PHC INDICATORS..”
Response: Corrected.

18) DESCRIPTIVE STATISTICS AND TABLE 2 – PERCENTAGES SHOULD ONLY BE GIVEN TO 1 DECIMAL PLACE.
Response: This has been corrected. We would like underline that the Table 2 has been replaced by a new, more complex table which includes the age groups of the patients as well.

19) REMOVE SPACES IN NUMBERS E.G. 4 567 SHOULD BE 4567.
Response: This has been corrected.

DISCUSSION
20) “CAREFUL INTERPRETATION IS REQUIRED DUE TO THE LARGE SAMPLE SIZE” WHY? LARGE SAMPLE SIZES PROVIDE MORE ACCURATE ESTIMATES THAN SMALL SAMPLE SIZES AND PROVIDE THE ABILITY TO DETECT SMALL EFFECTS
Response: It has been rephrased and moved to the Limitation section: „Although large sample size can detect even the smallest differences, the interpretation needs to consider that this small differences may have little importance at population level.”

21) IT WOULD BE HELPFUL TO HAVE A SUBHEADING “COMPARISON WITH OTHER STUDIES”
Response: The Discussion has been completed with a subheading of „Comparison with other studies”.

22) “THE SIZE OF THIS GENDER EFFECT PROVED TO BE UNNEGLECTABLE” – DO YOU MEAN THAT THE GENDER EFFECT WAS SIZABLE/NOTABLE?
Response: The „unneglectable” word has been replaced „notable”.

23) “[INSERT "A" STRENGTH OF THIS STUDY…
Response: This has been corrected.

24) ….. GENDER EFFECT, WHICH HAS RARELY BEEN INVESTIGATED [DELETE "DUE TO" INSERT "AS REFLECTED BY"] A SCARCITY OF RELEVANT PUBLICATIONS.” (SCARCITY OF EVIDENCE IS A REASON TO FIND EVIDENCE)
Response: This has been corrected.

25) “AS A CONSEQUENCE OF RESTRICTED AVAILABILITY OF NIHIFM’S DATA, THE RELATIONSHIP BETWEEN GP GENDER AND QUALITY OF CARE COULD BE EXCLUSIVELY EXAMINED USING AGGREGATED DATA AT GMP LEVEL, WHICH APPROACH IGNORES CLUSTERING OF PATIENTS.” IT’S NOT CLEAR WHAT THIS MEANS. THE NIHIFM COULD ONLY PROVIDE AGGREGATE DATA AT GMP LEVEL? CAN YOU SUGGEST EXAMPLES WHERE CLUSTERING MIGHT BE AN ISSUE.
Response: The statistical analysis has been revised.
NIHIFM provides data aggregated by the gender and age groups of patients. We disaggregated the data, and used multilevel analysis. Because of the hierarchical structure of our data where patients are nested in GPs, adjustment of the logistic model for the clustering effect was necessary. Therefore, we believe this is more appropriate approach to handle clustered data and helps to prevent incorrect or potentially misleading results.
Regarding to the revision of the statistical analysis, the following paragraph has been deleted from the Strength and limitations section: ”The major limitation of this study was that the GMPs were used as the unit of analysis, instead of patients. As a consequence of restricted availability of NIHIFM’s data, the relationship between GP gender and quality of care could be exclusively examined using aggregated data at GMP level, which approach ignores clustering of patients.”

