PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	Social relationships and GP use of middle-aged and older adults in Europe: a moderator analysis
AUTHORS	Bremer, Daniel; Lüdecke, Daniel; Vonneilich, Nico; von dem Knesebeck, Olaf

VERSION 1 – REVIEW

REVIEWER	Dr Caroline S Clarke UCL, UK
REVIEW RETURNED	14-Aug-2017

GENERAL COMMENTS	bmjopen-2017-018854 Social relationships and GP use of older adults in Europe: a moderator analysis Comments for the authors: This was a very interesting paper and I am very pleased to have had the opportunity to read it. I have a number of comments and suggestions that I hope will improve this report, and I hope that they are helpful to the authors. I have tried to be concise, so hope that my wording does not seem brusque. - Abstract and elsewhere: Using (1), (2), (3) to denote three ideas is a little confusing as they look like citations. Please change to something like (i), (ii), (iii). - Page 4, line 16: I'm not sure that "older adults" would normally include people in their 50s – this term generally refers to those aged over 65. - Page 6, line 15: Please can "contact rate" be explained – does it mean the rate of finding and/or re-contacting people, compared to the previous Wave, or compared to the numbers in Wave 1, or something else? - Page 6, line 21: Refusal rates seem high (22% - 49%) – what assumptions are made around this (e.g. missing (not) at random?), and how are the missing data dealt with? Do we know anything about the people who have refused? Please discuss this in the paper. - General comment: Twelve months is a long time for people to recall visits – how is any

	bias created by this being dealt with? What validation has there been that the figures recalled by people in this way are accurate? Are they given categories, or just asked to write down a number? [also see later note on this] - General comment: What about seeing the nurse at their GP practice? Is this thought to be important? This might vary by country. Please consider discussing this. - Page 5-6: Are the eligibility criteria simply that respondents must be "non-institutionalized adults aged 50 years or older and their spouses"? If not, please clarify. How were people found and recruited into the study? - Page 7: How did people respond if they lived with someone but were not married to them (e.g. co-habiting partner, sibling, children, etc.)? How might the lack of an option like this have generated bias in the study? Would this vary by country/language? Please discuss this in the paper. - Page 7, line 37 It says "in any of the six social groups" and then lists 5. Please check. - Last paragraph of page 7 – first paragraph of page 8: Were the questions contact frequency part of a validated questionnaire? If so, please give the reference and, if not, where did the questions come from? Please comment on this in the paper. - Page 8, lines 19-22: The question on health asks about self-perceived health status, not health needs. The two concepts are related but not the same. Please discuss this in the paper. - Page 8, lines 32-36: The not-employed group is a very broad group. I am not sure that grouping all these people together is very meaningful. Perhaps better to keep their analyses separate, as is done later? - Page 8, lines 38-42: Is the "making ends meet" question a validated one? Where did it come from? Please comment on this in the paper. - General comment on methods: Please consider any correlation between similar variables. Please also consider where confounding might take place (e.g. education and income). Please discuss this in the paper. - Page 8, line 51. Number of GP visits is treated as a count variable. What would happen if the responses were grouped into categories, depending on the range of responses available. Does that change the results? It seems that this kind of sensitivity analysis could potentially be informative, given the 12-month recall period. - Table 1: Please state that Age is in years.

	Please add a reminder of the scale for self-perceived health status and social integration index. Please state the denominator for contact frequency (number of contacts per week/month/year?). Please explain somewhere in the body of the paper what the difference between n and N is. Please change all thousand separators to commas instead of full-stops. How does this group of people compare to "average" people in this region? Is the population similar across all 16 countries? Please comment on the representativeness of the study population. - General comment: Please list the countries in the study. Should there be a variable for country in any of the analyses, or if not, please describe how the countries are incorporated into the analysis. If the analyses are done separately for each country, do the results match the overall results presented in the paper? This might be an important sensitivity analysis to perform. - Table 2: How were these models chosen? Are they the best fit models? E.g. please state AIC/BIC in related models. Which is the best fit – the basic model or the interaction model? As gender is not significant, is the fit better if that is taken out? What does that mean? Please explain alpha. The scale for health status is given as 0-4 in the table, but 1-5 in the text. Please streamline. Are all the variables with scales treated as continuous variables, e.g. self-perceived health status, education, social integration index, contact frequency, and is this appropriate? In treating them as continuous, there is an assumption made that the scale has interval properties. Please consider using categorical variables, and discuss and justify the choices made in the paper. The same goes for Age. - General comment on methods: It does not seem entirely clear why simultaneously including the three interaction terms between health status and the three social network/integration variables should help elucidate the role that social network/integration plays in the relationship between health status and frequency of GP visits. Why these three? Are they thought to be independent of each other, or are they correlated at all? Why include the interaction terms simultaneously? Which model actually gives the best fit to the data? Was more work done before arriving at these as the models? What if only one interaction term were included, for example? - Page 11, line 7: This mentions models 1a and 1b as looking at employed respondents, but it seems from Table 2 that models 1a and 1b include the whole sample. Please check numbering of models.

	- General comment on methods: As mentioned above, I am not sure that grouping together retirees, home-makers, those too ill to work, those with disabilities such that they cannot work, and those simply unemployed into a single category is meaningful (models 2a and 2b). Please discuss this in the paper. - Page 16, lines 14-18: I am not very familiar with this literature, but is it also possible that there is a variety of definitions used in the research, and variable data quality, gathered using a variety of different questions/questionnaires and in different contexts? This could also lead to inconsistencies in the results found. - Page 16, lines 23-49: The last sentence here is intriguing – perhaps clarify how this idea fits together with the rest of the paragraph. - Page 16, last line: Please reference which table this result is in. - Page 17, line 5: The health question is about health status, and not health needs. Please discuss the relationship between these two concepts. - Page 18, line 39: The reference cited states that "Older age was generally associated with underreporting". Please therefore discuss this bias with consideration of the direction of bias, rather than just mentioning "risk of memory bias". - Page 18, line 46: It is also only a single question. Self-reported health status is often assessed using validated questionnaires such as the SF-12, or the EQ-5D, etc., rather than with a single question. Please comment on this in the paper. - General comment: There is also no mention of visits to the nurse at the GP practice. This would be an important thing to include if this study were to take place in the UK – please consider commenting on this if it is relevant to the countries in the study. - Page 19, line 49: "GPs should take "employment status" and "social relationships" of their patients into account." This is very bold. Is it truly justified by this analysis? How could it be done in a fair way, and what would be the benefit to patients? - Page 19, lines 52-56: This is also bold: "Finally, our results indicate the necessity to integrate information on social relationships and employment status into debates on needs-based access to health care and adequate levels of treatment." Please discuss this further and state the justification. - STROBE checklist: Please address the impact of having missing data. They seem to have been ignored, which is ok if we know they are missing at random, but I am not sure that we do know this. If they are assumed

	to be missing at random, then the rationale for assuming this should be discussed. Please comment on this in the paper. Was any information available on people's reasons for not taking part? Please describe any efforts made during the analysis to address sources of bias – see various comments above. Please consider the direction of all biases – see comments above. Please describe what sensitivity analyses were done – see various comments above. The interpretation is not always cautious. There are some limitations on the source data and analysis methods, so perhaps more caution could be used? - ETHICS Research Ethics approval for the original study or for this analysis is not mentioned. - English: Some slightly unnatural phrases, e.g. "…moderation of social relationships on…" could perhaps be phrased as "….moderating action of social relationships on…". Also, "SR did matter in diverse facets in the not-employed-group" sounds a bit strange. Perhaps something like: "SR did appear to be significant in some sections of the not-employed group" (if that is what is meant). Decimal points should be full-stops (and not commas), and thousand separators should be commas (and not full-stops), e.g. "56,989 interviews". Please change this in all figures, and all tables. "The Social Integration Index by Berkmann et al. [30] was shown to be a reliable and robust approach…" – this implies that it was shown to be reliable by this study. If this is not the case, then perhaps re-phrase to something like: "The Social Integration Index by Berkmann et al. has been shown [30] to be a reliable and robust approach…", if this is what was meant.

REVIEWER	Chiu-Yueh Hsiao Asia University, Taiwan
REVIEW RETURNED	14-Sep-2017

GENERAL COMMENTS	1. It seems that the process of the IRB approval process was lacking in this manuscript. Please provide it. 2. In the anstract, one of objectives was ....(2) if SR moderate the association between health needs and GP visits. However, on page 5, line 24, (2) if social ties moderate the association between health needs and GP use. Please make it consistent. 3. Regarding the objectives of this study, please add more evidence-based information in the literature reivew. 4. Please provide a more clear rationale about why this study attempted to investigate the moderate effect of social relationships on the link between health needs and GP visits.

REVIEWER	Joshua Betz Johns Hopkins Bloomberg School of Public Health Department of Biostatistics
REVIEW RETURNED	29-Sep-2017

GENERAL COMMENTS	I am a statistician by training, so my review is mainly focused on the analysis and interpretation of results: I hope you find my comments helpful. The analysis performed (a negative binomial GLM using probability weights to account for the survey design) assumes that all responses are independent. While this could be a reasonable approximation, it's conceivable that households share health, behavioral, and social factors that might induce correlation among the outcomes, resulting in a lower "effective sample size" for a given sample size. Standard errors derived using the assumption of independence inherent in the GLM may be too narrow (anti-conservative) as a result. Multilevel models should be considered, using random effects to account for within-household correlations in outcome, and if reasonable to the analysts, nested random effects to account for correlations between households within a country. While the point estimates may be similar, the standard errors and confidence intervals may be more accurate if participants within households and households within countries are not effectively independent. I believe gllamm and other procedures for GLMEs in Stata handle survey weights and nested random effects. To the best of my knowledge at the time of this review, glmer in R's lme4 package only supports frequency weights, not probability weights. Mixed models for negative binomial outcomes in either package are more recent developments, so their ability to handle both survey weights and nested random effects may be limited. One assumption of statistical models that could be further described is how the assumption of linearity was addressed: common methods include residual diagnostic plots (added variable or component-residual plots) or the use of smoothing spline terms (mkspline in Stata; bs(), ns(), pspline() or other functions in R). For ordered categorical predictors (perceived health, social integration, etc.), graphical comparisons of the fitted categorical coefficients can be used. It is plausible that the effect of an increment in important variables like age or self-assessed health may vary in magnitude depending on the location of the increment (compare ages 50-60 vs. 60-70 or self-assessed health "1. Excellent"-"2. Very Good" vs. "4. Fair"-"5. Poor"). Linearity is more difficult to assess in interactions, but the main effects (especially age and self-assessed health status) should be checked. Inadequate fit (either ignoring relevant confounding variables or modeling of non-linear associations as linear) may result in residual confounding. Scaling age in terms of decades instead of years may make the coefficient easier to interpret. Changes in statistical significance can arise when interaction terms are introduced even in the absence of a genuine interaction because of variance inflation: interaction terms are correlated with the main effects, resulting in wider standard errors and confidence intervals. This variance inflation would be more pronounced if there were appreciable correlation between the three aspects of social

