Following recent comments on PubPeer https://pubpeer.com/publications/CCAB01FD663B858B53CD2DEE251AC6 regarding our paper[1], we further clarify our statistical methods. To reiterate we examined mental health service contacts in an adolescent cohort (n=1238) where there were 1190 (96%) participants with useable data, but a small percentage of those had a mental disorder (n=126 (11%) of 1190). Of the 126 with mental disorder, 48 (38%) had a past year mental health service contact at baseline. Our investigation tested whether self-reported depression scores were lower 3 years later amongst those with a mental illness at baseline who received a mental health service contact compared with those who did not. The statistical methods chosen aimed to account for: i) the confounding effects of a set of fixed covariates that may have accounted for the relationship of service contact with depression; ii) absence of randomisation to mental health service use at recruitment. We point out that much of the justification for our data analytic strategy is published online in the appendix.

We note in the paper that findings from the multilevel models unadjusted by covariates are presented in the online Supplement (Supplementary Table 4). These findings show that unadjusted imputed and raw depression scores improved more quickly among the disorder-and-services group than the disorder-only group, consistent with the adjusted findings. Employing propensity analysis following the multilevel modelling was a way of more robustly checking the findings from the multilevel modelling. In our supplement we have given the rationale for the appropriateness of using propensity scoring given the sample size. Whilst the supplement is fully available online (http://www.thelancet.com/journals/lanpsy/article/PIIS2215-0366(17)30002-0/supplemental), we are pleased to highlight, and expound upon, our rationale here for the reader's ease.

We used a propensity score method appropriate for small sample sizes as discussed and published in the literature: 'The propensity score method used to check covariate balance between groups and weight the data was inverse probability of treatment weighting (IPTW). IPTW gives correct estimations of treatment effect in small sample sizes…[2]' (Supplement 2). The appropriateness of this method has been demonstrated down to n=40

by Pirracchio et al[2] who '…conducted a series of Monte Carlo simulations to evaluate the influence of sample size, prevalence of treatment exposure, and strength of the association between the variables and the outcome and/or the treatment exposure, on the performance of (IPTW)'. Their findings show that 'Decreasing the sample size from 1,000 to 40 subjects did not substantially alter the Type I error rate, and led to relative biases below 10%'.[2]

We took additional steps to decrease bias in specifying our propensity score model. Pirracchio et al[2] state: 'Including variables unrelated to the exposure but related to the outcome in the PS (Propensity Score) model decreased the bias and the variance as compared to models omitting such variables'. Other findings from Monte Carlo simulation experiments concur with Pirracchio et al[2] and are cited in our appendix. In Supplement 2 we state that: '…baseline covariates correlated to the outcome (MFQ clinical cut-off age 17> ρ=0·10) (were) used to predict baseline mental health service contact…', and indeed several of these are unrelated to the exposure (Supplementary Table 2b).

Furthermore, we ensured we were not overparameterizing the logistic regression model used in generating the propensity score. Whilst many may be familiar with the rule of thumb that logistic models must have a minimum of 10 events per predictor variable (EPV), Vittinghoff & McCulloch[3] cite this rule as being based on simulation studies which only vary the number of events. Their simulation study not only varied the number of events, but also the number of predictor variables, sample sizes, values of the regression coefficient for the primary predictor, multiple correlation of the primary predictor with the model covariates, and prevalence of a binary primary predictor.[3] After examining 9,328 and 3,392 respective scenarios with binary and continuous primary predictors, problematic scenarios (ie: confidence interval coverage less than 93%, type I error rate greater than 7%, or relative bias greater than 15%) were encountered in 7% or less of the models with 5-9 EPV, predominantly with different numbers of predictors than used in the present study. Indeed, such problems were still observed in models with 10–16 EPV. This led them to conclude that: '…systematic discounting of results…from any model with 5–9 EPV does not appear to

be justified',[3] and that relaxing the rules to 5-9 EPV is appropriate. We have not gone below this figure in models used our paper.