26) “A FURTHER ISSUE REGARDING TO CLINICAL RELEVANCE IS THE LARGE SAMPLE SIZE, WHICH CAN THEORETICALLY DETECT STATISTICALLY SIGNIFICANT BUT VERY SMALL EFFECTS, WITH LIMITED CLINICAL/PRACTICAL RELEVANCE.” THIS IS ANOTHER CRITICISM OF THE LARGE SAMPLE SIZE. SEE MY COMMENTS ABOVE. CLINICAL SIGNIFICANCE AND STATISTICAL SIGNIFICANCE ARE DIFFERENT ISSUES. A STATISTICALLY SIGNIFICANT DIFFERENCE SIMPLY MEANS THAT IT WAS UNLIKELY TO HAVE OCCURRED BY CHANCE. IT DOESN’T NECESSARILY MEASURE THE CLINICAL SIGNIFICANCE WHICH IS ABOUT THE CLINICAL IMPORTANCE OF THIS DIFFERENCE FOR PATIENT CARE. THE AUTHORS SHOULD BE CONSIDERING THE ”STATISTICALLY SIGNIFICANT VS CLINICALLY SIGNIFICANT” ISSUE WHEN INTERPRETING THEIR FINDINGS BY HIGHLIGHTING THAT SOME GENDER DIFFERENCES THAT THEY FOUND WERE STATISTICALLY SIGNIFICANT BUT NOT CLINICALLY SIGNIFICANT.
Response: The clinically significant differences have been highlighted in the Main findings: ”Gender effect seems to have a clinical relevance mostly on hypertension and diabetes care related indicators, considering both the higher effect of the GP's gender (HbA1c measurement: OR=1.18, 95%CI [1.14-1.23], serum creatinine: OR=1.14, 95%CI [1.12-1.17] and lipid measurement: OR=1.14, 95%CI [1.11-1.16]) and size of the affected population compared with other indicators (mammography screening, eye examination, management of COPD) where we also found statistically significant differences.”
The following sentence has been deleted from the Limitations: „A further issue regarding to clinical relevance is the large sample size, which can theoretically detect statistically significant but very small effects, with limited clinical/practical relevance.”

27) NEED TO ACKNOWLEDGE THE LIMITATION IN THAT THE ANALYSIS WAS BASED ON DATA ON SOLO GP PRACTICES, WHICH ARE SPECIFIC TO THE HUNGARIAN PRIMARY HEALTH CARE SYSTEM AND OTHER SOLO-PROVIDER PRIMARY HEALTH CARE SYSTEMS.
Response: The following sentence was added to the Limitation section: „However the gender of GP is a significant predictor of receiving guideline-recommended care according to our results, it needs to be account that the analysis was based on data on solo GP practices which are specific to the

Hungarian primary health care system and other solo-provider primary health care systems."

FURTHER RESEARCH NEED
28) THE GENDER DIFFERENCES WERE ONLY STATISTICALLY AND CLINICALLY SIGNIFICANT FOR SERUM CREATININE AND LIPID MEASUREMENT. IN REFERRING TO THE "STRONG ASSOCIATION" IT WOULD BE BETTER TO BE MORE SPECIFIC.
Response: You are absolutely right. Remarkable differences were detected in case of HbA1c measurement, serum-creatinine and lipid measurement as it was mentioned in the Main findings. We have been corrected the sentence in the Further research needs, as follows: „Our findings on impact of GP gender (mainly for HbA1c, serum creatinine and lipid measurement) suggest that further consideration of the effect is needed to identify the details and mechanisms behind the gender effect in order to improve the adequacy of targeted interventions."

IMPLICATIONS
29) THE SCARCITY OF RESOURCES IN HUNGARY IS NOT RELEVANT TO THE DISCUSSION ABOUT CHANGES TO THE MEDICAL SCHOOL CURRICULUM AND CHANGES TO THE ATTITUDES OF GPS UNLESS THESE REQUIRE GREATER INVESTMENT, AND IN THAT CASE, THE RESOURCE IMPLICATIONS OF THE CHANGES NEED TO BE MENTIONED AFTERWARDS.
Response: The „where resources are truly scarce" part in the above mentioned paragraph has been deleted.

CONCLUSION
30) "PROVISION OF GUIDELINE-RECOMMENDED CARE, WHICH IS EXPRESSED IN THE HIGHER PROPORTION OF CARE EVENTS, WAS OBSERVED MORE OFTEN IN PATIENTS OF FEMALE GPS." HIGHER PROPORTION OF WHAT? THE SENTENCE WOULD READ MORE CLEARLY AS "PROVISION OF GUIDELINE-RECOMMENDED CARE WAS OBSERVED MORE OFTEN IN PATIENTS OF FEMALE GPS."
Response: The sentence has been corrected: „Provision of guideline-recommended was observed more often in patients of female GPs."

TABLE 2
31) "NUMBER OF GMPS' PATIENTS"
Response: Corrected.

32) AGE OF GP "X-44 YEARS" SHOULD BE "LESS THAN 44 YEARS" OR "≤ 44 YEARS"
Response: The statistical analysis has been revised, and according to the request of the Reviewer 4 the age of the GP is used in the analysis as continuous variable.