	relationships measured. Inference should focus on the direction and magnitude of interaction terms and their confidence intervals, as change in the statistical significance of the main effect once an interaction is introduced is not necessarily meaningful (See Gelman and Stern's 'The Difference Between "Significant" and "Not Significant" is not Itself Statistically Significant' (http://www.tandfonline.com/doi/abs/10.1198/000313006X152649), especially when they may naturally arise from variance inflation. Other considerations and clarifications: Please state whether or not subgroup analyses were pre-specified in the analytic plan. In table 1 (page 9), the cohort is described in aggregate: would it be possible to provide the aggregate results as well additional columns describing individuals by employment status (employed/retired/unemployed/disabled/homemaker)? These may be useful in understanding the stratified results. In addition, education and self-assessed health status are categorical variables - Reporting the category frequencies may be more interpretable than the mean and standard deviation of the numeric category labels. The intercept may be made more meaningful by centering the covariates (subtracting a reference value - either some clinically meaningful values or the sample mean value), giving the average number of visits for individuals with reference levels for all covariates. Please mention if centering either has or has not been performed. It's also noted that if respondents had more than 98 contacts with their doctor about their health, the number was censored at 98. Please list the frequency and proportion of such events. If very rare, censoring may have minimal impact on results, but if appreciable, this can bias means and variances downward, and confidence intervals may be too narrow as a result. What were the correlations between social integration, average frequency of contact, and number of 'very close'/'extremely close' people in social networks? If these are moderately or strongly correlated, having all of these terms plus interactions with all of these terms in the same model makes collinearity a potential concern. Polychoric correlations (polychoric in Stata) might be useful here because of the ordinal nature of the covariates. On page 8, self-assessed heath is rated on a scale of "1. Excellent" to "5. Poor," with higher numbers indicating worse health. In regression models, if self-assessed health was treated as a linear covariate in model 1a, each unit increase on this scale (towards poorer health) was associated with a 26% reduction in the rate of GP visits (IRR=0.74). It seems counterintuitive to me that those reporting worse self-assessed health would report lower frequency of GP utilization (although my intuition might be in error). Please confirm that this variable has not been re-scaled such that higher values indicate better self-assessed health. The buffer function of social integration is one of many possible explanation of the negative association with GP usage. Since the social integration includes participation in charity work, education, social clubs, religious services, or community/political organizations, could this also represent a healthy user effect whereby those who are less infirmed or spend less days in healthcare utilization are

	more able to participate in such events or maintain/expand their social networks? Consider a mention of multiplicity to accompany interpretation: three interaction terms were estimated in several models: in aggregate, Employed vs. Unemployed, and within the Unemployed (Retired, Unemployed, Sick/Disabled, and Homemakers). In most places, results are discussed in the language of associations, but on page 17 (lines 22-28) and 19 (lines 32-36), the language sounds more causal in tone ("lowers", "increase", "decrease"). Please consider "was associated with lower/higher" instead, which is consistent with the rest of the discussion.

REVIEWER	Molly Rosenberg Department of Epidemiology and Biostatistics Indiana University School of Public Health-Bloomington USA
REVIEW RETURNED	05-Oct-2017

GENERAL COMMENTS	I reviewed the manuscript "Social relationships and GP use of older adults in Europe: A moderator analysis" specifically for statistical methodology and analysis. Overall I found the methods section of the manuscript to be well-written and generally justifiable. Negative binomial models are appropriate for the count outcomes with overdispersion. I note two methodological concerns, one more major, one more minor, below. 1. My primary concern is that the way the models were constructed and how they were presented makes it difficult to understand the bivariate relationships between each exposure of interest and the GP visits outcome as they are only shown adjusted for blocks of other variables that may or may not be confounders of the relationship of interest. This is particularly true of the three inter-related social relationship variables. Including them all in the same models when they are likely highly collinear with each other is likely to produce biased estimates (e.g. What does the coefficient for social integration index adjusted for number of extremely close people in social network mean?). Further, pulling each of these estimates out from the same model and interpreting them as estimates of unconfounded relationships is problematic. See: Westreich, Daniel, and Sander Greenland. "The table 2 fallacy: presenting and interpreting confounder and modifier coefficients." American journal of epidemiology 177.4 (2013): 292-298. 2. The authors run multiple models in several different subpopulations without adjustment for multiple comparisons.

VERSION 1 – AUTHOR RESPONSE

Response to Review Comments
(bmjopen-2017-018854)

We would like to thank the reviewers and editor for the comments. We have revised the manuscript accordingly. We hope that the manuscript has improved to a level of your satisfaction.

On the following pages you will find the comments with our responses and corresponding revisions. Revised paragraphs in our paper are presented in our responses following "please see p. XY". All changes in the manuscript are highlighted in the revised manuscript marked changes.

Editor #1

1. Please include some quantitative results or statistics in the abstract results section.

Responses to the comments:
Please see our updated abstract.

Reviewer #1

1. Abstract and elsewhere:

Using (1), (2), (3) to denote three ideas is a little confusing as they look like citations. Please change to something like (i), (ii), (iii).

Responses to the comments:
We changed the numbering to (i), (ii) and (iii).

2. Page 4, line 16:

I'm not sure that "older adults" would normally include people in their 50s – this term generally refers to those aged over 65.

Responses to the comments:
We changed the wording in the title, abstract and introduction accordingly to the definition of the Oxford English Dictionary to "middle-aged and older adults".

Please see footnote on p. 5:
Due to the readability, we refer to "middle-aged and older adults" or "adults 50 years or older" when we write about "older adults" in this paper.

3. Page 6, line 15:

Please can "contact rate" be explained – does it mean the rate of finding and/or re-contacting people, compared to the previous Wave, or compared to the numbers in Wave 1, or something else?

Responses to the comments:
The proportion of eligible households that were contacted (including panel samples and refresher samples).

Reference:

- 38. Malter, F.; Börsch-Supan, A.; (2013). SHARE Compliance Profiles – Wave 4. Munich: MEA, Max Planck Institute for Social Law and Social Policy.
- 37. Malter, F., Börsch-Supan, A.(Eds.)(2013). SHARE Wave 4: Innovations & Methodology. Munich: MEA, Max Planck Institute for Social Law and Social Policy.

4. Page 6, line 21:

Refusal rates seem high (22% - 49%) – what assumptions are made around this (e.g. missing (not) at random?), and how are the missing data dealt with? Do we know anything about the people who have refused? Please discuss this in the paper.

Responses to the comments:

Members of the SHARE-team conducted non-response analyses and "concluded that there is little evidence for non-response bias amongst the variables we examined here" [gender, age, health, employment, number of children, income] (Malter & Börsch-Supan 2013).

Please see p. 18:
Due to unit non-response and panel attrition, sample selection bias is a potential problem limiting the representativeness of the data and the generalizability of results [35]. However, non-response analyses taking various variables into account (gender, age, health, employment, number of children, and income) showed only little evidence for non-response bias (e.g., a slightly larger number of males among respondents than non-respondents) [37].

Reference:

- 35. Börsch-Supan A, Brandt M, Hunkler C, et al. Data Resource Profile: the Survey of Health, Ageing and Retirement in Europe (SHARE). Int J Epidemiol. 2013;42:992-1001.
- 37. Malter, F., Börsch-Supan, A.(Eds.)(2013). SHARE Wave 4: Innovations & Methodology. Munich: MEA, Max Planck Institute for Social Law and Social Policy.

5. General comment:
Twelve months is a long time for people to recall visits – how is any bias created by this being dealt with? What validation has there been that the figures recalled by people in this way are accurate? Are they given categories, or just asked to write down a number? [also see later note on this]

Responses to the comments:
Respondents could name the number of visits.

Please see p. 18 and 19:
The time span covering the GP contacts is quite long, and considering the older age of the interviewed individuals, risk of memory bias is existent with regard to self-reported utilization data [62]. Bhandari and Wagner found in their systematic review on self-reported utilization of health care services that "[…] age was the most consistent demographic factor associated with self-report inaccuracy […]" by older adults underreporting their use [62]. Consequently, intercepts and age coefficients in our models could be potentially underestimated.

Reference:

- 62. Bhandari A, Wagner T. Self-reported utilization of health care services: improving measurement and accuracy. Medical Care Research and Review. 2006;63:217-35.

6. General comment:
What about seeing the nurse at their GP practice? Is this thought to be important? This might vary by country. Please consider discussing this

Responses to the comments:
We agree. Focusing on GPs and doctors at health care centers could create an underestimation of primary health care services in some countries. The magnitude may vary by country. Unfortunately, we do not have information on the proportion of nurse-specific services in general practices by country.

Please see p. 18 and 19:
The question used in SHARE to cover the use of GP services across 12 months is established in health services research [4, 20, 21, 61], it has some methodological drawbacks. The question is narrowed to contacting a GP or doctor in a health care center. Contacts with nurses at GP practices are not taken into account. Potentially, the level of using primary care is underestimated.

References:

- 4. Korten AE, Jacomb PA. Predictors of GP service use: a community survey of an elderly Australian sample. Australian & New Zealand Journal of Public Health. 1998;22:609-15.
- 20. Li Y, Chi I. Correlates of physician visits among older adults in China: the effects of family support. J Aging Health. 2011;23:933-53.
- 21. Miltiades HB, Wu B. Factors affecting physician visits in Chinese and Chinese immigrant samples. Soc Sci Med. 2008;66:704-14.
- 61. Gobbens RJJ, van Assen MALM. Frailty and its prediction of disability and health care utilization: The added value of interviews and physical measures following a self-report questionnaire. Archives of Gerontology and Geriatrics. 2012;55:369-79.

7. Page 5-6:

Are the eligibility criteria simply that respondents must be "non-institutionalized adults aged 50 years or older and their spouses"? If not, please clarify. How were people found and recruited into the study?

Responses to the comments:

Please see p. 6:

Based on population registers, SHARE uses probability samples within the countries and includes non-institutionalized adults aged 50 years or older and their partners. Further exclusion criteria are being incarcerated, moved abroad, unable to speak the language of questionnaire, deceased, hospitalized, moved to an unknown address or not residing at sampled address [35, 37].