After estimating the propensity score, we took further steps to reduce estimate bias: 'Stabilized IPTWs were used to reduce impact of extreme weights, thus reducing estimate bias[4]' (Supplement 2). In Supplement 2 we also indicate we ensured the propensity model was correctly specified, by checking and ensuring the balance of all covariates (not just those in the propensity model) following weighting, as exhorted by Austin.[5] Finally, as explained and cited in our methods[1] "we restricted the analysis to the region of common support—the range of propensity scores which were observed in both treated and untreated individuals" to further reduce estimate bias.

In sum, by using methodology appropriate given the small sample size and taking additional steps to reduce estimate bias, we conclude that our findings can be defended. We did however cite limitations related to a small sample size in our discussion, and exhort that future research assessing the relationship of mental health service contact with subsequent mental health should employ larger sample sizes where possible.

We acknowledge that an odds ratio over-estimates a relative risk when the disease is not rare in the population, typically >10%. Whilst depression was uncommon in the total sample of 1190, at 3% (n=31), in the propensity sample consisting of those only with a mental disorder it was not (25% [31/126]). Therefore our findings from the propensity score analysis should read "the odds of reporting clinical depression were more than sevenfold…" as opposed to "…seven times higher…".

We also wish to take this opportunity to clarify our sample size. Whilst there are 1238 participants recruited to the cohort, we explain in the results that: "1190 adolescents had data for T1 current mental disorder and past-year mental health service contact". There was however additional missing data from the outcome and covariates bringing numbers of respondents with complete data for outcome and all covariates to 983 (83%) for T1, 717 (60%) for T2, and 769 (65%) for T3 (See Supplementary Table 2 column 6 for baseline sample size separate by each covariate). The numbers listed in the main text in the first

paragraph of the results section are however different (being 995, 778 and 856 respectively) and we acknowledge this error in reporting. Listed in column 5 of Table 1 are the sample sizes for completed data for outcome and all covariates and they add up as expected to the total across all time points (2257+140+72 = 2469).

To expand upon this sample size with imputation, we required each self-report questionnaire to be at least partially completed at one time point to be used in imputation. These imputed measures were computed (explained in Supplement 2) across all three time-points with other measures related to attrition. This resulted in the imputed sample size being smaller than 1190. The imputed sample size with covariates was in fact 1120 and we should have noted that in the main paper.

In Table 1 columns 2, 3, and 4 involving imputed adjusted data, the sample sizes indeed do not add up (ie: 2965+202+126 = 3293 not 3302). This n=9 difference was due to differential loss of participants for each of the sub samples as a result of inexact merging of different imputation files for some covariates, resulting in some of the original raw data with missing values being randomly included in some imputations. When this was rectified, for the first set of analyses (by age) in Table 1 this gave n=3360, the second set (by unaffected group) gave n=3015, the third (the disorder-only group), n=213 and the fourth (the disorder-and-services group), n=132.

We recomputed our findings in column 3 and 4 of Table 1 and these are virtually unaltered: if anything they become slightly stronger. As before, 14-year-old adolescents who had contact with mental health services in the past year had a greater decrease in depressive symptoms than those without contact (linear adjusted coefficient –1·68, 95% CI –3·18 to –0·19; p=0·028; quadratic adjusted coefficient –0·56, 95% CI –1·06 to -0·05; p=0·031). By T3, patients in the disorder-and-services group reported significantly fewer symptoms than did those in the disorder-only group (adjusted coefficient -4·76, 95% CI -8·75 to -0·77; p=0·020).

References

1. Neufeld SAS, Dunn VJ, Jones PB, Croudace TJ, Goodyer IM. Reduction in adolescent depression after contact with mental health services: a longitudinal cohort study in the UK. *Lancet Psychiatry.* 2017;4:120–27.
2. Pirracchio R, Resche-Rigon M, Chevret S. Evaluation of the propensity score methods for estimating marginal odds ratios in case of small sample size. *BMC Med Res Methodol*. 2012;12:1–10.
3. Vittinghoff E, McCulloch CE. Relaxing the rule of ten events per variable in logistic and Cox regression. *Am J Epidemiol.* 2007;165:710-8.
4. Robins JM, Hernan MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology.* 2000;11:550–60.
5. Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behav Res.* 2011;46:399–424.