33) REMOVE SPACES IN NUMBERS E.G. 4 567 SHOULD BE 4567.
Response: This has been corrected.

TABLE 3
34) NEED TO STATE IN THE TITLE WHAT IS MEANT BY ACHIEVEMENT. DON'T NEED TO STATE THE P-VALUE IN THE TITLE.
Response: The p-value has been deleted from the title. The title has been rephrased, as follow: "The number of patients received the care, the number of people in the target groups, and the proportion of patients received the care in 2016 for the whole country by gender of GPs with 95% confidence intervals."

SUPPLEMENTARY TABLE 1
35) THERE IS NO NEED TO LIST THE COUNTIES IN THE FOOTNOTE.
Response: The list of the counties has been removed. Furthermore, the Supplementary table has

been modified based on the new analysis.

Thank you for the careful review and for the suggestions to improve the manuscript!

RESPONSE TO REVIEWER 4:

Please leave your comments for the authors below
1) THIS PAPER LOOKS TO INVESTIGATE THE EFFECTS OF A PRIMARY CARE PHYSICIAN GENDER AND THE QUALITY OF CARE BEING PROVIDED. THE AUTHORS USE ONE YEAR OF DATA FROM A NATIONAL DATABASE, EXAMINING HUNGARIAN PRIMARY CARE PHYSICIANS. WHILE THE QUESTION THAT THE AUTHORS ARE TRYING TO ASSESS IS AN IMPORTANT ONE, THERE ARE A FEW WAYS THAT THE AUTHORS CAN HELP IMPROVE THE ANALYSIS THEY'VE RUN.

2) THE AUTHORS NEED TO ADD IN A FIXED EFFECT FOR EACH PRIMARY CARE PHYSICIAN WITHIN THEIR STUDY. THE CARE THAT IS BEING PROVIDED BY EACH PHYSICIAN IS GOING TO BE DIFFERENT FROM THAT OF THEIR PEERS. THIS NEEDS TO BE ACCOUNTED FOR.
Response: Thank you for your comments! Considering your advice regarding to analysis, we have consulted with a statistician. As a result, we applied another (hierarchical) data structure which allows to fit multilevel analysis on our data. Consequently, the statistical analyses were revised. After „disaggregation" of the data, we used multilevel logistic regression models because of the binary outcome variable, and in order to account for the clustering effect of patients within physicians. The random-effect was used instead of fixed-effect, because of the invariant property of the GP and the practice. Consequently, effects of these variables could not estimated in a physician fixed effects model. Moreover, we had wanted to estimate the total variation explained by the physician groups by calculating the intraclass correlation coefficient (ICC).
Accordingly, the Statistical analysis section in Methods has been corrected: „We used multilevel logistic regression models because of the binary outcome variables and in order to account for the clustering effect of patients within physicians. The models were adjusted for characteristics of patients (age), GPs (gender, age) and GMPs (practice size, types of settlement, geographical location and relative education) to assess the effect of the gender of the GP on their performance in case of each indicator. Odds ratios (ORs) with the corresponding 95% confidence intervals (CIs), and robust standard errors were estimated. We presented intra-class correlation coefficient (ICC), which shows the proportion of total variance explained by the physician as grouping factor.
The Results section has also been corrected: "The female gender of GPs was associated with hypertension and diabetes care related indicators (HbA1c measurement: OR=1.18, 95%CI [1.14-1.23]; serum creatinine measurement: OR=1.14, 95%CI [1.12-1.17]; lipid measurement: OR=1.14, 95%CI [1.11-1.16]; eye examination: OR=1.06, 95%CI [1.03-1.08]), mammography screening (OR=1.05, 95%CI [1.03-1.08]), management of COPD patients (OR=1.05, 95%CI [1.01-1.09]), and the composite indicator (OR=1.08, 95%CI [1.07-1.1])."
The Table 4 has been corrected.

3) ARE THERE OTHER GP PRACTICE EFFECTS THAT YOU CAN ACCOUNT FOR? THINGS LIKE THE NUMBER AND QUALITY OF STAFF?
Response: Considering the limited data availibility regarding to primary care practices in Hungary, there were no more available variables which we could have accounted for. The lack of other GP practice effects has been indicated in the Limiation section „ ... range of variables in the NIHIFM database which could be included in the analysis is limited..." "Although the analyses were controlled for GPs' and patients' age and gender, education-indicated socio-economic status, GMP practice size, types of settlement and regional location, there were confounding factors that were not included in our models, limiting the reliability of the presented risk measures."