References:

- 35. Börsch-Supan A, Brandt M, Hunkler C, et al. Data Resource Profile: the Survey of Health, Ageing and Retirement in Europe (SHARE). Int J Epidemiol. 2013;42:992-1001.
- 37. Malter F, Börsch-Supan A. SHARE Wave 4: Innovations & Methodology. Munich: MEA, Max Planck Institute for Social Law and Social Policy; 2013.

8. Page 7:

How did people respond if they lived with someone but were not married to them (e.g. co-habiting partner, sibling, children, etc.)? How might the lack of an option like this have generated bias in the study? Would this vary by country/language? Please discuss this in the paper.

Responses to the comments:

Partners were interviewed regardless of their legal status.

Please see:

comment nr. 7 above.

9. Page 7, line 37

It says "in any of the six social groups" and then lists 5. Please check.

Responses to the comments:

We checked and corrected the number.

Please see p. 8:

Third domain: the affiliation with voluntary organizations was measured by activities in any of the five social groups: 'Which of the activities have you done in the past twelve months? 1. Done voluntary or charity work, 2. Attended an educational or training course, 3. Gone to a sport, social or other kind of

club, 4. Taken part in activities of a religious organization (church, synagogue, mosque etc.), 5. Taken part in a political or community-related organization'.

10. Last paragraph of page 7 – first paragraph of page 8:
Were the questions contact frequency part of a validated questionnaire? If so, please give the reference and, if not, where did the questions come from? Please comment on this in the paper.

Responses to the comments:
Please see p. 8:
This module was based on other similar studies, such as the National Social Life, Health, and Aging Project (NSHAP) [40], the American General Social Survey and the Longitudinal Aging Study Amsterdam [41-43].

References:

- 40. Cornwell B, Schumm LP, Laumann EO, et al. Social Networks in the NSHAP Study: rationale, measurement, and preliminary findings. Journals of Gerontology Series B: Psychological Sciences and Social Sciences. 2009;64:i47-i55.
- 41. Van Tilburg TG. Delineation of the social network and differences in network size. Living arrangements and social networks of older adults. 1995:83-96.
- 42. Burt RS. A note on sociometric order in the general social survey network data. Social Networks. 1986;8:149-89.
- 43. Burt RS, Guilarte MG. A note on scaling the general social survey network item response categories. Social Networks. 1986;8:387-96.

11. Page 8, lines 19-22:
The question on health asks about self-perceived health status, not health needs. The two concepts are related but not the same. Please discuss this in the paper.

Responses to the comments:
Please see p. 19:
The limited level of information of self-reported data holds also true for all other variables in our analyses, especially for the variable "self-rated health" which is culturally sensitive [63]. Although, self-rated health status is based on a single item, it is a suitable summary of health [64]. Studies on several representative samples showed that health ratings can be used as valid measures of health regardless of different cultures and social conditions [65-67]. Furthermore, self-rated health is used as a substitute for health needs in this study. To predict need for and use of health care services, perceived health status corresponds well to the objective health status [68, 69]. Consequently, using self-perceived health only represents an approximation of health needs, since SHARE did not include items on (perceived) health needs.

Reference:

- 63. Jylhä M, Guralnik JM, Ferrucci L, et al. Is Self-Rated Health Comparable across Cultures and Genders? The Journals of Gerontology: Series B. 1998;53B:S144-S52.
- 64. Singh-Manoux A, Martikainen P, Ferrie J, et al. What does self rated health measure? Results from the British Whitehall II and French Gazel cohort studies. Journal of Epidemiology & Community Health. 2006;60:364-72.
- 65. Miilunpalo S, Vuori I, Oja P, et al. Self-rated health status as a health measure: the predictive value of self-reported health status on the use of physician services and on mortality in the working-age population. Journal of clinical epidemiology. 1997;50:517-28.
- 66. Mossey JM, Shapiro E. Self-rated health: a predictor of mortality among the elderly. American journal of public health. 1982;72:800-8.

• 67. DeSalvo KB, Bloser N, Reynolds K, et al. Mortality prediction with a single general self-rated health question. Journal of general internal medicine. 2006;21:267-75.
• 68. Hunt SM, McKenna S, McEwen J, et al. A quantitative approach to perceived health status: a validation study. Journal of Epidemiology & Community Health. 1980;34:281-6.
• 69. Johnson RJ, Wolinsky FD. The structure of health status among older adults: disease, disability, functional limitation, and perceived health. Journal of health and social behavior. 1993:105-21.

12. Page 8, lines 32-36:
The not-employed group is a very broad group. I am not sure that grouping all these people together is very meaningful. Perhaps better to keep their analyses separate, as is done later?

Responses to the comments:
We followed your recommendation and dissolved the not-employed group. Originally, our didactical idea of presenting that broad group was to traceably show the reader the interesting differences between “not-employed” and “employed”. Instead, we decided to integrate more information on the employment status into the full sample model and include three-way interactions.

Please see p. 9:
Employment status was split into five categories (0 = employed, 1 = retired, 2= unemployed, 3= permanently sick or disabled and 4 = homemaking respondents).

Please see p. 10:
Finally, three-way interactions were computed to elaborate the role of the employment status within the interaction between health and social relationships (health*social relationship*employment status).

Please see results section:
Table 1 and Figures

13. Page 8, lines 38-42:
Is the “making ends meet” question a validated one? Where did it come from? Please comment on this in the paper.

Responses to the comments:
Please see p. 19:
Ultimately, caution is needed drawing conclusions from analyses using self-rated health. The same holds true for the variable of self-perceived economic status, even though the assessment of the ease with which households can “make ends meet” compared to income represents an adequate and direct measure of the economic situation of individuals, especially among older individuals [70].

Reference:
• 70. Börsch-Supan A, Brugiavini A, Jürges H, et al. First results from the Survey of Health, Ageing and Retirement in Europe (2004-2007). Starting the longitudinal dimension Mannheim: MEA. 2008.

14. General comment on methods:
Please consider any correlation between similar variables. Please also consider where confounding might take place (e.g. education and income). Please discuss this in the paper.

Responses to the comments:

Due to the revised analyses, the social relationship variables (social integration index, contact frequency and emotional closeness) are now presented in separate models. Pairwise correlations were low to moderate and inconspicuous.

Please see p. 9:
Sociodemographic (gender, age) and socioeconomic (education, employment status, make ends meet) factors were used as covariates (Supplement Table 1).

15. Page 8, line 51.
Number of GP visits is treated as a count variable. What would happen if the responses were grouped into categories, depending on the range of responses available. Does that change the results? It seems that this kind of sensitivity analysis could potentially be informative, given the 12-month recall period.

Responses to the comments:
For the purpose of sensitivity analysis, we grouped the count variable and recalculated our models (please see APPENDIX 1). For the vast majority of results, the "ologit" results are similar to the negative binomial regression models.

16. Table 1:
Please state that Age is in years.
Please add a reminder of the scale for self-perceived health status and social integration index.
Please state the denominator for contact frequency (number of contacts per week/month/year?).
Please explain somewhere in the body of the paper what the difference between n and N is.
Please change all thousand separators to commas instead of full-stops.

Responses to the comments:
Please see p. 10:
Age in years

Please see p. 11:
Self-perceived health (0=poor – 4=excellent)
Social integration index (0=low – 6=high)

Please see p. 11:
Average of contact frequency in social network (0=less than once per month or never – 5=daily)

Please see p. 11:
n = Unweighted sample (= number of observations)
N = Weighted sample (= population size based on survey design)

Please see p. 10 and 11.

17. How does this group of people compare to "average" people in this region? Is the population similar across all 16 countries? Please comment on the representativeness of the study population.

Responses to the comments:
Please see p. 7:
To handle possible selection and participation biases, SHARE offers sample design weights [34, 37] (for further details please see analyses section).

Reference:

- 34. Börsch-Supan A. Survey of Health, Ageing and Retirement in Europe (SHARE) Wave 4. Release version: 5.0.0. SHARE-ERIC. Data set. 2016.
- 37. Malter F, Börsch-Supan A. SHARE Wave 4: Innovations & Methodology. Munich: MEA, Max Planck Institute for Social Law and Social Policy; 2013.

Please see p. 9 and 10:
Due to the complex sample structure, including individual level, household level and country level, a survey design was implemented [34, 45].To account for within-household correlations and between-country differences, households were defined as primary sampling unit and countries as strata. Furthermore, to adjust for variation in selection probabilities by design and for variation in participation probabilities caused by non-response, sample design weights were used [37]. Consequently, the Stata survey command, respectively the survey-package in R were used to handle weighted and stratified data adequately [46-48].

Reference:
- 34. Börsch-Supan A. Survey of Health, Ageing and Retirement in Europe (SHARE) Wave 4. Release version: 5.0.0. SHARE-ERIC. Data set. 2016.
- 45. Lumley T, Scott A. Fitting regression models to survey data. Statistical Science. 2017;32:265-78.
- 37. Malter F, Börsch-Supan A. SHARE Wave 4: Innovations & Methodology. Munich: MEA, Max Planck Institute for Social Law and Social Policy; 2013.
- 46. Stata A. STATA SURVEY DATA REFERENCE MANUAL RELEASE 13. 2011.
- 47. Lumley T. Analysis of complex survey samples. Journal of Statistical Software. 2004;9:1-19.
- 48. Lumley T. Survey: analysis of complex survey samples. R package version 3.32-1. 2017.

18. General comment:
Please list the countries in the study. Should there be a variable for country in any of the analyses, or if not, please describe how the countries are incorporated into the analysis. If the analyses are done separately for each country, do the results match the overall results presented in the paper? This might be an important sensitivity analysis to perform.

Responses to the comments:
Please see p. 6:
Data was collected in 2010 and 2011 from sixteen European countries (Austria, Germany, Sweden, Netherlands, Spain, Italy, France, Denmark, Switzerland, Belgium, Czech Republic, Poland, Hungary, Portugal, Slovenia and Estonia).

The country level is covered by the survey design structure of our models. Country is defined as strata, while the household is the primary sampling unit. Moreover, design weights were applied.

Please see p. 9 and 10:
Due to the complex sample structure, including individual level, household level and country level, a survey design was implemented [34, 45].To account for within-household correlations and between-country differences, households were defined as primary sampling unit and countries as strata. Furthermore, to adjust for variation in selection probabilities by design and for variation in participation probabilities caused by non-response, sample design weights were used [37]. Consequently, the Stata survey command, respectively the survey-package in R were used to handle weighted and stratified data adequately [46-48].