4) HOW ARE GPS WHO PRACTICE IN MULTIPLE LOCATIONS BEING ACCOUNTED FOR?
Response: In Hungary, GPs are working in solo practices, so one GMP is owned and operated by only one physician. Consequently, the characteristics of the GP (e.g. the age and the gender) can be assigned to the given GMP. The subject of the study was the GMP, so if a GP was contracted to more than one GMP, the age and gender of the GP has been taken into consideration in each GMP.

5) AS FAR GOES THE AGE CATEGORIES, I WOULD ADVISE AGAINST FURTHER CATEGORIZATION FOR THE PATIENTS AGES AND RUN THE REGRESSION USING THE DATA AS GIVEN. I UNDERSTAND THAT THE GP AGE IS A VARIABLE OF INTEREST TO YOU, BUT YOU MAY ALSO WISH TO RUN THE ANALYSIS WITH THE GP AGE AS A CONTINUOUS VARIABLE.
Response: The analysis run using the age of the GP as a continous variable (instead of the age categories, which we used in the previous version). The age of patients were provided by the NIHIFM as categorical variable, therefore we used age groups of patients in the analysis. The following sentences has been added to the Methods section: „The age of the GP was continuous variable."

6) DUE TO THE CLUSTERING OF YOUR DATA AND THE HETEROSCEDASTICITY OF THE DATA, I WOULD ADVISE THE AUTHORS TO RUN THE ANALYSIS USING ROBUST STANDARD ERRORS. THIS WILL HELP THE AUTHORS OBTAIN UNBIASED STANDARD ERRORS.
Response: The statistical analysis has been revised. The Table 4 and the Supplementary tables have been completed with robust standard errors.

7) IN EITHER AN APPENDIX OR ADDITIONAL TABLES, PLEASE PRESENT REGRESSION RESULTS. IT MADE IT CHALLENGING TO UNDERSTAND WHAT VARIABLES WERE IN EACH MODEL BEING RUN.
Response: The Supplementary table 1-2 present the results of regression models with the whole range of variables.

8) PLEASE JUSTIFY HOW THE PHYSICIAN PRACTICE SIZES WERE BEING CATEGORIZED.
Response: The practice size was categorized by the NIHIFM based on the number of adults in each practices. The average number of patients in practices is 1600 in Hungary.
The following additional information has been added to the part in the Methods section which is listing the practice size categories: „the categories are defined by NIHIFM."

9) ARE YOU ABLE TO ESTABLISH HOW LONG THE PROVIDER AND PATIENT HAVE BEEN SEEING EACH OTHER? THINGS LIKE PHYSICIAN PATIENT RELATIONSHIP HAVE AN IMPORTANT IMPACT ON HOW THE PROVIDER PROVIDES CARE TO THE PATIENT.
Response: Unfortunately, the availability of practice-level data in primary care is limited in Hungary, due to lack of information, we are not able to adjust the models for the consultation length.
The Limitation has been completed with the above mentioned issue: „Since most of the GPs in Hungary are self-employed workers organizing the provision on their own authority, neither the length of their working hours nor the length of the consultation time were mesurable."
In any case, the average GP consultation length was 5.5 minutes in Hungary in 2018 according to the NIHIFM. This physician consultation time is relatively short similarly to other low- and middle-income countries.1
1Irving G, Neves AL, Dambha-Miller H, et al. International variations in primary care physician consultation time: a systematic review of 67 countries. BMJ Open2017;7:e017902. doi:10.1136/bmjopen-2017-01790 https://bmjopen.bmj.com/content/bmjopen/7/10/e017902.full.pdf

10) PLEASE CORRECT MULTIVARIATE TO MULTIVARIABLE REGRESSION.
Response: The analysis has changed to multilevel logistic regression.

11) HOW ARE MISSED CARE EVENTS BEING DEFINED? IS THIS JUST THAT THE PATIENT DIDN'T RECEIVE THE CARE OR HOW DO YOU KNOW THAT THE PATIENT WASN'T ADVISED TO DO SO, AND THEN REFUSED? OR ARE THE MISSED CARE EVENTS JUST NOT MEETING THE NUMBER OF EXPECTED EVENTS FROM THE REGRESSION?
Response: The missed care events are not meeting the number of events which would be expected if the care was provided by female GP. There is no information whether patient refused the care. The following sentence was added to the detailed description of calculation of missed care events in Methods section: „We calculated the missed care events, that is the number of care events not meeting the number of events which would be expected if the care was provided by female GP."

12) I'M HESITANT TO STATE THAT PROVIDER GENDER ALONE CAN EXPLAIN SOME OF THE DIFFERENCES OF CARE, ESPECIALLY WHEN YOU'RE ONLY ABLE TO EXPLAIN 5-27% OF THE VARIABILITY FOR EACH CONDITION.
Response: Obviously there are several factors can explain the differences of care, but according to our study one of them is the gender of the provider. We believe that studies focusing on gender effect can contribute to the improvement of national health care systems by reflecting reality in a more accurate way. Raising awareness on this particular gender issue may improve the quality of services in primary care.
Further, it is mentioned in the text that further research is needed to explore the mechanisms responsible for the GP-gender effect. Actually, the paper was to demonstrate that indicators show dependency on GP-gender which is not neglectable and worth to investigate in details.

13) FURTHER, THIS STUDY ONLY LOOKS AT ONE YEAR OF CARE BEING PROVIDED. MANY PATIENTS MAY NOT SEEK THE CARE BEING ADVISED TO THEM RIGHT AWAY, THUS NOT SHOWING UP IN THE SINGLE YEAR OF DATA.
Response: The one year period represents the interval in which the patients have to receive the care according to the guidelines. The patients eligible for the care in each indicators should seek the doctor within a year. Moreover, the studied indicators are used for GP performance assessment purposes by the NIHIFM in Hungary, measuring the proportion of patients who received the care within the given period. The following sentence has been added to the Limitation section: " A limitation of the study is that we were not able to monitor the changes over time due to the cross-sectional design."

14) HOWEVER, YOUR FINDINGS DO SHOW THAT THIS MAY BE AN AREA OF INTERESTING RESEARCH TO BE INVESTIGATED FURTHER.
Response: Thank you for this comment.

Thank you for the careful review and for the suggestions to improve the manuscript!

RESPONSE TO REVIEWER 5:

1) THE SOURCE OF PERFORMANCE INDICATORS IS NOT CLEAR. WERE THE DATA OBTAINED DIRECTLY FROM THE GPS?
Response: The performance indicator data were provided directly from the GMP by the NIHIFM, what we analyzed. The source of data is indicated in Study design: "Demographic data of 7 207 186 clients (above 18 years) and 4 575 GPs, the GMPs' organizational characteristics, and data on performance indicators for GMPs were provided by NIHIFM."

2) ARE THE INDICATORS VALIDATED/ESTABLISHED PARAMETERS WITHIN THE HUNGARIAN HEALTH SYSTEM OR WERE THEY DEVISED AD HOC FOR THIS STUDY?

Response: The indicators are validated parameters within the Hungarian primary health care system. The performance indicators applied by NIHIFM are used to monitor and assess the performance of GPs. In Performance indicators for GMPs section in Methods contains the following sentence: ”Each routine indicator of NIHIFM on immunization, cancer screening, and chronic disease management were used to assess the performance of GMPs.”

3) THE BASELINE CHARACTERISTICS OF FEMALE-LED GMPS ARE SYSTEMATICALLY AND CONSISTENTLY DIFFERENT COMPARED WITH MALE-LED GMPS. FOR ALL, MALE-LED GMPS ARE CONSISTENTLY LARGER, WHICH IS LIKELY A MAJOR DETERMINANT OF WHETHER ADEQUATE CARE IS GIVEN TO THE INDIVIDUAL PATIENT. ALTHOUGH THE STATISTICAL ANALYSIS IS ADJUSTED FOR SUCH DIFFERENCES, IT IS STILL INDICATIVE OF THE POSSIBILITY THAT RESIDUAL CONFOUNDING OR SYSTEMATIC BIAS NOT OTHERWISE MEASURABLE COULD EXPLAIN THE DIFFERENCES IN SEX