Reference:

• 34. Börsch-Supan A. Survey of Health, Ageing and Retirement in Europe (SHARE) Wave 4. Release version: 5.0.0. SHARE-ERIC. Data set. 2016.
• 45. Lumley T, Scott A. Fitting regression models to survey data. Statistical Science. 2017;32:265-78.
• 37. Malter F, Börsch-Supan A. SHARE Wave 4: Innovations & Methodology. Munich: MEA, Max Planck Institute for Social Law and Social Policy; 2013.
• 46. Stata A. STATA SURVEY DATA REFERENCE MANUAL RELEASE 13. 2011.
• 47. Lumley T. Analysis of complex survey samples. Journal of Statistical Software. 2004;9:1-19.
• 48. Lumley T. Survey: analysis of complex survey samples. R package version 3.32-1. 2017.

19. Table 2:
How were these models chosen?
Are they the best fit models? E.g. please state AIC/BIC in related models.
Which is the best fit – the basic model or the interaction model?
As gender is not significant, is the fit better if that is taken out? What does that mean?
Please explain alpha.
The scale for health status is given as 0-4 in the table, but 1-5 in the text. Please streamline.
Are all the variables with scales treated as continuous variables, e.g. self-perceived health status, education, social integration index, contact frequency, and is this appropriate?
In treating them as continuous, there is an assumption made that the scale has interval properties. Please consider using categorical variables, and discuss and justify the choices made in the paper. The same goes for Age.

Responses to the comments:
The models are built by theoretical considerations following the Behavioral Model of R. M. Andersen (1995). Age and gender are defined as “predisposing characteristics”. Education, employment and economic situation as well as social relationship variables are specified as “enabling resources”. Self-perceived health status is used as “need”.
Since survey designs apply pseudo-likelihoods, there is no “true” likelihood, which is needed to compute AIC.

Reference:
• Andersen RM. Revisiting the Behavioral Model and Access to Medical Care: Does it Matter? Journal of Health and Social Behavior. 1995;36:1-10.

Please see Supplement Table 1-3:
Alpha = estimate of the dispersion parameter

Please see p. 9:
0. Poor, 1. Fair, 2. Good, 3. Very good, 4. Excellent

The variables age, social integration index, contact frequency and emotional closeness have limited range, but can be considered as continuous by nature. Regarding the ordered variables health and education, we are following the intervalist view, on treating ordinal variables as continuous when – at least – five response categories are available (Rhemtulla et al. 2012; Bollen & Barb 1981; Gaito & Yokubynas 1986). Our data and methods are robust for treating the named variables as continuous.

Reference:

• Rhemtulla, M., Brosseau-Liard, P. É., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. Psychological methods, 17(3), 354.
• Bollen, K. A., & Barb, K. H. (1981). Pearson's r and coarsely categorized measures. American Sociological Review, 232-239.
• Gaito, J., & Yokubynas, R. (1986). An empirical basis for the statement that measurement scale properties (and meaning) are irrelevant in statistical analyses. Bulletin of the Psychonomic Society, 24(6), 449-450.

20. General comment on methods:
It does not seem entirely clear why simultaneously including the three interaction terms between health status and the three social network/integration variables should help elucidate the role that social network/integration plays in the relationship between health status and frequency of GP visits. Why these three? Are they thought to be independent of each other, or are they correlated at all? Why include the interaction terms simultaneously? Which model actually gives the best fit to the data? Was more work done before arriving at these as the models? What if only one interaction term were included, for example?

Responses to the comments:
Due to the revised analyses, the social relationship variables (social integration index, contact frequency and emotional closeness) are now presented in separate models.

Following Holt-Lunstad et al. (2010), social relationships can be categorized into structural and functional dimensions. The “social integration index” represents the structural dimension including partnership, size of social network and number of activities. The “contact frequency within the social network” is defined as structural as well, though it implies functional elements of social relationships as more contacts means more access to resources and social support and has an impact on health (Cornwell et al. 2009). The variable “number of emotionally close ties” stands for the functional dimension of social relationships. “High-quality relationships are the most likely to provide individuals with a sense of belonging and self-esteem” (Cornwell et al. 2009).

References:
• Holt-Lunstad, J., Smith, T. B., & Layton, J. B. (2010). Social relationships and mortality risk: a meta-analytic review. PLoS medicine, 7(7), e1000316.
• Cornwell, B., Schumm, L. P., Laumann, E. O., & Graber, J. (2009). Social Networks in the NSHAP Study: rationale, measurement, and preliminary findings. Journals of Gerontology Series B: Psychological Sciences and Social Sciences, 64(suppl_1), i47-i55.

Please see also: comment nr. 14.

21. Page 11, line 7:
This mentions models 1a and 1b as looking at employed respondents, but it seems from Table 2 that models 1a and 1b include the whole sample. Please check numbering of models.

Responses to the comments:
Please see Supplement Table 1-3:
We renumbered the models in the text and tables.

22. General comment on methods:
As mentioned above, I am not sure that grouping together retirees, home-makers, those too ill to work, those with disabilities such that they cannot work, and those simply unemployed into a single category is meaningful (models 2a and 2b). Please discuss this in the paper.

Responses to the comments:
We followed your recommendation and dissolved the not-employed group. Originally, our didactical idea of presenting that broad group was to traceably show the reader the interesting differences between “not-employed” and “employed”. Instead, we decided to integrate more information on the employment status into the full sample model and include three-way interactions.

Please see p. 9:
Employment status was split into five categories (0 = employed, 1 = retired, 2= unemployed, 3= permanently sick or disabled and 4 = homemaking respondents).

Please see p. 10:
Finally, three-way interactions were cpmputed to elaborate the role of the employment status within the interaction between health and social relationships (health*social relationship*employment status).

Please see results section:
Table 1 and Figures

23. Page 16, lines 14-18:
I am not very familiar with this literature, but is it also possible that there is a variety of definitions used in the research, and variable data quality, gathered using a variety of different questions/questionnaires and in different contexts? This could also lead to inconsistencies in the results found.

Responses to the comments:
We agree. Partly, inconsistencies could be related to variations in definitions and questions, survey design and heterogeneity in data and context.

24. Page 16, lines 23-49:
The last sentence here is intriguing – perhaps clarify how this idea fits together with the rest of the paragraph.

Responses to the comments:
Due to missing lines, we could not find the paragraph you were referring to – sorry.

25. Page 16, last line:
Please reference which table this result is in.

Responses to the comments:
Please see p. 16 and 17:
We added the model/figure information.

26. Page 17, line 5:
The health question is about health status, and not health needs. Please discuss the relationship between these two concepts.

Responses to the comments:
Please see p. 19:
The limited level of information of self-reported data holds also true for all other variables in our analyses, especially for the variable “self-rated health” which is culturally sensitive [63]. Although, self-rated health status is based on a single item, it is a suitable summary of health [64]. Studies on several representative samples showed that health ratings can be used as valid measures of health

regardless of different cultures and social conditions [65-67]. Furthermore, self-rated health is used as a substitute for health needs in this study. To predict need for and use of health care services, perceived health status corresponds well to the objective health status [68, 69]. Consequently, using self-perceived health only represents an approximation of health needs, since SHARE did not include items on (perceived) health needs.

Reference:

- 63. Jylhä M, Guralnik JM, Ferrucci L, et al. Is Self-Rated Health Comparable across Cultures and Genders? The Journals of Gerontology: Series B. 1998;53B:S144-S52.
- 64. Singh-Manoux A, Martikainen P, Ferrie J, et al. What does self rated health measure? Results from the British Whitehall II and French Gazel cohort studies. Journal of Epidemiology & Community Health. 2006;60:364-72.
- 65. Miilunpalo S, Vuori I, Oja P, et al. Self-rated health status as a health measure: the predictive value of self-reported health status on the use of physician services and on mortality in the working-age population. Journal of clinical epidemiology. 1997;50:517-28.
- 66. Mossey JM, Shapiro E. Self-rated health: a predictor of mortality among the elderly. American journal of public health. 1982;72:800-8.
- 67. DeSalvo KB, Bloser N, Reynolds K, et al. Mortality prediction with a single general self-rated health question. Journal of general internal medicine. 2006;21:267-75.
- 68. Hunt SM, McKenna S, McEwen J, et al. A quantitative approach to perceived health status: a validation study. Journal of Epidemiology & Community Health. 1980;34:281-6.
- 69. Johnson RJ, Wolinsky FD. The structure of health status among older adults: disease, disability, functional limitation, and perceived health. Journal of health and social behavior. 1993:105-21.

27. Page 18, line 39:

The reference cited states that “Older age was generally associated with underreporting”. Please therefore discuss this bias with consideration of the direction of bias, rather than just mentioning “risk of memory bias”.

Responses to the comments:

Please see p. 18 and 19:

The time span covering the GP contacts is quite long, and considering the older age of the interviewed individuals, risk of memory bias is existent with regard to self-reported utilization data [62]. Bhandari and Wagner found in their systematic review on self-reported utilization of health care services that “[…] age was the most consistent demographic factor associated with self-report inaccuracy […]” by older adults underreporting their use [62]. Consequently, intercepts and age coefficients in our models could be potentially underestimated.

Reference:

- 62. Bhandari A, Wagner T. Self-reported utilization of health care services: improving measurement and accuracy. Medical Care Research and Review. 2006;63:217-35.

28. Page 18, line 46:

It is also only a single question. Self-reported health status is often assessed using validated questionnaires such as the SF-12, or the EQ-5D, etc., rather than with a single question. Please comment on this in the paper.

Responses to the comments:

Unfortunately, SHARE did not assess health status by instruments such as the SF-12 or the EQ-5D. Instead, summary measures such as self-perceived health status were included. It has been

recommended by the WHO and used by the Health and Retirement Survey (HRS) and the English Longitudinal Study of Ageing (ELSA).

Please see also: comment nr. 26 above.

Please see p. 19:
Although, self-rated health status is based on a single item, it is a suitable summary of health [64]. Studies on several representative samples showed that health ratings can be used as valid measures of health regardless of different cultures and social conditions [65-67].

Reference:

- 64. Singh-Manoux A, Martikainen P, Ferrie J, et al. What does self rated health measure? Results from the British Whitehall II and French Gazel cohort studies. Journal of Epidemiology & Community Health. 2006;60:364-72.
- 65. Miilunpalo S, Vuori I, Oja P, et al. Self-rated health status as a health measure: the predictive value of self-reported health status on the use of physician services and on mortality in the working-age population. Journal of clinical epidemiology. 1997;50:517-28.
- 66. Mossey JM, Shapiro E. Self-rated health: a predictor of mortality among the elderly. American journal of public health. 1982;72:800-8.
- 67. DeSalvo KB, Bloser N, Reynolds K, et al. Mortality prediction with a single general self-rated health question. Journal of general internal medicine. 2006;21:267-75.

29. General comment:
There is also no mention of visits to the nurse at the GP practice. This would be an important thing to include if this study were to take place in the UK – please consider commenting on this if it is relevant to the countries in the study.

Responses to the comments:
We agree. Focusing on GPs and doctors at health care centers could create an underestimation of primary health care services in some countries. The magnitude may vary by country. Unfortunately, we do not have information on the proportion of nurse-specific services in general practices by country.

Please see p. 18 and 19:
The question used in SHARE to cover the use of GP services across 12 months is established in health services research [4, 20, 21, 61], it has some methodological drawbacks. The question is narrowed to contacting a GP or doctor in a health care center. Contacts with nurses at GP practices are not taken into account. Potentially, the level of using primary care is underestimated.

References:

- 4. Korten AE, Jacomb PA. Predictors of GP service use: a community survey of an elderly Australian sample. Australian & New Zealand Journal of Public Health. 1998;22:609-15.
- 20. Li Y, Chi I. Correlates of physician visits among older adults in China: the effects of family support. J Aging Health. 2011;23:933-53.
- 21. Miltiades HB, Wu B. Factors affecting physician visits in Chinese and Chinese immigrant samples. Soc Sci Med. 2008;66:704-14.
- 61. Gobbens RJJ, van Assen MALM. Frailty and its prediction of disability and health care utilization: The added value of interviews and physical measures following a self-report questionnaire. Archives of Gerontology and Geriatrics. 2012;55:369-79.

30. Page 19, line 49:

"GPs should take "employment status" and "social relationships" of their patients into account." This is very bold. Is it truly justified by this analysis? How could it be done in a fair way, and what would be the benefit to patients?

Responses to the comments:
We rewrote the conclusion section.
Please see p. 20 and 21

31. Page 19, lines 52-56:
This is also bold: "Finally, our results indicate the necessity to integrate information on social relationships and employment status into debates on needs-based access to health care and adequate levels of treatment." Please discuss this further and state the justification.

Responses to the comments:
We rewrote the conclusion section.
Please see p. 20 and 21

32. STROBE checklist:
Please address the impact of having missing data. They seem to have been ignored, which is ok if we know they are missing at random, but I am not sure that we do know this. If they are assumed to be missing at random, then the rationale for assuming this should be discussed. Please comment on this in the paper. Was any information available on people's reasons for not taking part?
Please describe any efforts made during the analysis to address sources of bias – see various comments above.
Please consider the direction of all biases – see comments above.
Please describe what sensitivity analyses were done – see various comments above.
The interpretation is not always cautious. There are some limitations on the source data and analysis methods, so perhaps more caution could be used?

Responses to the comments:
We updated the STROBE checklist with regard to your comments. Furthermore, we gathered information on your comments (please see above) and added potential aspects of bias and limitations to the limitations section (please see p. 18-20). Unfortunately, the SHARE consortium does not offer information to every discussed aspect in needed details.

33. ETHICS
Research Ethics approval for the original study or for this analysis is not mentioned.

Responses to the comments:
Please see p. 6:
"SHARE has been submitted to, and approved by, the ethics committee at the University of Mannheim which was the legally responsible entity for SHARE during wave four" [37].

Reference:
• 37. Malter, F., Börsch-Supan, A.(Eds.)(2013). SHARE Wave 4: Innovations & Methodology. Munich: MEA, Max Planck Institute for Social Law and Social Policy.

34. English:
Some slightly unnatural phrases, e.g. "…moderation of social relationships on…" could perhaps be phrased as "….moderating action of social relationships on…".

Also, “SR did matter in diverse facets in the not-employed-group” sounds a bit strange. Perhaps something like: “SR did appear to be significant in some sections of the not-employed group” (if that is what is meant).
Decimal points should be full-stops (and not commas), and thousand separators should be commas (and not full-stops), e.g. “56,989 interviews”. Please change this in all figures, and all tables.
“The Social Integration Index by Berkmann et al. [30] was shown to be a reliable and robust approach…” – this implies that it was shown to be reliable by this study. If this is not the case, then perhaps re-phrase to something like: “The Social Integration Index by Berkmann et al. has been shown [30] to be a reliable and robust approach…”, if this is what was meant.

Responses to the comments:
We changed the wording according to your suggestions.
We changed full-stops in all necessary places.

Reviewer #2
35. It seems that the process of the IRB approval process was lacking in this manuscript. Please provide it.

Responses to the comments:
Please see p. 6:
“SHARE has been submitted to, and approved by, the ethics committee at the University of Mannheim which was the legally responsible entity for SHARE during wave four” [37].

Reference:
• 37. Malter, F., Börsch-Supan, A.(Eds.)(2013). SHARE Wave 4: Innovations & Methodology. Munich: MEA, Max Planck Institute for Social Law and Social Policy.

36. In the abstract, one of objectives was ….(2) if SR moderate the association between health needs and GP visits. However, on page 5, line 24, (2) if social ties moderate the association between health needs and GP use. Please make it consistent.

Responses to the comments:
We made it consistent.

37. Regarding the objectives of this study, please add more evidence-based information in the literature review.

Responses to the comments:
Please see p. 4 and 5:
We rewrote the introduction and added evidence-based information to the introduction section:

• 11. Berkman, L. F., Leo-Summers, L., & Horwitz, R. I. (1992). Emotional support and survival after myocardial infarction. Annals of Internal Medicine, 117(12), 1003-1009.
• 12. Melchior, M., Berkman, L. F., Niedhammer, I., Chea, M., & Goldberg, M. (2003). Social relations and self-reported health: a prospective analysis of the French Gazel cohort. Social science & medicine, 56(8), 1817-1830.
• 27. Lin, N., Ye, X., & Ensel, W. M. (1999). Social support and depressed mood: A structural analysis. Journal of Health and Social behavior, 344-359.
• 28. Orth-Gomér, K. (2009). Are social relations less health protective in women than in men? Social relations, gender, and cardiovascular health. Journal of Social and Personal Relationships, 26(1), 63-71.

• 29. Von dem Knesebeck, O., & Geyer, S. (2007). Emotional support, education and self-rated health in 22 European countries. BMC Public Health, 7(1), 272.
• 30. Rosner, T. T., Namazi, K. H., & Wykle, M. L. (1988). Physician use among the old-old: Factors affecting variability. Medical care, 982-991.
• 31. Cafferata, G. L. (1987). Marital status, living arrangements, and the use of health services by elderly persons. Journal of Gerontology, 42(6), 613-618.
• 32. Krause, N. (1988). Stressful life events and physician utilization. Journal of Gerontology, 43(2), S53-S61.
• 33. Schafer, M. H. (2013). Discussion networks, physician visits, and non-conventional medicine: probing the relational correlates of health care utilization. Social Science & Medicine, 87, 176-184.

38. Please provide a more clear rationale about why this study attempted to investigate the moderate effect of social relationships on the link between health needs and GP visits.

Responses to the comments:
Please see p. 5:
We rewrote some paragraphs of the introduction to provide a more clear rationale. Furthermore, we added some literature and a figure.

Figure 1: Conceptual model of our research questions i), ii) and iii).

Reviewer #3
39. The analysis performed (a negative binomial GLM using probability weights to account for the survey design) assumes that all responses are independent. While this could be a reasonable approximation, it's conceivable that households share health, behavioral, and social factors that might induce correlation among the outcomes, resulting in a lower "effective sample size" for a given sample size. Standard errors derived using the assumption of independence inherent in the GLM may be too narrow (anti-conservative) as a result. Multilevel models should be considered, using random effects to account for within-household correlations in outcome, and if reasonable to the analysts, nested random effects to account for correlations between households within a country. While the point estimates may be similar, the standard errors and confidence intervals may be more accurate if participants within households and households within countries are not effectively independent.
I believe gllamm and other procedures for GLMEs in Stata handle survey weights and nested random effects. To the best of my knowledge at the time of this review, glmer in R's lme4 package only supports frequency weights, not probability weights. Mixed models for negative binomial outcomes in either package are more recent developments, so their ability to handle both survey weights and nested random effects may be limited.

Responses to the comments:
Please see p. 9 and 10:
Due to the complex sample structure, including individual level, household level and country level, a survey design was implemented [34, 45].To account for within-household correlations and between-country differences, households were defined as primary sampling unit and countries as strata. Furthermore, to adjust for variation in selection probabilities by design and for variation in participation probabilities caused by non-response, sample design weights were used [37]. Consequently, the Stata survey command, respectively the survey-package in R were used to handle weighted and stratified data adequately [46-48].

Reference:
• 34. Börsch-Supan A. Survey of Health, Ageing and Retirement in Europe (SHARE) Wave 4. Release version: 5.0.0. SHARE-ERIC. Data set. 2016.

• 45. Lumley T, Scott A. Fitting regression models to survey data. Statistical Science. 2017;32:265-78.
• 37. Malter F, Börsch-Supan A. SHARE Wave 4: Innovations & Methodology. Munich: MEA, Max Planck Institute for Social Law and Social Policy; 2013.
• 46. Stata A. STATA SURVEY DATA REFERENCE MANUAL RELEASE 13. 2011.
• 47. Lumley T. Analysis of complex survey samples. Journal of Statistical Software. 2004;9:1-19.
• 48. Lumley T. Survey: analysis of complex survey samples. R package version 3.32-1. 2017.

Following your formulation concerning biased standard errors, we compared the results including standard errors of survey design models and negative binomial mixed models (packages lme4 und glmmTMB) (each with rescaled sample weights) on the basis of smaller samples. The standard errors of the survey design model are the most conservative (please see APPENDIX 2).

Nonetheless, we tried to compute multilevel models for negative binomial outcomes with the “svy” command in Stata. Stata could not perform that kind of model due to the vast number of primary sampling units. Furthermore, we also tried to replicate that in R, and we got a full load of warning messages, indicating computational problems, like convergence problems. As shown above, survey design still performs better than mixed models with rescaled sample weights.

40. One assumption of statistical models that could be further described is how the assumption of linearity was addressed: common methods include residual diagnostic plots (added variable or component-residual plots) or the use of smoothing spline terms (mkspline in Stata; bs(), ns(), pspline() or other functions in R).
For ordered categorical predictors (perceived health, social integration, etc.), graphical comparisons of the fitted categorical coefficients can be used.
It is plausible that the effect of an increment in important variables like age or self-assessed health may vary in magnitude depending on the location of the increment (compare ages 50-60 vs. 60-70 or self-assessed health "1. Excellent"-"2. Very Good" vs. "4. Fair"-"5. Poor").
Linearity is more difficult to assess in interactions, but the main effects (especially age and self-assessed health status) should be checked. Inadequate fit (either ignoring relevant confounding variables or modeling of non-linear associations as linear) may result in residual confounding.
Scaling age in terms of decades instead of years may make the coefficient easier to interpret.