Response: The statisitcal analyses were revised. We used multilevel logistic regression model which is a more precise approach to handle clustered data. Furthermore, instead of further categorization of our data, we run the regression models using the data as given (e.g. we used the age of the GP as a continous variable in the model instead of the age categories, which we used in the previous version). Obviously there are several factors besides confounding variables used in the analysis which can still explain the differences of care. We indicated the existence of other possible confounding factors in the Limitation: „Additionally, the fact that the range of variables in the NIHIFM database which could be included in the analysis is limited, careful interpretation is required. Although the analyses were controlled for GPs' and patients' age and gender, educational attainment, GMP practice size, types of settlement and regional location, there were confounding factors that were not included in our models, limiting the reliability of the presented risk measures. Selection of the GP may be affected by patients' preferences and expectations, as patients have free choice of health care provider in Hungary. Since most of the GPs in Hungary are self-employed workers organizing the provision on their own authority, neither the length of their working hours nor the length of the consultation time were mesurable.”

4) THE OUTCOME VARIABLE IS A PROPORTION (NUMBER OF PATIENTS RECEIVING A GIVEN INTERVENTION DIVIDED BY THE NUMBER OF PATIENTS ELIGIBLE FOR THAT INTERVENTION). THE USE OF A DIFFERENT REGRESSION MODEL SUCH AS NEGATIVE BINOMIAL OR POISSON WOULD BE PREFERABLE COMPARED WITH LINEAR REGRESSION. ON TOP OF A POTENTIALLY BETTER FIT, THE USE OF SUCH MODELS WOULD ALLOW THE AUTHORS TO OBTAIN MORE UNDERSTANDABLE ESTIMATES IN THIS CONTEXT RATHER THAN BETAS

Response: : Thank you for your comments! Considering your advice regarding to analysis, we consulted with a statistician. Consequently, the statistical analyses have been revised. As a result, we applied another (hierarchical) data structure which allows to fit multilevel analyses on our data. After „disaggregation” of the data, we used multilevel logistic regression models because of the binary outcome variable, and in order to account for the clustering effect of patients within physicians. Moreover, we were able to estimate the total variation explained by the physician groups by calculating the intraclass correlation coefficient (ICC).

Accordingly, the Statistical analysis section in Methods has been corrected: „We used multilevel logistic regression models because of the binary outcome variables and in order to account for the clustering effect of patients within physicians. The models were adjusted for characteristics of patients (age), GPs (gender, age) and GMPs (practice size, types of settlement, geographical location and relative education) to assess the effect of the gender of the GP on their performance in case of each indicator. Odds ratios (ORs) with the corresponding 95% confidence intervals (CIs), and robust standard errors were estimated. We presented intra-class correlation coefficient (ICC), which shows the proportion of total variance explained by the physician as grouping factor.”

The Results section has also been corrected: ”The female gender of GPs was associated with hypertension and diabetes care related indicators (HbA1c measurement: OR=1.18, 95%CI [1.14-1.23]; serum creatinine measurement: OR=1.14, 95%CI [1.12-1.17]; lipid measurement: OR=1.14, 95%CI [1.11-1.16]; eye examination: OR=1.06, 95%CI [1.03-1.08]), mammography screening (OR=1.05, 95%CI [1.03-1.08]), management of COPD patients (OR=1.05, 95%CI [1.01-1.09]), and the composite indicator (OR=1.08, 95%CI [1.07-1.1]).”
The Table 4 has been corrected.