Responses to the comments:
Following your recommendation we provide our assessment of linearity in APPENDIX 3. After analysing the graphs we decided that there is no urgent need of integrating spline terms into our models.

41. Changes in statistical significance can arise when interaction terms are introduced even in the absence of a genuine interaction because of variance inflation: interaction terms are correlated with the main effects, resulting in wider standard errors and confidence intervals.
This variance inflation would be more pronounced if there were appreciable correlation between the three aspects of social relationships measured.
Inference should focus on the direction and magnitude of interaction terms and their confidence intervals, as change in the statistical significance of the main effect once an interaction is introduced is not necessarily meaningful (See Gelman and Stern's 'The Difference Between "Significant" and "Not Significant" is not Itself Statistically Significant' (http://www.tandfonline.com/doi/abs/10.1198/000313006X152649), especially when they may naturally arise from variance inflation.

Responses to the comments:
Although, the correlations are only low to moderate, we followed your recommendation and only investigate the three social relationship variables and their interactions in separate models (please see: Figures 2-8 and Supplement Tables 2-4).
Furthermore, we anticipated your suggestion of focussing on the direction and magnitude of interaction terms and their confidence intervals. This perspective is supported by our approach to present figures to visualize the results, especially the 2-way and 3-way interactions.

42. Please state whether or not subgroup analyses were pre-specified in the analytic plan.

Responses to the comments:
While we were investigating i) the association between social relationship variables and GP visits and ii) potential moderation of social relationship variables on the association between health and GP visits, we also controlled for age and gender and we found inconsistent age coefficients. Consequently, we assumed that life events (in this case represented by employment status such as retirement or unemployment) were a possible explanation, worthwhile to look at. Throughout the course of our study, the original research question remained the same.

43. In table 1 (page 9), the cohort is described in aggregate: would it be possible to provide the aggregate results as well additional columns describing individuals by employment status (employed/retired/unemployed/disabled/homemaker)? These may be useful in understanding the stratified results.
In addition, education and self-assessed health status are categorical variables - Reporting the category frequencies may be more interpretable than the mean and standard deviation of the numeric category labels.

Responses to the comments:
Following the reviewers' comments we are not conducting the subgroup analyses via separate models. Since we integrated the employment status into the full sample model, we added information on employment within table 1.
We added the information on education and health status in table 1.

44. The intercept may be made more meaningful by centering the covariates (subtracting a reference value - either some clinically meaningful values or the sample mean value), giving the average number of visits for individuals with reference levels for all covariates. Please mention if centering either has or has not been performed.
It's also noted that if respondents had more than 98 contacts with their doctor about their health, the number was censored at 98. Please list the frequency and proportion of such events. If very rare, censoring may have minimal impact on results, but if appreciable, this can bias means and variances downward, and confidence intervals may be too narrow as a result.

Responses to the comments:
Following your recommendation, we centered the variables age, social integration index, contact frequency and emotional closeness at theirs means.

0.25% of all observations (in our analysed models) were censored at 98 (= 110 observations out of 44,133).

45. What were the correlations between social integration, average frequency of contact, and number of 'very close'/'extremely close' people in social networks?
If these are moderately or strongly correlated, having all of these terms plus interactions with all of these terms in the same model makes collinearity a potential concern.

Polychoric correlations (polychoric in Stata) might be useful here because of the ordinal nature of the covariates.

Responses to the comments:
Due to the revised analyses, the social relationship variables (social integration index, contact frequency and emotional closeness) are now presented in separate models. Polychoric correlations were low to moderate between the three social relationship variables.

Please see p. 10:
Sociodemographic (gender, age) and socioeconomic (education, employment status, make ends meet) factors were used as covariates (Supplement Table 1).

46. On page 8, self-assessed heath is rated on a scale of "1. Excellent" to "5. Poor," with higher numbers indicating worse health. In regression models, if self-assessed health was treated as a linear covariate in model 1a, each unit increase on this scale (towards poorer health) was associated with a 26% reduction in the rate of GP visits (IRR=0.74). It seems counterintuitive to me that those reporting worse self-assessed health would report lower frequency of GP utilization (although my intuition might be in error). Please confirm that this variable has not been re-scaled such that higher values indicate better self-assessed health.

Responses to the comments:
Please see p. 8 – we corrected the scaling within the text:
0. Poor, 1. Fair, 2. Good, 3. Very good, 4. Excellent

47. The buffer function of social integration is one of many possible explanation of the negative association with GP usage. Since the social integration includes participation in charity work, education, social clubs, religious services, or community/political organizations, could this also represent a healthy user effect whereby those who are less infirmed or spend less days in healthcare utilization are more able to participate in such events or maintain/expand their social networks?

Responses to the comments:
Please see p. 18:
When interpreting the results, some methodological limitations need to be taken into account. Firstly, our analyses were based on cross-sectional data, forbidding statements on causal directions and changes over time. The cross-sectional design was chosen due to the inclusion of social relationship variables from SHARE's "social networks" module which was applied only in wave four [34, 35, 60]. Therefore, the postulated buffer function of social integration (of retirees and homemakers) on GP visits, for instance, is only one possible explanation. Another scenario may be the healthy user effect due to volunteering activities which are included in the social integration index. Healthier people with less GP visits have more resources to invest into their social integration.

Reference:

- 34. Börsch-Supan A. Survey of Health, Ageing and Retirement in Europe (SHARE) Wave 4. Release version: 5.0.0. SHARE-ERIC. Data set. 2016.
- 35. Börsch-Supan A, Brandt M, Hunkler C, et al. Data Resource Profile: the Survey of Health, Ageing and Retirement in Europe (SHARE). Int J Epidemiol. 2013;42:992-1001.
- 60. Litwin H, Stoeckel K, Roll A, et al. Social Network Measurement in SHARE Wave Four. 2013. In: SHARE Wave 4: Innovations & Methodology [Internet]. Munich: MEA, Max Planck Institute for Social Law and Social Policy; [18-37].

48. Consider a mention of multiplicity to accompany interpretation: three interaction terms were estimated in several models: in aggregate, Employed vs. Unemployed, and within the Unemployed (Retired, Unemployed, Sick/Disabled, and Homemakers).

Responses to the comments:
Due to the revised analyses, subgroup analyses on employment status are integrated into the aggregate model. The social relationship variables (social integration index, contact frequency and emotional closeness) are now presented in separate models.

49. In most places, results are discussed in the language of associations, but on page 17 (lines 22-28) and 19 (lines 32-36), the language sounds more causal in tone ("lowers", "increase", "decrease"). Please consider "was associated with lower/higher" instead, which is consistent with the rest of the discussion.

Responses to the comments:
We made the discussion and conclusion consistent with regard to the tone.

Please see p. 17:
A higher level of social integration was associated with lower rates of GP use for retirees, but was associated with a higher frequency of visits for unemployed older adults, especially for unemployed older people with a poor self-rated health (Figure 6).

Please see p. 20:
Our results demonstrate that different indicators of social relationships are not associated with higher or lower frequency of GP visits. The magnitude of the associations is relatively low and the minority of the investigated associations is statistically significant.

Reviewer #4
50. My primary concern is that the way the models were constructed and how they were presented makes it difficult to understand the bivariate relationships between each exposure of interest and the GP visits outcome as they are only shown adjusted for blocks of other variables that may or may not be confounders of the relationship of interest. This is particularly true of the three inter-related social relationship variables. Including them all in the same models when they are likely highly collinear with each other is likely to produce biased estimates (e.g. What does the coefficient for social integration index adjusted for number of extremely close people in social network mean?). Further, pulling each of these estimates out from the same model and interpreting them as estimates of unconfounded relationships is problematic.
See: Westreich, Daniel, and Sander Greenland. "The table 2 fallacy: presenting and interpreting confounder and modifier coefficients." American journal of epidemiology 177.4 (2013): 292-298.

Responses to the comments:
Due to the revised analyses, the social relationship variables (social integration index, contact frequency and emotional closeness) are now presented in separate models. Pairwise correlations were low to moderate and inconspicuous.
Instead, we decided to integrate more information on the employment status into the full sample model and include three-way interactions.

Please see p. 9:
Employment status was split into five categories (0 = employed, 1 = retired, 2= unemployed, 3= permanently sick or disabled and 4 = homemaking respondents).

Please see p. 10:

Finally, three-way interactions were computed to elaborate the role of the employment status within the interaction between health and social relationships (health*social relationship*employment status).

Please see p. 10:
Sociodemographic (gender, age) and socioeconomic (education, employment status, make ends meet) factors were used as covariates (Supplement Table 1).

Please see results section:
Table 1 and Figures

51. The authors run multiple models in several different subpopulations without adjustment for multiple comparisons.

Responses to the comments:
Due to the revised analyses, subgroup analyses on employment status are integrated into the aggregate model. The social relationship variables (social integration index, contact frequency and emotional closeness) are now presented in separate models.

VERSION 2 – REVIEW

REVIEWER	Molly Rosenberg Indiana University School of Public Health-Bloomington, USA
REVIEW RETURNED	27-Nov-2017

GENERAL COMMENTS	I reviewed the resubmission of the manuscript “Social relationships and GP use of older adults in Europe: A moderator analysis” for statistical methodology and analysis. The authors have sufficiently addressed my previous comments.

REVIEWER	Chiu-Yueh Hsiao Asia University
REVIEW RETURNED	12-Dec-2017

GENERAL COMMENTS	Clearly, the authors made efforts to address the reviewers' comments and suggestions. I do not have further question.