VERSION 3 – REVIEW

REVIEWER	Dr Kathryn Taylor University of Oxford UK
REVIEW RETURNED	28-Jun-201

GENERAL COMMENTS	The authors have done a good job of revising the manuscript, which has made reviewing easier and, from my perspective, the process is converging. My remaining concerns stem from the way that the authors have presented a study of solo GP practices. The redraft of the abstract does not make it clear that the study is about a health system that will be different and unfamiliar to most of those who read the journal. I would expect that decision making in group GP practices will differ from that of solo practices, as in the former, testing will involve group discussion and the existence of practice-based policies which will reduce the effects of any non-clinical factors in decision making (e.g. gender). I would also expect that the availability of support staff in solo practices could be will be an important factor in determining whether or not a patients is tested (therefore its absence from the analysis should be explicitly stated). The conclusions need to be moderated to account for the limitations of the study. The level of English is generally acceptable but there a few English errors which are listed below with a few other minor issues (page numbers refer to the tracked version in the proof copy): 1. Page 34, lines 44-45 (missing articles): "was used as a reference... The age of the GP was a continuous variable....were used as a reference.." 2. Page 41, line 53: "Intercooled STATA" is unfamiliar. "STATA IC version 13.0" is more standard. 3. Page 44, line 25: Supplementary tables present results of models, not details of models. 4. Page 44, lines 7-16: The results reported in the text are repeated in Table 4. The ORs should be removed from the text. 5. Page 46, Table 4: I would think that the audience of BMJ Open will know that a CI of an OR that does not contain 1.0 indicates a statistically significant result and also given that the text clearly states what’s significant and what’s not, the use of bold font is not necessary. 6. Page 50, line 51: “measurable” (spelling correction). The manuscript should be spell-checked. 7. Page 50, lines 57-58: The redraft of this sentence has lost the point about statistical significance vs clinical significance and a key problem which has confused matters is that this sentence is in the wrong place. The interpretation of results belongs in the implications section. Interpretations have to take account of the

	study limitations. 8. Page 51, line 3: Misuse of “However” and long sentence. The sentence could be simplified by leading on the point about the study being based on solo practices and how the results may differ in group GP practices. 9. Page 51, lines 52-54: “Training in effective communication……suggest that the scope of this training should be” (grammar) 10. Page 52, line 4: Given the limitations of the study, the results only suggest the existence of a gender effect.

REVIEWER	Pietro Manuel Ferraro Fondazione Policlinico Universitario A. Gemelli IRCCS, Roma, Italia
REVIEW RETURNED	03-Jul-2019

GENERAL COMMENTS	I appreciate the new statistical approach of the authors, which is much more suited to their data. I also appreciate the added limitations related to unmeasured confounding

VERSION 3 – AUTHOR RESPONSE

RESPONSE TO REVIEWER 3:

THE AUTHORS HAVE DONE A GOOD JOB OF REVISING THE MANUSCRIPT, WHICH HAS MADE REVIEWING EASIER AND, FROM MY PERSPECTIVE, THE PROCESS IS CONVERGING. MY REMAINING CONCERNS STEM FROM THE WAY THAT THE AUTHORS HAVE PRESENTED A STUDY OF SOLO GP PRACTICES. THE REDRAFT OF THE ABSTRACT DOES NOT MAKE IT CLEAR THAT THE STUDY IS ABOUT A HEALTH SYSTEM THAT WILL BE DIFFERENT AND UNFAMILIAR TO MOST OF THOSE WHO READ THE JOURNAL. I WOULD EXPECT THAT DECISION MAKING IN GROUP GP PRACTICES WILL DIFFER FROM THAT OF SOLO PRACTICES, AS IN THE FORMER, TESTING WILL INVOLVE GROUP DISCUSSION AND THE EXISTENCE OF PRACTICE-BASED POLICIES WHICH WILL REDUCE THE EFFECTS OF ANY NON-CLINICAL FACTORS IN DECISION MAKING (E.G. GENDER). I WOULD ALSO EXPECT THAT THE AVAILABILITY OF SUPPORT STAFF IN SOLO PRACTICES COULD BE WILL BE AN IMPORTANT FACTOR IN DETERMINING WHETHER OR NOT A PATIENTS IS TESTED (THEREFORE ITS ABSENCE FROM THE ANALYSIS SHOULD BE EXPLICITLY STATED). THE CONCLUSIONS NEED TO BE MODERATED TO ACCOUNT FOR THE LIMITATIONS OF THE STUDY.
Response: Thank you for highlighting this point. The abstract has been modified as follows:
„Setting and participants: The study covered all general medical practices in Hungary (N=4575) responsible for the provision of primary health care for adults. All GPs in their private practices are solo practitioners."
„Conclusion: …Factors behind the gender effect should receive more attention in quality improvement particularly in countries where the primary care is organised around solo practices.”
The Limitation section has been completed: " Since most of the GPs in Hungary are self-employed workers organizing the provision on their own authority, the length of their working hours, the length of the consultation time and the availability of support staff were not measurable. The availability of support staff in solo practices or work in group practice could be an important factor in determining whether and how a patient will be treated.”
The following sentence has been modified in the Conclusion: „Our results suggest the existence of a gender effect in Hungary on performance of GPs working in solo practices.”
The following sentence has been added to the Conclusion: „However further studies involving a

broader range of clinical and non-clinical factors are needed to allow more precise assessment of gender differences."