REVIEWER	Dr Caroline S Clarke UCL, UK
REVIEW RETURNED	16-Dec-2017

GENERAL COMMENTS	bmjopen-2017-018854 - revised Social relationships and GP use of middle-aged and older adults in Europe: a moderator analysis Comments for the authors: Thanks for the amendments that you have made to your models and write-up. The paper is greatly improved. There are a number of points raised by the original reviewers which have not been addressed, especially the use of linear scales for age. Why is age not put into categories? By leaving it as just the age in years, you are assuming that a change from e.g. 50 years of age to 51 years of

	age is equivalent to a change of 85 to 86 years of age. Why are you assuming this? This is a very strong assumption to make. Please either use categories, or give a clear justification for not doing so. There is the same issue with number of GP contacts. The page numbering here relates to the page numbers in the tracked version of the paper. - Throughout: "Contact frequency" is ambiguous – please consider calling it "social contact frequency" to remove any confusion with this possibly meaning the frequency of GP contacts or any other type of contacts. This is especially an issue in the Results section of the Abstract, as it seems from there that you are saying that people who contacted the GP more had more GP visits, which is obviously a redundant point to make. Calling it "social contact frequency" would fix this. - Throughout: Please do not confuse health needs and self-rated health status. Please refer to self-rated health status where this is what you are discussing (i.e. the answer to the single question about health status). Please also discuss the limitations of this approach to using this variable as the only measure of people's health status, and why you are interested in using this as a proxy for health needs. It is conceivable that the number of times people have gone to the GP is just as good a marker for "need" as that single question on their health status. Please carefully consider this, and discuss it in the paper, and justify the methods you are using. - Throughout: You refer to the rate and frequency of GP visits at points through the paper, but I think what you are referring to actually is the number of times people have reported visiting the GP over the last 12 months. Please clarify this, and be careful with the wording. - Page 6: "On the other hand, social relationships are closely linked to health needs." Please give a reference for this. - First paragraph of Data and Methods (page 7): The eligibility criteria imply that people must have spouses to be eligible for the study. Is this right? If not, please re-word. Please also add that the ethical approval for the SHARE study also provides for this analysis (if this is true). - Page 9: How did people respond if they lived with someone but were not married to them (e.g. co-habiting partner, sibling, children, etc.)? How might the lack of an option like this have generated bias in the study? Would this vary by country/language? Please discuss this in the paper. - Page 11: "Consequently, the Stata survey command, respectively the survey-package in R were used to handle weighted and stratified data adequately" Please fix the language in this sentence.

	- General comment on methods: Please consider any correlation between similar variables. Please also consider where confounding might take place (e.g. education and income). Please discuss this in the paper. - Page 8, line 51. Number of GP visits is treated as a count variable. What would happen if the responses were grouped into categories, depending on the range of responses available? Does that change the results? It seems that this kind of sensitivity analysis could potentially be informative, given the 12-month recall period. - Table 1: What is N for? It is not mentioned in the paper except in this table. Can it be omitted? How does this group of people compare to "average" people in this region? Is the population similar across all 16 countries? Please comment on the representativeness of the study population. When giving the median, please also include the inter-quartile range. For the Education categories, please explain approximately what they are, i.e. "none" [code 0] should be "pre-primary"; code 1 = primary, etc. - Page 14: It could be more meaningful to give (IRR=X.XX, 95%CI Y.YY – Z.ZZ) instead of the p-values in the text. Please consider doing this. - Throughout The use of "predicted" is problematic, as we don't know how good a fit any of the models actually are. Please discuss how good the models are. - Page 15: "This association is statistically significant for people with a very good and excellent health, although the slope differences are relatively small (approx. 0.5 GP visits per year)." Are you talking about differences in gradient (i.e. slope) here? Please clarify the wording. - Page 15: Please change "contrarily to" to "in contrary to". - Top of page 16: Are these results from sub-groups of the study population who had ticked different employment types? Please clarify. - Page 16: Could some of the differences be due to other things, e.g. people who work have less time available to go to the GP? Please consider and discuss. The actual differences in numbers of GP contacts between the groups are interesting. Can you compare your findings for these to other published work?

	- Page 26: "On the one hand, our results are in line with a number of studies on structural and functional aspects of social ties [5, 24, 25, 51-53]." Please expand on this. - Page 27 Please change "kids" to "children". - Throughout Please do not start a sentence with "Though," or "Although,". - Page 28: "Consequently, emotional closeness within social networks enables and fosters the utilization of GP services, especially for those individuals that report high levels of health needs." Please give a reference for this. You cannot infer this from your analysis as it is presented here. - Page 28: "All in all, the groups of retired, unemployed, permanently sick/disabled and homemaking people show a higher predicted number of GP visits, especially, if they are unemployed, permanently sick/disabled or homemakers." This is a circular argument. Please remove/clarify what you mean. - Page 29 Please change "GP services" to "GP doctors" as you have not looked at other primary care services. - Page 29: "Homemakers use more GP visits, if their social contact frequency is higher, especially, if their health status is rated as intermediate" Please change "intermediate" to either low or high, and check that this is the right way round. (social contact frequency lower?) - Page 31 On this page you start talking about just "health". Do you mean self-rated health status? Please clarify. - Page 31: "Furthermore, self-rated health is used as a substitute for health needs in this study." Yes, but is this actually a reasonable thing to do, in this context? Please justify this and give appropriate references – see my earlier comments. What's wrong with just discussing self-rated health status? - Page 31: "To predict need for and use of health care services, perceived health status corresponds well to the objective health status" This is confusing. There are a lot of different concepts in this sentence and I'm not sure what you are trying to relate to what. - Page 31 "The same holds true for the variable of self-perceived economic status, even though the assessment of the ease with which households can "make ends meet" compared to income represents an adequate and direct measure of the economic situation of individuals, especially among older individuals" This is a bit contradictory. What are you trying to say?

	- Page 32: ".. the minority… is statistically significant." Please fix the language in this sentence. - Page 32-33 paragraph starting "Since…" This is much better, thanks very much! - General comment on methods: Regarding the three interaction terms between health status and the three social network/integration variables: their use in the new separate models is slightly clearer now, but it is still not clear which model actually gives the best fit to the data? If this is not what you are interested in, then please explain further. - General comment on methods: Thanks for splitting the non-employed people into their own categories. It is now not clear how that information has been included. It seems from Supplementary Table 2 that there are now 5 separate yes/no variables – one for each employment category. Why was it done this way instead of having a single 5-category factor variable? Please clarify and justify your modelling choice. - STROBE checklist: Thanks for using the checklist. - ETHICS Please state that this analysis is provided for under the Ethics approval for the original study (if this is true). - Figure 1 This is a bit confusing still. What is arrow (ii) indicating? Is it the effect of the relationship between social relationships and GP visits on the relationship between health needs (which you are not looking at) and GP visits? And then arrow (iii) is the effect of employment status on said effect (ii)? - Figures 3-8 Please put in the actual words in the keys for these figures, rather than -1 SD and +1 SD. - Supplementary table 4 There are a lot of variables for each model! Are they all really separate variables or are some multi-category factor variables? Please clarify.

VERSION 2 – AUTHOR RESPONSE

	Reviewer #1	Responses to the comments
1.	Why is age not put into categories? By leaving it as just the age in years, you are assuming that a change from e.g. 50 years of age to 51 years of age is equivalent to a change of 85 to 86 years of age. Why are you assuming this? This is a very strong assumption to make. Please either use categories, or give a clear justification for not doing so. There is the same issue with number of GP contacts.	We agree. We performed an assessment of linearity. Our assessment of linearity showed that there is no need to transform variables, e.g. age, by integrating spline terms into our models (please see APPENDIX 3). Please see our feedback on comment number 10.
2.	Throughout: "Contact frequency" is ambiguous – please consider calling it "social contact frequency" to remove any confusion with this possibly meaning the frequency of GP contacts or any other type of contacts. This is especially an issue in the Results section of the Abstract, as it seems from there that you are saying that people who contacted the GP more had more GP visits, which is obviously a redundant point to make. Calling it "social contact frequency" would fix this.	We changed the wording throughout the document.
3.	Throughout: Please do not confuse health needs and self-rated health status. Please refer to self-rated health status where this is what	Following your suggestion we changed the wording throughout.

	you are discussing (i.e. the answer to the single question about health status). Please also discuss the limitations of this approach to using this variable as the only measure of people's health status, and why you are interested in using this as a proxy for health needs. It is conceivable that the number of times people have gone to the GP is just as good a marker for "need" as that single question on their health status. Please carefully consider this, and discuss it in the paper, and justify the methods you are using.	Please see p. 20: *The limited level of information of self-reported data holds also true for all other variables in our analyses, especially for the variable "self-rated health" [65]. Self-rated health status is based on a single item, it is a suitable summary of health status [66]. Studies on several representative samples showed that self-rated health ratings can be used as valid measures of health status regardless of different cultures and social conditions [67-69] and that they may correspond well to the objective health status [70, 71]. Caution is needed drawing conclusions from analyses using self-rated health.*
4.	Throughout: You refer to the rate and frequency of GP visits at points through the paper, but I think what you are referring to actually is the number of times people have reported visiting the GP over the last 12 months. Please clarify this, and be careful with the wording.	Please see p. 7, p. 10, p. 11, p. 19: We changed the wording accordingly to your suggestion at central points of the manuscript to guide the reader adequately.
5.	Page 6: "On the other hand, social relationships are closely linked to health needs." Please give a reference for this.	Please see p. 5: *On the other hand, social relationships are closely linked to health [10, 12, 34].* References: 10. Hemingway H, Marmot M. Evidence based cardiology: psychosocial factors in the aetiology and prognosis of coronary heart disease. Systematic review of prospective cohort studies. BMJ. 1999;318:1460-7. 12. Melchior M, Berkman LF, Niedhammer I, et al. Social relations and self-reported health: a prospective analysis of the French Gazel cohort. Social science & medicine.

		2003;56:1817-30. 34. Almedom AM. Social capital and mental health: An interdisciplinary review of primary evidence. Social science & medicine. 2005;61:943-64.
6.	First paragraph of Data and Methods (page 7): The eligibility criteria imply that people must have spouses to be eligible for the study. Is this right? If not, please re-word. Please also add that the ethical approval for the SHARE study also provides for this analysis (if this is true).	Please see p. 6: *Based on population registers, SHARE uses probability samples within the countries and includes non-institutionalized adults aged 50 years or older and, if available, their partners.* Please see p. 6: *"SHARE has been submitted to, and approved by, the ethics committee at the University of Mannheim which was the legally responsible entity for SHARE during wave four" [38]. Following the SHARE conditions of use, the ethical approval for the SHARE study also applies to this analysis [39].*
7.	Page 9: How did people respond if they lived with someone but were not married to them (e.g. co-habiting partner, sibling, children, etc.)? How might the lack of an option like this have generated bias in the study? Would this vary by country/language? Please discuss this in the paper.	Please see p. 21: *Another point that can be discussed is that one out of three domains of the social integration index focused on marital and partnership status and cohabitation. That focus cannot capture the whole variety of non-married or non-partner cohabiting household structures. Potentially, this lack of information is buffered by the other two domains, and especially, by the second domain of the social integration index by including of the number of social ties. Nevertheless, the level of social integration could be slightly higher than illustrated by our index. In particular, this could be true for countries with a higher number of "non-traditional" living arrangements.*
8.	Page 11: "Consequently, the Stata survey command, respectively the survey-package in R were used to handle weighted and stratified data adequately" Please fix the language in this sentence.	Please see p. 10: *In the case of Stata the survey command and in R the survey-package were used to adequately handle weighted and stratified data [48-50].*