THE LEVEL OF ENGLISH IS GENERALLY ACCEPTABLE BUT THERE A FEW ENGLISH ERRORS WHICH ARE LISTED BELOW WITH A FEW OTHER MINOR ISSUES (PAGE NUMBERS REFER TO THE TRACKED VERSION IN THE PROOF COPY):
1. PAGE 34, LINES 44-45 (MISSING ARTICLES): "WAS USED AS A REFERENCE... THE AGE OF THE GP WAS A CONTINUOUS VARIABLE....WERE USED AS A REFERENCE.."
Response: Corrected.

2. PAGE 41, LINE 53: "INTERCOOLED STATA" IS UNFAMILIAR. "STATA IC VERSION 13.0" IS MORE STANDARD.
Response: Corrected.

3. PAGE 44, LINE 25: SUPPLEMENTARY TABLES PRESENT RESULTS OF MODELS, NOT DETAILS OF MODELS.
Response: Corrected.

4. PAGE 44, LINES 7-16: THE RESULTS REPORTED IN THE TEXT ARE REPEATED IN TABLE 4. THE ORS SHOULD BE REMOVED FROM THE TEXT.
Response: Corrected.

5. PAGE 46, TABLE 4: I WOULD THINK THAT THE AUDIENCE OF BMJ OPEN WILL KNOW THAT A CI OF AN OR THAT DOES NOT CONTAIN 1.0 INDICATES A STATISTICALLY SIGNIFICANT RESULT AND ALSO GIVEN THAT THE TEXT CLEARLY STATES WHAT'S SIGNIFICANT AND WHAT'S NOT, THE USE OF BOLD FONT IS NOT NECESSARY.
Response: Corrected.

6. PAGE 50, LINE 51: "MEASURABLE" (SPELLING CORRECTION). THE MANUSCRIPT SHOULD BE SPELL-CHECKED.
Response: Corrected.

7. PAGE 50, LINES 57-58: THE REDRAFT OF THIS SENTENCE HAS LOST THE POINT ABOUT STATISTICAL SIGNIFICANCE VS CLINICAL SIGNIFICANCE AND A KEY PROBLEM WHICH HAS CONFUSED MATTERS IS THAT THIS SENTENCE IS IN THE WRONG PLACE. THE INTERPRETATION OF RESULTS BELONGS IN THE IMPLICATIONS SECTION. INTERPRETATIONS HAVE TO TAKE ACCOUNT OF THE STUDY LIMITATIONS.
Response: The sentence has been moved to the Implication section: „Keeping in mind that large sample size can detect even the smallest differences, the interpretation needs to consider that this small differences may have limited importance at population level; our results indicate the significance and urge the expansion…"

8. PAGE 51, LINE 3: MISUSE OF "HOWEVER" AND LONG SENTENCE. THE SENTENCE COULD BE SIMPLIFIED BY LEADING ON THE POINT ABOUT THE STUDY BEING BASED ON SOLO PRACTICES AND HOW THE RESULTS MAY DIFFER IN GROUP GP PRACTICES.
Response: The sentence has been simplified: "Our analysis, which was based on data on solo GP practices, may differ from findings in group GP practices where professional cooperation with other providers may mitigate the effects of any non-clinical factors (e.g. gender of GP)."

9. PAGE 51, LINES 52-54: "TRAINING IN EFFECTIVE COMMUNICATION……SUGGEST THAT THE SCOPE OF THIS TRAINING SHOULD BE" (GRAMMAR)
Response: Corrected.

10. PAGE 52, LINE 4: GIVEN THE LIMITATIONS OF THE STUDY, THE RESULTS ONLY SUGGEST THE EXISTENCE OF A GENDER EFFECT.
Response: The original sentence has been modified: „Our results suggest the existence of a gender effect in Hungary on performance of GPs working in solo practices."

Thank you for the careful review and for the suggestions to improve the manuscript!

RESPONSE TO REVIEWER 5:
I APPRECIATE THE NEW STATISTICAL APPROACH OF THE AUTHORS, WHICH IS MUCH MORE SUITED TO THEIR DATA. I ALSO APPRECIATE THE ADDED LIMITATIONS RELATED TO UNMEASURED CONFOUNDING.