9.	General comment on methods: Please consider any correlation between similar variables. Please also consider where confounding might take place (e.g. education and income). Please discuss this in the paper.	Please see p. 9: *The correlation matrix of the covariates did not reveal strong or very strong associations between similar variables (Supplement Table 1). The highest correlation was found between education and financial distress (r = 0.22). Hence, the level of confounding within the following analyses can be rated as low to moderate.*
10.	Page 8, line 51. Number of GP visits is treated as a count variable. What would happen if the responses were grouped into categories, depending on the range of responses available? Does that change the results? It seems that this kind of sensitivity analysis could potentially be informative, given the 12-month recall period.	Please see p. 10: *Regression models were used to analyze the associations between GP use and the predictors. The dependent variable "reported number of GP visits in the last 12 months" is a discrete count variable following a Poisson distribution. As the variance of the dependent variable is greater than its mean, negative binomial regression was used to account for the significant evidence of overdispersion. Furthermore, negative binomial regression models include a parameter that reflects unobserved heterogeneity among observations [46].* Please see p. 19 & p. 20: *The question to assess the use of GP services across 12 months is established in health services research [4, 20, 21, 63], but has some methodological drawbacks. The question is narrowed to the number of times people have reported visiting a GP or doctor in a health care center. Contacts with nurses at GP practices are not taken into account. Potentially, the level of using primary care is underestimated. The time span covering the GP contacts is quite long, and considering the older age of the interviewed individuals, risk of memory bias is existent with regard to self-reported utilization [64]. Bhandari and Wagner found in their systematic review on self-reported utilization of health care services that "[…] age was the most consistent demographic factor associated with self-report inaccuracy […]" by older adults underreporting their use [64]. Consequently, intercepts and age coefficients in our models could be potentially underestimated.* Furthermore, we followed your suggestion of grouping the count variable into categories with regard to sensitivity analysis. We applied ordered logistic regressions on our models (please see APPENDIX 1 of this cover letter). All in all, the results were similar to the negative binomial regression models.
11.	Table 1:	We omitted the N-row.

	What is N for? It is not mentioned in the paper except in this table. Can it be omitted?	
	How does this group of people compare to "average" people in this region? Is the population similar across all 16 countries? Please comment on the representativeness of the study population.	We were able to compare our study population with EUROSTAT data and its population census 2011 (= year of SHARE wave 4) for the variables age, gender and educational attainment (ISCED). Analyses showed that our study population is comparable to the EUROSTAT information, and consequently, it may be rated as representative. Only for the age groups 50-64 and 65-84, we observed moderate differences. In comparison to the EUROSTAT data, the ratio of age group 65-84 is higher in SHARE wave 4 (and lower in the age group 50-64). Please see the detailed data in APPENDIX 4. Reference: https://ec.europa.eu/CensusHub2/query.do?step=selectHyperCube&qhc=false
	When giving the median, please also include the inter-quartile range.	We added the inter-quartile range.
	For the Education categories, please explain approximately what they are, i.e. "none" [code 0] should be "pre-primary"; code 1 = primary, etc.	We added the suggested explanation.
12.	Page 14: It could be more meaningful to give (IRR=X.XX, 95%CI Y.YY – Z.ZZ) instead of the p-values in the text. Please consider doing this.	Please see Abstract and p. 12: Now, we report the 95%CI and omit the p-values.
13.	Throughout The use of "predicted" is	We removed the term "predicted", where appropriate or changed the wording to „estimated" or "expected" as results report marginal means. All in all, our models do not claim

	problematic, as we don't know how good a fit any of the models actually are. Please discuss how good the models are.	transferability and generalizability. Since we are interested in associations, the models are built by theoretical considerations following the Behavioral Model of R. M. Andersen (1995). Age and gender are defined as "predisposing characteristics". Education, employment and economic situation as well as social relationship variables are specified as "enabling resources". Self-perceived health status is used as "need". Since survey designs apply pseudo-likelihoods, there is no "true" likelihood, which is needed to compute AIC. **Reference:** • Andersen RM. Revisiting the Behavioral Model and Access to Medical Care: Does it Matter? Journal of Health and Social Behavior. 1995;36:1-10.
14.	Page 15: "This association is statistically significant for people with a very good and excellent health, although the slope differences are relatively small (approx. 0.5 GP visits per year)." Are you talking about differences in gradient (i.e. slope) here? Please clarify the wording.	Please see p 13 & 14: *All in all, the patterns are similar to Figure 2, but the slopes of the groups with lower and higher contact frequencies are the other way round. The slope of estimated number of GP visits on self-rated health is steeper for those with lower social contact frequency. This association is statistically significant for people with a very good and excellent health, although the differences in the slopes are relatively small.*
15.	Page 15: Please change "contrarily to" to "in contrary to".	We changed the wording.
16.	Top of page 16: Are these results from sub-groups of the study population who had ticked different employment types? Please clarify.	Please see p14: *iii) Moderation of social relationships and employment types on health and GP visits* *To answer research question (iii), Figures 5-7 incorporate the three way interactions between health, social relationships and employment status in relation to the number of GP visits. Figure 5 shows the expected number of GP visits depending on the three-way interaction between health, social integration index and employment status based on the full sample (Supplement Table 4).*
17.	Page 16:	Please see p. 19; we added your suggestion to the section

	Could some of the differences be due to other things, e.g. people who work have less time available to go to the GP? Please consider and discuss. The actual differences in numbers of GP contacts between the groups are interesting. Can you compare your findings for these to other published work?	"limitations": *Furthermore, some of the differences between employment types may be related to temporary resources, since employed people have less time available to consult their GP.* We agree, but we could not find comparable studies including that detailed information on employment status.
18.	Page 26: "On the one hand, our results are in line with a number of studies on structural and functional aspects of social ties [5, 24, 25, 51-53]." Please expand on this.	Please see p. 16: *On the one hand, our results are in line with a number of studies on structural and functional aspects of social ties [5, 24, 53-55]. Studies on structural aspects of social relationships, e.g. marital status, living arrangements and family size, showed no statistically significant associations with the frequency of physician use [53-55]. Furthermore, studies on functional aspects of social relationships, e.g. social anchorage, social support and emotional, instrumental and informational support, demonstrated no statistically significant associations with regard to the use of primary care services [5, 24].*
19.	Page 27 Please change "kids" to "children".	We changed the wording.
20.	Throughout Please do not start a sentence with "Though," or "Although,".	We changed the wording.
21.	Page 28: "Consequently, emotional closeness within social networks enables and fosters the utilization of GP services, especially for those individuals that report high levels of health needs." Please give a reference for this. You cannot infer this from your	We agree and we omitted that sentence, since we could not provide an adequate reference for our statement.

	analysis as it is presented here.	
22.	Page 28: "All in all, the groups of retired, unemployed, permanently sick/disabled and homemaking people show a higher predicted number of GP visits, especially, if they are unemployed, permanently sick/disabled or homemakers." This is a circular argument. Please remove/clarify what you mean.	We changed the wording.
23.	Page 29 Please change "GP services" to "GP doctors" as you have not looked at other primary care services.	We changed the wording.
24.	Page 29: "Homemakers use more GP visits, if their social contact frequency is higher, especially, if their health status is rated as intermediate" Please change "intermediate" to either low or high, and check that this is the right way round. (social contact frequency lower?)	Please see p. 18: *Homemakers use more GP visits, if their social contact frequency is higher, especially, if their health status is rated as fair or good.*
25.	Page 31 On this page you start talking about just "health". Do you mean self-rated health status? Please clarify.	Please see p. 20: *Self-rated health status is based on a single item, it is a suitable summary of health status [66]. Studies on several representative samples showed that self-rated health ratings can be used as valid measures of health status regardless of different cultures and social conditions [67-69] and they may correspond well to the objective health status [70, 71]. Caution is needed drawing conclusions from analyses using self-rated health.*
26.	Page 31:	We agree, omitted the sentences including "health needs"

	"Furthermore, self-rated health is used as a substitute for health needs in this study." Yes, but is this actually a reasonable thing to do, in this context? Please justify this and give appropriate references – see my earlier comments. What's wrong with just discussing self-rated health status?	and focused on discussing self-rated health status.
27.	Page 31: "To predict need for and use of health care services, perceived health status corresponds well to the objective health status" This is confusing. There are a lot of different concepts in this sentence and I'm not sure what you are trying to relate to what.	We simplified the structure; please see p. 20: *Studies on several representative samples showed that self-rated health ratings can be used as valid measures of health status regardless of different cultures and social conditions [67-69] and that they may correspond well to the objective health status [70, 71].*
28.	Page 31 "The same holds true for the variable of self-perceived economic status, even though the assessment of the ease with which households can "make ends meet" compared to income represents an adequate and direct measure of the economic situation of individuals, especially among older individuals" This is a bit contradictory. What are you trying to say?	Please see p. 20: *The same holds true for the variable "make ends meet", since the assessment of self-perceived financial distress compared to income represents an adequate and direct measure of the economic situation of individuals, especially among older individuals [72].*
29.	Page 32: ".. the minority… is statistically significant."	Please see p. 21: *The magnitude of the associations is relatively low and most of the investigated associations are statistically insignificant.*

	Please fix the language in this sentence.	
30.	Page 32-33 paragraph starting "Since…" This is much better, thanks very much!	Ok, thanks.
31.	General comment on methods: Regarding the three interaction terms between health status and the three social network/integration variables: their use in the new separate models is slightly clearer now, but it is still not clear which model actually gives the best fit to the data? If this is not what you are interested in, then please explain further.	Please see our comment number 13.
32.	General comment on methods: Thanks for splitting the non-employed people into their own categories. It is now not clear how that information has been included. It seems from Supplementary Table 2 that there are now 5 separate yes/no variables – one for each employment category. Why was it done this way instead of having a single 5-category factor variable? Please clarify and justify your modelling choice.	We agree, "employment status" is one variable including five categories. Since it is a categorical variable on a nominal scale we had to create dummy-variables and chose a reference group. To guide the reader we reformatted the supplementary tables 2-4.
33.	STROBE checklist: Thanks for using the checklist.	Ok, thanks.
34.	ETHICS Please state that this analysis is provided for under the Ethics approval for the original study (if this is true).	Please see our comment number 6.
35.	Figure 1 This is a bit confusing still. What	Since figure 1 seems to be confusing, we decided to omit it.

	is arrow (ii) indicating? Is it the effect of the relationship between social relationships and GP visits on the relationship between health needs (which you are not looking at) and GP visits? And then arrow (iii) is the effect of employment status on said effect (ii)?	
36.	Figures 3-8 Please put in the actual words in the keys for these figures, rather than -1 SD and +1 SD.	We changed that.
37.	Supplementary table 4 There are a lot of variables for each model! Are they all really separate variables or are some multi-category factor variables? Please clarify.	We reformatted the tables to clarify the structure and to guide the reader. Please see supplement tables 2-4.